Probing Intuitive Physics Understanding with Natural Language

Modeling Human Intuition Using Flexible Queries to Simulated Worlds

Julius Heitkoetter (JULIUSH@Mit.Edu)
Massachusetts Institute of Technology, 77 Massachusetts Avenue
Cambridge, MA 02139 USA

Abstract

Human intuitions concerning physics are remarkable in their performance and ability to be expressed in natural language. In a heartbeat, humans are able to predict *which direction a tree may fall* even if they have seen very few examples of trees falling. Furthermore, humans are able to express and share these intuitions easily through natural language queries: *How will the tree fall?*, *What will the tree look like after it has fallen?* The ability of humans to reason and communicate physical intuition with relatively few examples outperforms all comparable AI based multimodal and physics engine models. Furthermore, it contributes to one of the biggest open questions in cognitive science: *How do we learn so much with so little?* Modern day research in cognitive science points to hierarchical Bayesian inference as an avenue to solve this problem. In this study, we propose a framework for modeling human intuition using natural language queries which are transformed into a probabilistic language of thought and are used to flexibly probe bayseian inference over a set of simulated worlds. We then show that this framework is able to outperform state-of-the-art multi-modal baselines and exhibit strong correlations to human behavior.

Keywords: Intuitive Physics; Simulated Worlds; Probabilistic Language of Thought; Bayesian World Modeling, Natural Language Processing.

Code: https://github.com/julius-heitkoetter/NaLaPIP

Introduction

Our intuitive physics understanding is fast and can be efficiently probed using natural language. At a glance, we can tell whether *two cars will collide, a stack of boxes will fall*, or *a child is likely to fall off a playground*. Not only can we make these predictions, we can infer qualitative properties of the world by observing these situations: *how fast are the cars moving, how heavy are the boxes,* or *how strong is the child.* Moreover, these predictions and inferences are able to be communicated and probed extremely efficiently through natural language and with few examples beforehand.

How are we able to perform this inference in so little time, with so few examples? Modern AI models such as state-of-the-art Large Language Models (LLMs) still fall short when it comes to replicating human-like behavior in out-of-distribution tasks (Collins, Wong, Feng, Wei, & Tenenbaum, 2022). However, work from Tenenbaum (1999) outlines a Bayesian framework to answer the question of *how do we learn so much from so little?* More modern work articulates this Bayesian framework as a probabilistic language of thought used to convert word models to world models (Wong et al., 2023). Battaglia, Hamrick, and Tenenbaum (2013) apply an extension of this Bayesian framework to intuitive physics, showing evidence that human intuitive physics is modeled as probabilistic inference over a mental physics engine. Further work shows evidence of grounding intuitive physics understanding in natural language through the use of probabilistic programs and simulated worlds (Zhang, Wong, Grand, & Tenenbaum, 2023).

In this study, we expand upon the work of these findings to see if we can probe intuitive physics understandings with natural language through a probabilistic language of thought. We seek to build a model using Bayesian frameworks which we can probe with natural language questions in order to extract evidence of human-like intuitive physics in simplified cases. This study seeks to make two main contributions in this direction:

1. We propose a framework for converting questions in natural language to a probabilistic programming language (PPL) which is then used to probe intuitive physics models that use simulated worlds and probabilistic language of thought. A high level overview of this framework is illustrated in figure 1

2. We develop NaLaPIP, a model for *Natural Language Probing of Intuitive Physics* which uses the above framework to answer questions containing a natural language and a visual stimulus.

Both of these contributions are inspired by stimuli from Battaglia et al. (2013) and approach from (Zhang et al., 2023).

Related Work

Preview work has focused on constructing linguist meaning from cognitive representations in a compositional language of thought (Fodor, 1975; Jackendoff, 1985) and further work which addresses the vagueness often found in natural language (Van Eijck & Lappin, 2012; Cooper, Dobnik, Lappin, & Larsson, 2015; Goodman & Lassiter, 2015a). In our study, we related these findings and developments to intuitive physics problems.

Other work in intuitive physics has explored the ability to use probabilistic inference approaches (K. A. Smith & Vul, 2013; K. Smith et al., 2024) to model human behavior as well

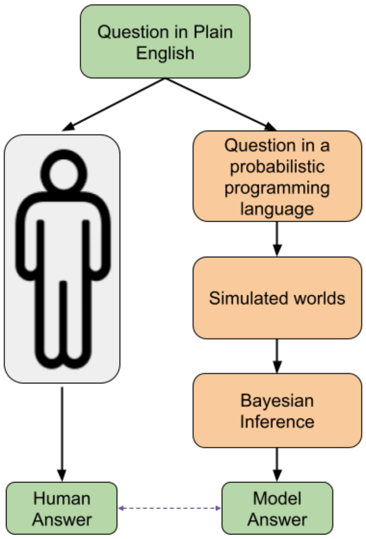

Question in Plain English
Question in a probabilistic programming language
Simulated worlds
Bayesian Inference
Human Answer
Model Answer

Figure 1: An overview of the framework presented in this study. The framework can be modeled as a pipeline, where an intuitive physics question in plain English is converted to a question in a probabilistic programming language, applied to extract answers from simulated worlds, and put through Bayesian inference over all possible worlds to get a distribution of model answers which are compared to human answers.

as using simulation as an engine of physics scene understanding (Battaglia et al., 2013). However, the questions asked in these experiments were often simple, with a the focus of the study on deriving the mechanisms of understanding. In this study, we examine more than just simple questions in simple scenes, but elevate the level of the questions such that they are difficult for humans to answer[1].

Lastly, modern approaches to the problem have been able to expand out from simple 2D scenes to complex 3D scenes (Xue et al., 2023). Other work done by Battaglia et al. (2013) pioneers the use of a single model for intuitive physics understanding across 5 different tasks and the work done by Zhang et al. (2023) creates a model which is able to process physical scene descriptions in natural language as input and output distributions of answers over a fixed question. In our study, we build off of similar ideas, but give a fixed scene representation and vary the underlying question.

Methods

We begin by presenting an overview of the experiment. Then we describe the stimuli used in the experiment and the process for human data acquisition. Following sections will then discuss our proposed model, NaLaPIP, as well as a baseline model derived directly for Large Language Model outputs. Lastly, we describe post processing applied to both NaLaPIP and baseline LLM data to prepare it for a better fit to human data.

Study Overview

We propose a natural language and intuitive physics task inspired by stimuli from Battaglia et al. (2013) and structure from Zhang et al. (2023), in which subjects were presented with visual scenes of various stable or unstable stacks of blocks and asked to predict the outcome of the scene after some time had passed. Our study expands on this work by allowing questions to vary flexibly in natural language, converting each question into a piece of probabilistic code and then running that code. Additionally, for simplicity, our scenes are 2-dimensional and we use a simplified approach to physics simulation.

Stimuli

Our stimuli are composed of a question in natural language and a 2D physical scene depicted through an image.

Each of the 2D physics scenes are composed of a series of blocks, colored red and purple, which are stacked on top of each other on a grey platform. The scenes are designed such that the blocks are stacked in a manner where it is not immediately obvious what will happen after some time has passed and gravity is allowed to act on the stack.[2] A full set of visual stimuli are included in the appendix section .

Each of the questions are designed to probe intuitive physics understanding of models. The base of the questions always query's what the physical state of the system will be (*will there be more than 3 boxes on the platform*, *will there be an even number of purple boxes on the platform*, etc.) and then is preface with the phrase *After some time has passed* to emphasise the probing of intuitive physics after the simulation has completed.

Questions were built using base units representing conceptual categories widely studied in both cognitive science and natural language processing contexts:

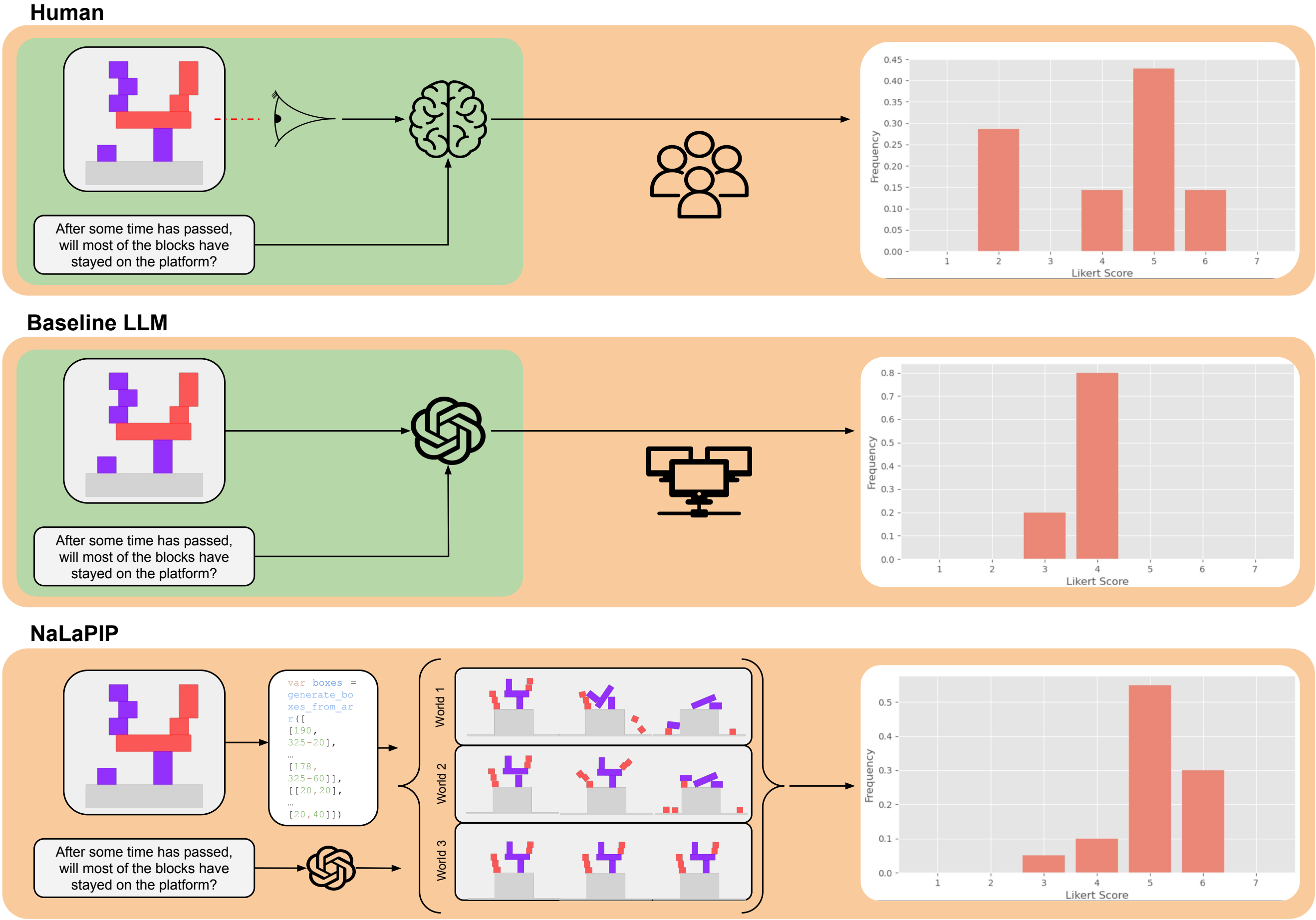
1. **Number:** Asking a question on the number of blocks, such as *three blocks* or asking a question about a very simple arithmetic property of the number of blocks, such as *divisible by 3* (Bartsch, 1973; Gelman & Gallistel, 1978).

2. **Spatial relations:** Asking a question on the block's position relative to other objects in the space, such as *to the left of the platform* (Landau & Jackendoff, 1993).

3. **Quantifiers:** Involving statements which imply reasoning about the number of blocks, but do no directly qiery what that number is, such as *some*, *most*, *few*, etc. (Montague, 1973; van Tiel, Franke, & Sauerland, 2021; Barwise & Cooper, 1981).

[1] Average human response time in the study we conduct is between 20 seconds and a minute

[2] To check this, we record the average response time from humans when they answer questions about the stack and verify that it is on the order of 10 seconds.


Human
After some time has passed, will most of the blocks have stayed on the platform?
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
1
2
3
4
5
6
7
Likert Score
Baseline LLM
After some time has passed, will most of the blocks have stayed on the platform?
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5
6
7
Likert Score
NaLaPIP
After some time has passed, will most of the blocks have stayed on the platform?
World 1
World 2
World 3
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
Likert Score

```
var boxes =
generate_bo
xes_from_ar
r([
[190,
325-20],
…
[178,
325-60]],
[[20,20],
…
[20,40]])
```

Figure 2: An outline of how the 3 agents involved in this study respond to stimuli. (Top) A human's visual perception system inputs a visual stimulus which is paired with a natural language stimulus to extract an answer. The answer is aggregated over multiple people to gain a distribution over results. (Middle) A multi-modal large language model is fed an image and natural language text and asked to make an inference of the result. Many calls are used to reconstruct the distribution of logit answers. (Bottom) NaLaPIP converts the inputs to probabilistic language of thought with help by a translating agent and a deterministic visual system. It then performs inference over all possible worlds to generate a distribution over results.

4. **Gradable adjectives:** Involving adjectives which exist on a scale, such as *tall*, *large*, etc. (Klein, 1980; Lassiter & Goodman, 2017).

We design 8 of these questions for training and then use a separate 6 for testing. The full set of questions used in this study is included in the appendix section .

For the stimuli used in the study, we pair each of the 8 visual stimuli with all 6 of the natural language stimuli, resulting in 48 total stimuli.

Human data acquisition

In total, 12 responses were gathered from 12 different participants using a online app tracking responses and response time. Participants were asked to rate their responses on a scale from 1 (*Definitely no*) to 7 (*Definitely yes*), known as a Likert (1932) scale. To insure the quality of the responses, each participant was first presented with a comprehension check before being presented with 24 different pairs of stimuli. More details on the application are included in the appendix section .

The NaLaPIP model

In this study, we create and evaluate NaLaPIP: a cognitive model which translates a visual stimuli and a natural language question into probabilistic language of thought (PLoT) (Goodman, Tenenbaum, & Gerstenberg, 2014) and produces an answer on a Likert (1932) scale. The model can be broken up into 3 different components: creating probabilistic generative models from visual stimuli, inserting probabilistic language of thought derived from natural language stimuli, and synthesizing an answer using a physics simulator. The probabilistic generative models are inspired by those proposed by Battaglia et al. (2013), the probalistic language of thought from natural language is based on the work of Goodman and Lassiter (2015b) on the use of probabilistic techniques in natural language pragmatics, and the fusion of these two techniques into one cohesive model is inspired from work done by Zhang et al. (2023).

Probabilistic generative models from visual stimuli: We create a probabilistic generative model which generates all possible worlds over a perception prior on the location and size of the blocks. Simple models of vision priors can derived from Gaussian distributions (Mamassian, Landy, & Maloney, 2002; Seriès & Seitz, 2013). Therefore, our prior over all possible worlds is based on Gaussian priors on the position and size of each of the blocks. These Gaussian priors are centered at the true value of their corresponding property. Additionally, the position priors are given width σ_{pos} and the size prior are given width σ_{size}. These widths remain as free parameters in the NaLaPIP model. The generative model is written in a probabilistic programming language (PPL) named WebPPL, which uses JavaScript based syntax to stochastically generate and sample from all possible worlds (Goodman & Stuhlmüller, 2014).

PLoT from natural language stimuli: In order to process the question in the context of our generative model, we need a translation function f where f : Natural Language $\rightarrow$ PLoT. As demonstrated by OpenAI's Codex, LLMs provide a model for approximating this function to a high degree (Chen et al., 2021). For this study, we use GPT-4, a more powerful successor of Codex. Using few-shot prompting (Brown et al., 2020), we give 8 translation examples and some other context and find that the LLM was able to correctly generate 48 out of 48 questions used in the final experiment with very little prompt engineering. More information on prompting can be found in the appendix section .

Answer synthesis with a physics simulator: Once the possible worlds have been generated, they are simulated using a WebPPL based physics model, Box2D (Catto, 2023). The default physics parameters are maintained and the only force which acts is gravity. After the simulation, we create the set of all possible worlds *after* gravity has acted, which we can query and sample from using the PLoT code generated in the previous section. Lastly, the results are interpreted on a Likert (1932) scale.

LLM Baseline

To establish a reference for our results, we create a baseline model. For this baseline model, we choose the mutlimodal version of the LLM used in the NaLaPIP model, GPT 4V. We prompt using a zero-shot prompt and with in-context examples to generate a few-shot prompt (Brown et al., 2020). The few-shot prompt is created with only one image, where the in-context examples are composed of 8 questions from the training set and corresponding answers retrieved from the mean of the human data distribution. More information on the prompts for the baseline can be found in the appendix section . Once we insert the natural language question and image stimulus properly into the prompt, we feed this through the LLM to get a result. In order to achieve a distribution, we pass through the LLM multiple times with *tempaure*=t and combine all the results. The temperature, t, is a free parameter if the LLM model and can be tuned accordingly.

Free Parameter	Description
σ_{pos}	Uncertainty from perception on position
σ_{size}	Uncertainty from perception on size
a	Sharpness of logistic function
b	Model data offset of the logistic function
c	Scale of logistic function
d	Human data offset of the logistic function
t	Variability in baseline LLM model answers

Table 1: Summary of all free parameters used in the experiment and their interpretation. All parameters are free in NaLaPIP model except t and only t, a, b, c, and d are free in the LLM baseline model.

Post Processing

After distributions have been obtained from NaLaPIP and the Baseline LLMs, they are transformed through a logistic function in order to help transform the probability space outputted by the models to a human scale. Additionally, this logistic function accounts for non-linearity in human judgement (Kim, Yang, & Kim, 2008). We parameterize the logistic function by (a,b,c,d) such that if we have initial model rating of r, our post-process model rating r' is given by Equation 1.

$$r' = \frac{c}{1+e^{-(ax-b)}} + d \tag{1}$$

We add free variables (a,b,c,d) to the models which will, along with the other free parameters, be fit to human data. A full summary of free parameters of the study can be seen in Table 1.

Results

In this study, we evaluate the quality of our cognitive model, NaLaPIP, by comparing it to human responses. For reference, we also compare the output of bleeding edge mutlimodal generative model, GPT-4V, to human responses. We evaluate the comparison to human responses by (1) comparing the distance between the average human response and the average model response over all possible stimuli and (2) comparing the distance between the human and model distributions using the Wasserstein metric. In evaluating these metrics, we obtain two main results[3]:

1. NaLaPIP is a **significant improvement** over state-of-the-art mutlimodal AI models.

[3] Note, all results shown are with tuned free parameters. The parameters found for NaLaPIP are $\sigma_{\text{pos}} = 3, \sigma_{\text{size}} = 1, a = 0.79, b = 3.978, c = 6.00, d = 1.00$. The free parameters found for GPT-4V are surprisingly close at $a = 0.78, b = 3.975, c = 6.00, d = 1.00, t = 1.5$

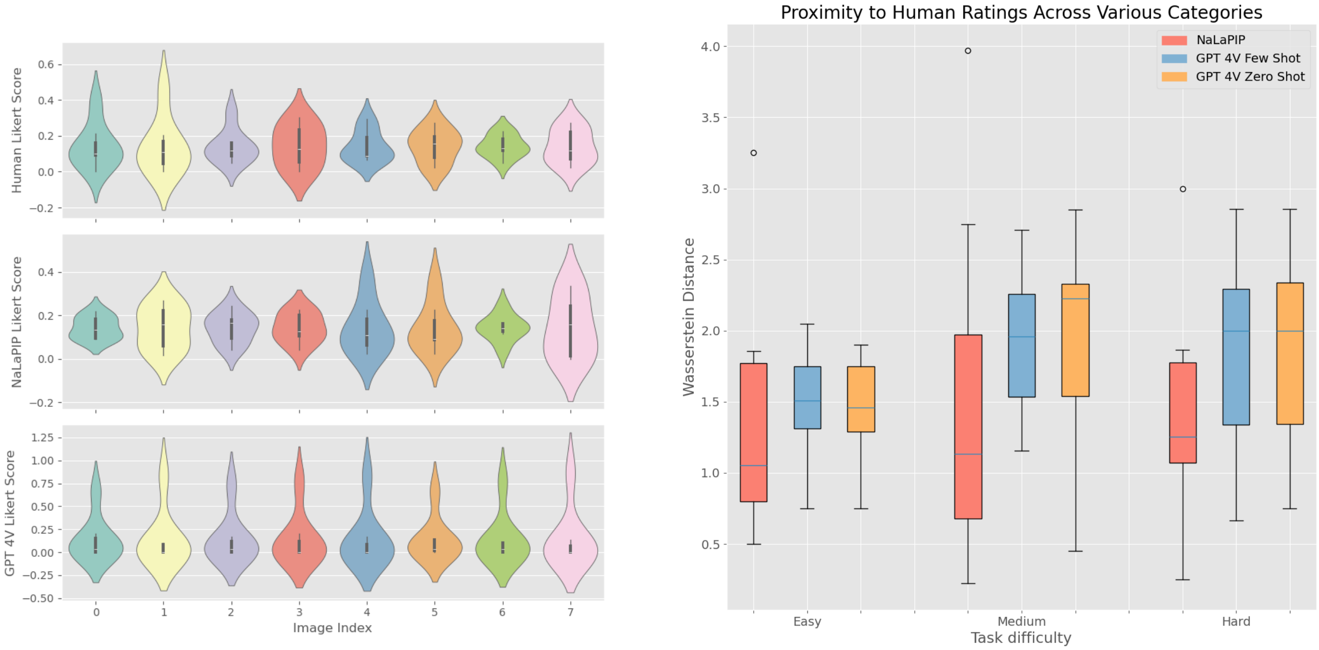


Figure 3: (Left) Violin plot display distribution over human answers (top) NaLaPIP answers (middle) and LLM baseline answers (bottom). Each violin collumn corresponds to a different image stimulus and is averaged over all questions asked about that image stimulus. (Right) Wasserstein distance when compared to human ratings across various difficulties of tasks and various models. Lower Wasserstein distance is interpreted as closer to human ratings.

2. Whil there is a **discrepancy** between exact human and NaLaPIP responses, we can identify a **strong positive correlation** between the two samples.

The section on the significant improvement of NaLaPIP focuses on comparing the Wasserstein metric derived from NaLaPIP and GPT-4V responses and outlines the statistical significance of this comparison. The section on the discrepancy between human and NaLaPIP responses focuses on comparing the mean of responses from NaLaPIP and GPT-4V directly as well as their correlation.

The improvement of NaLaPIP over GPT 4V:

We observe significant improve of our model, NaLaPIP, over a baseline state-of-the-art multimodel transformer, GPT-4V. Qualitatively, this is best seen in Figure , which shows a comparison between distributions of likert scores aggregated over various images[4] [5]. From this plot, one can qualitatively see that GPT-4V is not able to capture differences between images, while NaLaPIP does a much better job matching the general shape of the human distribution. This provides evidence that compared to the baseline, NaLaPIP is able to much better account for physical scene understanding.

To assess this qualitative trend qualitatively, we introduce the Wasserstein metric, which serves as a distance metric between two distributions. The key property of this metric is that distributions which are more similar to one another have lower Wasserstein metrics and the Wasserstein metric is 0 if and only if the two distributions are the same. We compute 3 Wasserstein metrics for each of the 48 tasks[6] to create a quantitative measurement of the difference between human/NaLaPIP, human/few-shot GPT-4V, and human/zero-shot GPT-4V response distributions. We show the distribution of the Wasserstein metric for each task, grouped by task difficulty, in Figure . We see that across all levels of task difficulty, NaLaPIP shows significant improvement over GPT 4V Few Shot and GPT 4V Zero Shot across all task difficulties.

To quantify this observation, we perform a T-test of significance on the Wasserstein scores of NaLaPIP compared to baselines. Due to a low number of tasks, this statistical test is done on all the tasks simultaneously. While the distribution over Wasserstein metrics is not normal, the difference between the Wasserstein scores of NaLaPIP and GPT-4V Few Shot is approximately normal[7]. Therefore, we can perform a 1-tailed T-test of differences with null hypothesis *there is no difference between Wasserstein metrics of GPT-4V Few Shot and NaLaPIP*. The result of this test gives a value

[4] The mapping between image index and image can be found in the appendix section

[5] The y-axis labels are not shown on the violin plot is not shown because the distributions are shown after logistic re-scaling, which transforms the axis. Due to the monotonicity of the transformation, higher likert scores are still shown as higher on the axis.

[6] A task is a question/image pair. 48 tasks are created from 8 images and 6 questions.

[7] See appendix section for distributions over Wasserstein metrics

of $p = 0.1\%$, showing strong evidence that NaLaPIP is a significant improvement over the state-of-the-art mutlimodel AI baseline in modeling human responses to our stimuli.

Correlation between NaLaPIP and human responses.

While we are able to use the Wasserstein metric to make claims on the significant improvement of NaLaPIP compared to GPT-4V Few Shot, the irregularity of the distribution obtained on the Wasserstein metric when comparing human to NaLaPIP responses makes it difficult to make conclusions using the metric. Therefore, we propose an analysis on the means of the likert scores obtained from humans and from the NaLaPIP model, asserting that a proximity of means suffices to demonstrate accurate modeling of human responses.

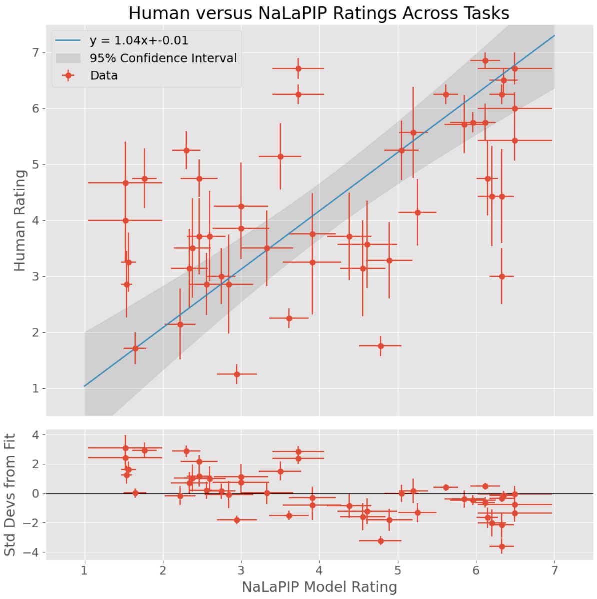


Figure 4: Direct comparison of mean NaLaPIP ratings versus mean human rating (top) and their residuals to line of best fit (bottom). Line of best fit is depicted with a 95% confidence interval assuming data is linear. Confidence interval and residual signify a non-linear relationship.

Figure 4 demonstrates the relationship between the mean of the likert distribution of human and NaLaPIP responses over all 48 tasks. We begin analysis of this relationship by assuming a positive linear relationship between mean NaLaPIP likert ratings and mean human likert ratings, performing a linear fit, and propagating the errors of the fit along with the errors on each datapoint to establish a 95% confidence interval. While the best fit demonstrates near equivalence ($y = 1.04x - 0.01$), we see that far fewer than 95% of the data lies inside this confidence interval. Therefore, we assert that the assumption of a direct linear relationship does not hold for our data. This conclusion is qualitatively echoed by the relationship between residuals and mean NaLaPIP model rating, as displayed in Figure 4.

Instead, we demonstrate that there is a distinct positive correlation between mean NaLaPIP likert scores and mean human likert scores by calculating the R^2 metric and it's resultant p-value. We find the R^2 statistic to be 0.58 which, when combined with the uncertainty on each data point, results in significance of $p = 0.1\%$ that there is no correlation. This leads us to conclude with high confidence that there is a positive correlation between NaLaPIP and human responses[8].

Discussion

With high statistical significance, we are able to assert that there is a significant improvement of NaLaPIP over state-of-the-art multi-modal LLM baselines and that this is a positive relationship between mean human response and mean likert responses. With the improvement over multi-modal LLM baselines, we propose that people's physical scene understanding is explained significantly better by Bayesian inference over simulations on all possible worlds using a perception prior than it is by more general purpose machine learning techniques used to build today's transformers. Furthermore, we show that the framework demonstrated in this study suffices to create a model which correlates strongly with human data over a range of questions and scenes. While this study has a relatively small stimuli sample size, we use the robustness of hyper-parameter adjustments as evidence that the techniques can be applied to a much larger and more diverse set of of stimuli than demonstrated in this study. Lastly, our strong correlations show evidence that intuitive physics models can reflect human behavior by translating natural language questions into probabilistic language of thought sentences which probe Bayesian generative models.

Our analysis has several limitations. Firstly, we present relatively noisy data over only 48 different tasks. The noise is due a relatively low amount of human participation causing propagation of a large amount on statistical uncertainty throughout the analysis. This lack of human participation also causes the constraint to only 48 tasks, as any more tasks would contribute to a smaller sample size per task.

Another limitation of our model is the limitation of the prior created over all possible worlds and the simplicity of the Box2D engine used. Work in computational cognitive science has demonstrated several more complex methods for establishing perception and physics priors over intuitive physics models which out-perform the method used in this study (K. Smith et al., 2024). Furthermore, these models exist in 3D, which would allow the study to be more linked to real-world intuitive physics tasks (Xue et al., 2023). Additionally, this study could be expanded by removing the mapping from a visual image to code which rendered that image through use of inverse graphics methods to construct scene representations to feed into NaLaPIP (Yi et al., 2018).

[8] For comparison, the best LLM baseline result gave a result of $R^2 = 0.114$ with $p = 43\%$ significance that there is a correlation. Corresponding plots to this analysis can be found in the appendix section

More broadly, we acknowledge the limitations of this study which result weak conclusions about NaLaPIP's ability to exhibit human-like properties. However, we hope this framework will inspire further investigation into probing intuitive physics with natural language.

Acknowledgments

I'd like to thank the teaching team of MIT's 9.66 class in the fall of 2023 for their work and dedication to an amazing course in computation cognitive science. I'd also like to specifically thank Josh Tennebaum for his role in inspiring this study with his lectures, office hours advice, and published work. Lastly, I'd like to thank Cedegao Zhang, who's work inspired this study and who gave advice on the framework of the analysis.

Author Contribution:

This research paper is the result of the sole efforts and intellectual contribution of Julius Heitkoetter, the sole author of this work. The entirety of the project, including the conception and design of the study, data collection and analysis, interpretation of data, and the writing of the manuscript, was carried out by Julius Heitkoetter with help and guidance from MIT's 9.66 course staff. There were no additional contributors or collaborators involved in any stage of this research project.pro

References

Bartsch, R. (1973). The semantics and syntax of number and numbers. In *Syntax and semantics volume 2* (pp. 51–93). Brill.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language* (pp. 241–301). Springer.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Catto, E. (2023). *Box2d: a 2d physics engine for games.* http://box2d.org.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., … Zaremba, W. (2021). *Evaluating large language models trained on code.*

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *ArXiv*, *abs/2205.05718*. Retrieved from https://api.semanticscholar.org/CorpusID:248721753

Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, *10*.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Gelman, R., & Gallistel, C. (1978). Thechildsunderstanding of number. *Cambridge, Ma.: Harvard University*.

Goodman, N. D., & Lassiter, D. (2015a). Probabilistic semantics and pragmatics uncertainty in language and thought. *The handbook of contemporary semantic theory*, 655–686.

Goodman, N. D., & Lassiter, D. (2015b). Probabilistic semantics and pragmatics uncertainty in language and thought. In *The handbook of contemporary semantic theory* (p. 655-686). John Wiley Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118882139.ch21 doi: https://doi.org/10.1002/9781118882139.ch21

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. (Accessed: 2023-12-17)

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). *Concepts in a probabilistic language of thought* (Tech. Rep.). Center for Brains, Minds and Machines (CBMM).

Jackendoff, R. S. (1985). *Semantics and cognition* (Vol. 8). MIT press.

Kim, C. N., Yang, K. H., & Kim, J. (2008). Human decision-making behavior and modeling effects. *Decision Support Systems*, *45*(3), 517-527. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923607000966 (Special Issue Clusters) doi: https://doi.org/10.1016/j.dss.2007.06.011

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, *4*, 1–45.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, *16*(2), 255–265.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*, 3801–3836.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. *Probabilistic models of the brain: Perception and neural function*, 13–36.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 stanford workshop on grammar and semantics* (pp. 221–242).

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in human neuroscience*, *7*, 668.

Smith, K., Hamrick, J., Sanborn, A. N., Battaglia, P., Gerstenberg, T., Ullman, T., & Tenenbaum, J. (2024). Intuitive physics as probabilistic inference.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Van Eijck, J., & Lappin, S. (2012). Probabilistic semantics for natural language. *Logic and interactive rationality (LIRA)*, *2*, 17–35.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9), e2005453118.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). *From word models to world models: Translating from natural language to the probabilistic language of thought.*

Xue, H., Torralba, A., Tenenbaum, J., Yamins, D., Li, Y., & Tung, H.-Y. (2023). 3d-intphys: Towards more generalized 3d-grounded visual intuitive physics under challenging scenes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3624–3634).

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, *31*.

Zhang, C., Wong, L., Grand, G., & Tenenbaum, J. (2023). Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Application Given to Humans During Study

The online survey application built with jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023).

The instructions were as follows:

Welcome! We are conducting an experiment about how people evaluate intuitive physics situations when asked a question in plain english. Your answers will be used to inform computer science and cognitive science research. This experiment should take no more than 10 minutes

In this experiment, you will be presented with a series of images containing =stacks of blocks and some simple questions. The block are stacked on a grey platform in a configuration which may or may not be stable. In each situation, you will be asked to imagine what will occur after a sufficiently large amount of time has passed and gravity has finished acting on the stacks of blocks. , Your task will be to think about how these blocks will fall and their position after they have finished falling=. Note that sometimes, you may be presented with a stable stack where no blocks will fall. You will enter your answer for each question by clicking a rating on a ¡strong¿7-point multiple choice scale¡/strong¿ ranging from 1 (definitely no) to 7 (definitely yes). ,

You will see descriptions of ¡strong¿24 different scenarios¡/strong¿ in total.

The comprehension check was as follows:

Check your knowledge before you begin. If you don't know the answers, don't worry; we will show you the instructions again.

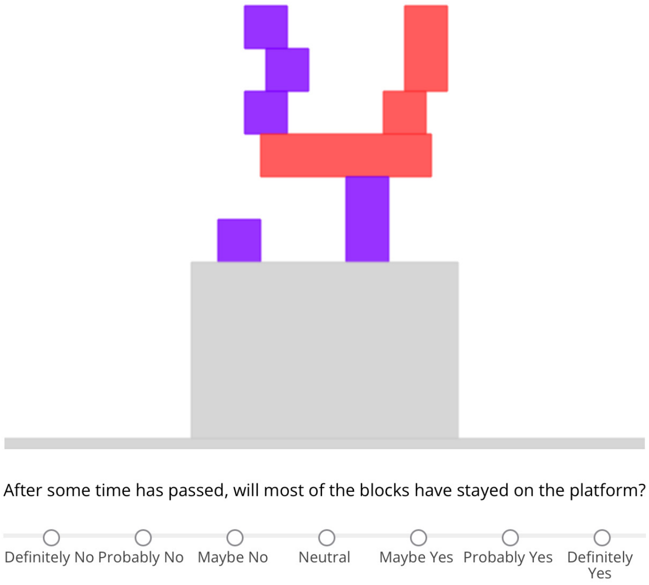
What will you be asked to do in this task?

1. Look at images containing blocks and answer questions about what is currently occurring in the image.
2. Look at images containing blocks and answer questions about what will happen in the image after gravity has acted on the blocks.
3. Look at images containing blocks and answer questions about what will happen in the image while gravity is acting on the blocks.

How will you be providing your answer?

1. By writing text.
2. By selecting an option from a multiple choice scale.
3. By moving a slider.

An example presentation of the stimulus was as follows:



After some time has passed, will most of the blocks have stayed on the platform?

○ Definitely No ○ Probably No ○ Maybe No ○ Neutral ○ Maybe Yes ○ Probably Yes ○ Definitely Yes

Prompts used by NaLaPIP to give to LLM

Below is the prompt that was fed into NaLaPIP's codex to generate sentences in WebPPL (a probalistic programming language) from sentences in natural language

> You will output code to a probabilistic programming language built in java script. The code you output will be a conditional for a world that starts with 8 purple and blue boxes, stacked on top of each other, on a platform. Please only output a single line of code, nothing else.
>
> Below are the following are the predicates you may use. You can only use these predicates, do not invent other predicates.
>
> var isOnPlatform = function(obj) return obj.y ¡ 325;
>
> var isNotOnPlatform = function(obj) return !(obj.y ¡ 325);
>
> var isLeftOfPlatform = function(obj) return obj.y ¿ 325 and obj.x ¡ 175;
>
> var isNotLeftOfPlatform = function(obj) return !(obj.y ¿ 325 and obj.x ¡ 175);
>
> var isRightOfPlatform = function(obj) return obj.y ¿ 325 and obj.x ¿ 425;
>
> var isNotRightOfPlatform = function(obj) return !(obj.y ¿ 325 and obj.x ¿ 425);
>
> var isPurple = function(obj) return obj.color=="purple";
>
> var isRed = function(obj) return obj.color=="red";
>
> var isLarge = function(obj) return obj.dims[0] ¿ 25 —— obj.dims[1] ¿ 25;
>
> var isNotLarge = function(obj) return !(obj.dims[0] ¿ 25 —— obj.dims[1] ¿ 25);
>
> Below are some examples:
>
> English: After some time has passed, will some of the purple blocks have fallen off the platform.
>
> Code: filter(isNotOnPlatform, filter(isPurple, finalBoxes)).length ¿ 0
>
> ... [MORE EXAMPLES HERE] ...
>
> English: After some time has passed, will there be more red boxes to the left of the platform than purple boxes on the platform?
>
> Code:

Prompts used by Baseline to give to LLM

Below are the prompts used to query the GPT-4V baseline LLM

.

Zero Shot Prompt

Look at the 2-D physics scenario in the image. It portrays a set of blocks (purple and red) sitting on a grey platform.

The blocks may be stacked stably or unstably. For a sufficiently long amount of time, gravity is allowed to act on the blocks.

Your job is to answer the following question on a 1-7 scale, with 1 being Definitely No and 7 being Definitely Yes.

Only answer ”1” or ”7” if you are extremely sure about the outcome of the question, as it pertains to the image.

Question as it pertains to image:

.

Few Shot

Look at the 2-D physics scenario in the image. It portrays a set of blocks (purple and red) sitting on a grey platform.

The blocks may be stacked stably or unstably. For a sufficiently long amount of time, gravity is allowed to act on the blocks.

Your job is to answer the following question on a 1-7 scale, with 1 being Definitely No and 7 being Definitely Yes.

Only answer ”1” or ”7” if you are extremely sure about the outcome of the question, as it pertains to the image.

Question as it pertains to image: ”After some time has passed, will there be any red blocks remaining on the platform?”

Rating: 5

... [MORE EXAMPLES HERE] ...

Question as it pertains to image:

Rating:

Natural Language Stimuli Used in Study

Below are the natural language stimuli used in the study: .

Testing Questions

1. After some time has passed, will there be any red blocks remaining on the platform?
2. After some time has passed, will there be more red blocks to the left of the platform that purple blocks to the right of the platform?
3. After some time has passed, will most of the blocks have stayed on the platform?
4. After some time has passed, will there be any purple blocks to the right of the platform?
5. After some time has passed, will the number of blocks on the platform be divisible by 3?
6. After some time has passed, will the majority of the large boxes be on the platform?

.

Training Questions

1. After some time has passed, will there be more than 3 boxes on the platform?
2. After some time has passed, wil there be any boxes to the left of the platform?
3. After some time has passed, will there be more than 2 red boxes on the platform?
4. After some time has passed, will there be more red boxes to the left of the platform than purple boxes?
5. After some time has passed, will there be any large boxes to the left of the platform?
6. After some time has passed, will most of the boxes have fallen off of the platform?
7. After some time has passed, will there be an even number of purple boxes on the platform?
8. After some time has passed, will some of the purple blocks have fallen off the platform.

Visual Stimuli Used in Study

Below are the visual stimuli used in the experiment. Each caption depicts the corresponding stimulus index which the ensemble of boxes is mapped to.

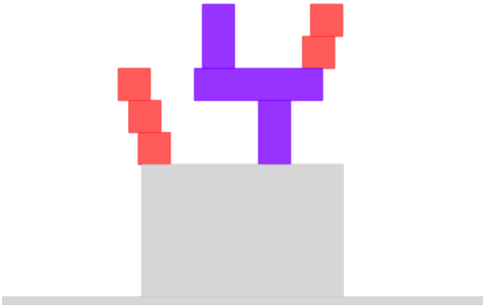


Figure 5: Visual Stimuli 00

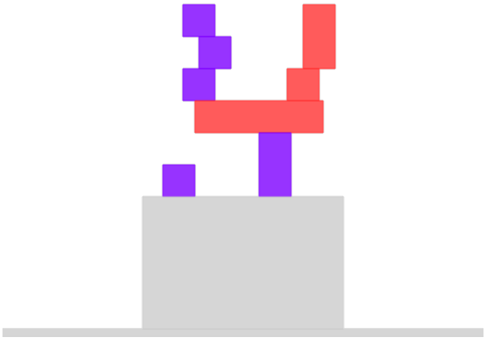


Figure 6: Visual Stimuli 01

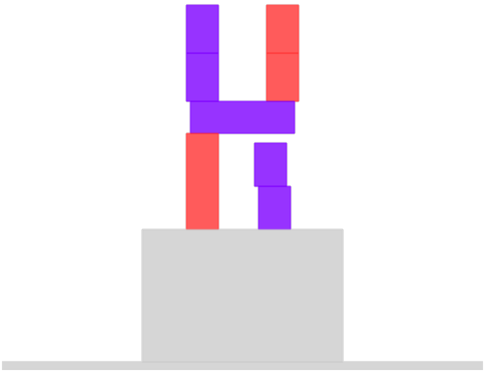


Figure 7: Visual Stimuli 02

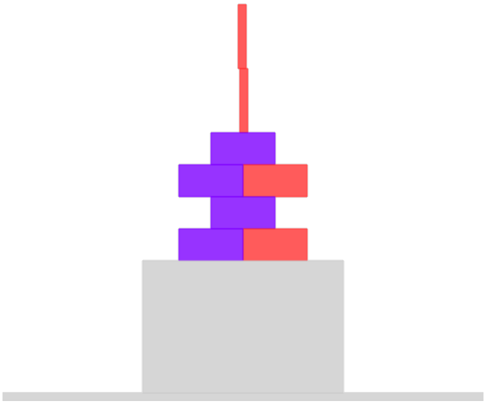


Figure 8: Visual Stimuli 03

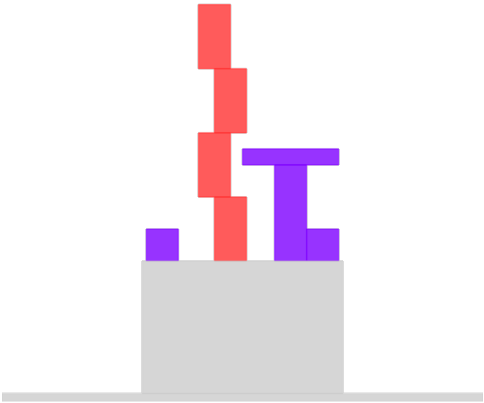


Figure 9: Visual Stimuli 04

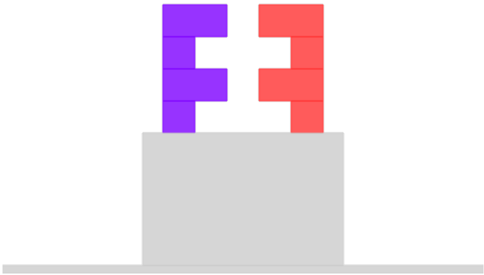


Figure 10: Visual Stimuli 05

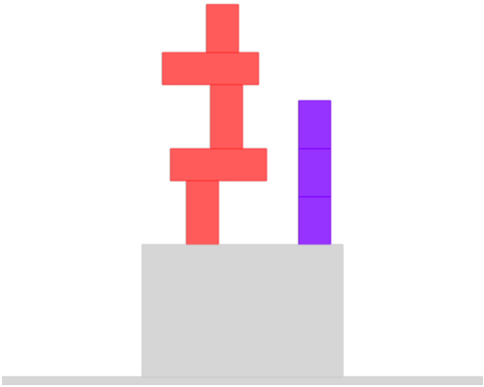


Figure 11: Visual Stimuli 06

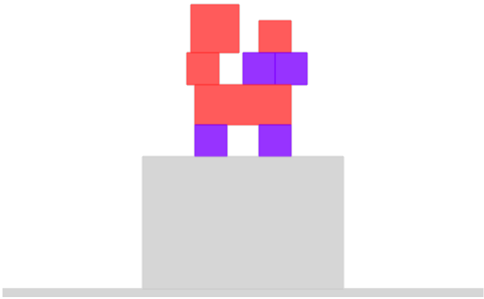


Figure 12: Visual Stimuli 07

Supplemental Plots

Below are supplemental plots mensioned in the study but not integral to its results.

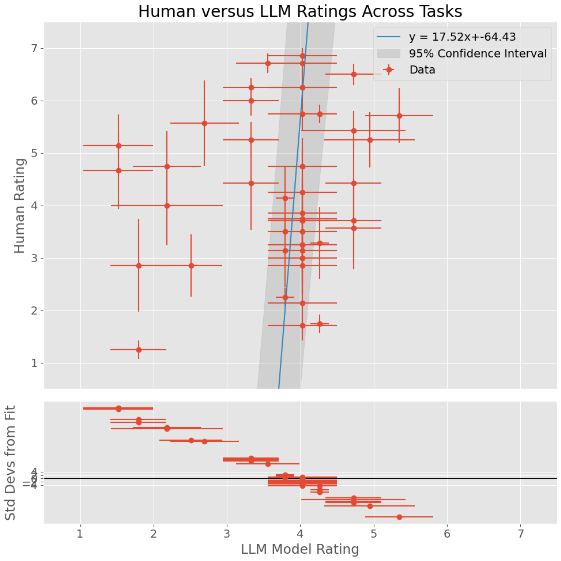


Figure 13: Direct comparison of mean GPT-4V Few-Shot ratings versus mean human rating (top) and their residuals to line of best fit (bottom). Line of best fit is depicted with a 95% confidence interval assuming data is linear. Confidence interval and residual signify a non-linear and uncorrelated relationship.

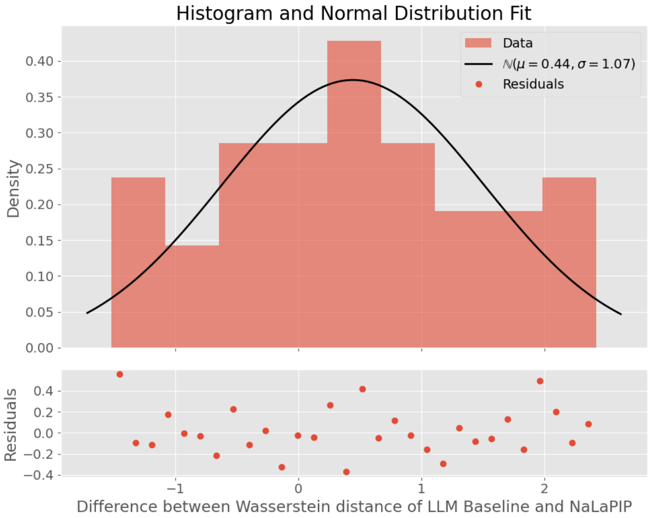


Figure 14: Histogram of task-by-task difference between Wasserstein metric of LLM GPT-4V Few Shot and NaLaPIP when compared with human ratings as well as normal distribution fit. Notice the randomness amoung the residuals signifying an approximately normal distribution.

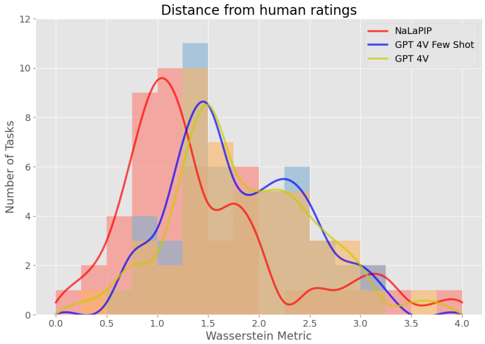


Figure 15: Plot of histograms and spline fits of Wasserstein distance for NaLaPIP (red), GPT-4 Few-Shot (yellow), and GPT Zero-Shot (blue) when they are compared to human data. NaLaPIP shows the best result, with GPT4V Few Shot only slightly outperforming GPT-4V Zero Shot. Note the extremely non-gaussian functional shape.