Large Language Models Are Challenged by Habitat-Centered Reasoning

Sadaf Ghaffari and **Nikhil Krishnaswamy**
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science
Colorado State University
Fort Collins, CO USA
{sadafgh,nkrishna}@colostate.edu

Abstract

In this paper we perform a novel in-depth evaluation of text-only and multimodal LLMs' abilities to reason about object *habitats* or conditions on how objects are situated in their environments that affect the types of behaviors (or *affordances*) that can be enacted upon them. We present a novel curated multimodal dataset of questions about object habitats and affordances, which are formally grounded in the underlying lexical semantics literature, with multiple images from various sources that depict the scenario described in the question. We evaluate 16 text-only and multimodal LLMs in a zero-shot manner on this challenging data. Our findings indicate that while certain LLMs can perform reasonably well on reasoning about affordances, there appears to be a consistent low upper bound on habitat-centered reasoning performance. We discuss how the formal semantics of habitats in fact predicts this behavior and propose this as a challenge to the community.

1 Introduction

Discussion of Large Language Models (LLMs) in both research and public media often gives the impression that they are capable of much more human-like reasoning than in they really are (Shanahan, 2024). This makes it even more important to rigorously examine the limitations of LLMs.

When it comes to multimodal large language models (MLLMs), particularly in the domain of integration with robotic systems, significant attention has been paid to *affordance* recognition and reasoning (Huang et al., 2024), particularly as it pertains to the ability to reason about changes enacted over objects in a scene. However, toward goals such as LLM-enabled robots, there remain many open problems left to be solved, from migrating out of tabletop scenarios to navigating dynamic and changing environments. One such under-addressed area is *habitats* (Pustejovsky, 2013). These are locally-scoped environments that contextualize an object and condition the behaviors it can participate in. They are closely related to *affordances* (Gibson, 1977), in that habitats enable or disable certain behaviors that may be executed with an object (Pustejovsky and Krishnaswamy, 2016). For example, in order to screw in a screw with a screwdriver, the tip of the screwdriver must be inserted into the head of the screw.

In this paper, we thoroughly evaluate the habitat and affordance reasoning capabilities of **16** open-weight text-only and multimodal LLMs. We find that performance on affordance-centering questions frequently outstrips performance on habitat-centering questions, although all models make significant errors on multiple classes of problems. We also find that while image information provided to MLLMs can help performance, they are highly sensitive to perturbations in the image input; often, image information actually hurts performance, indicating weaknesses in the reasoning capabilities of MLLMs.

Our findings indicate how habitats, which are typically implicitly indicated in linguistic descriptions, are more challenging for LLMs and MLLMs to reason about than affordances, which are typically selected for by the matrix verb of sentences involving object-action descriptions. More generally, we show how LLMs' reasoning is susceptible to strong biases toward typical or "canonical" object orientations, because these are the collocations that typically occur in free text and common image datasets, even if they do not reflect the reality of *in situ* object reasoning problems. Our evaluation data is collated into **HabitatQA**, a novel reasoning and question answering dataset focused on the habitats and affordances of common objects, with accompanying images of the situations described, from various sources, including crowd-sourcing, simulation, and generative text-to-image models. Our data and code is available at https://github.

com/csu-signal/Habitat-Reasoning.

2 Related Work

Pustejovsky (2013) motivates habitats by way of *event localization* within a minimal model that distinguishes the temporal traces made by certain linguistic inputs. Assuming certain agent-oriented cognitive constraints, such as an evidential point of view (POV), event localization requires constructing sets of contextual factors $\mathcal{C}$ for an item x, without which properties of x cannot be distinguished within the minimal model.

Pustejovsky and Krishnaswamy (2016, 2022) formally denote the relation between habitats and affordances as $\mathcal{C} \rightarrow [\pi]\mathcal{R}$, in terms of a conditioning environment $\mathcal{C}$, program π, and result R, such that if an object is in configuration $\mathcal{C}$, every time the afforded behavior program π is executed, result $\mathcal{R}$ will occur. $\mathcal{C}$, the habitat, may implicate object orientations, size constraints, or relations to other objects. Habitats are also categorized into *intrinsic* (e.g., a television has an intrinsic front) and *extrinsic* (e.g., to be rolled, a glass must be placed in an extrinsic horizontal orientation). These properties of objects are either selected for by their positioning in the environment, or inherent to the object regardless of how they are positioned. Relatedly, Barbu et al. (2019) identified how non-canonical object orientations in images (a proxy for habitats) challenge object classification models that are overwhelmingly trained on canonical or "stereotypical" object positionings in datasets like ImageNet. Despite this insight, such biases appear to have persisted in model multimodal LLMs, largely because their data requirements mean they are trained over "Internet-scale data" (Zitkovich et al., 2023) in which canonical object positionings remain overwhelmingly prevalent. As a result, recent investigations into LLMs and VLMs show that while certain models can perform reasonably well on correctly identifying object affordances (Jones et al., 2022; Ghaffari and Krishnaswamy, 2023; Qian et al., 2024), habitats remain challenging (Jones and Trott, 2024; Ghaffari and Krishnaswamy, 2024).

There is a wealth of datasets and LLM evaluations that address areas related to but non-overlapping with habitats. A sampling are discussed below.

The SPACE dataset (Duan et al., 2021) consists of synthetic video in a 3D environment, of containment, stability, and contact events. PHYSOBJECTS (Gao et al., 2023) is a large dataset centered on common household objects automatically annotated with physical concepts that capture human priors from the physical appearance. The HANDAL dataset (Guo et al., 2023) consists of images annotated with 6-DoF category-level pose and scale for robotic manipulation, annotated for affordances as manipulable parts of objects. The Physiclear dataset (Yu et al., 2024) contains both physical/object property reasoning tasks and annotated tactile videos obtained using a GelSight tactile sensor. Kondo et al. (2023) presented a dataset to investigate the ability of language models to predict size relationships between objects, which is a component of the habitat concept but does not cover it completely.

Kembhavi et al. (2017) focused on Multi-Modal Machine Comprehension (M3C) tasks given text, diagrams, and images. Bisk et al. (2020) introduced PIQA, a multiple choice question-answering dataset for physical interaction wherein the model must choose the best solution for a physical goal expressed in natural language. Aroca-Ouellette et al. (2021) used multiple choice questions to cover 10 basic concepts: direction, mass, height, circumference, and various object affordances. Hong et al. (2021) showed visual reasoning models underperform humans on part-based conceptual, relational, and physical reasoning. Krishnaswamy and Pustejovsky (2022) showed that embeddings trained over representations of affordances can be used to analogize between similar objects, such as in finding a substitute tool for a task. Similarly, Tian et al. (2023) explored the creative problem solving capabilities of LLMs, with an emphasis on unconventional tool usage and how objects can be combined to achieve a complex goal.

Collins et al. (2022)'s benchmark compares humans and distributional LLMs in planning and explanation generation. Wang et al. (2023) benchmark multimodal and text-only LLMs on physical attributes such as malleability, elasticity, and stiffness, using multiple choice questions. Yiu et al. (2023) perform an assessment of LLM problem solving and find them inferior to human children when it comes to choosing objects for a task based on their affordances. Zheng et al. (2024) assessed machine common sense reasoning in MLLMs and non-generative models over soft bodies and fluids using video question answering, wherein the model must have a deep understanding of physical scenes and their dynamics to succeed. Ser-

manet et al. (2024)'s RoboVQA is a large dataset for robotics-oriented pick-and-place tasks. Majumdar et al. (2024) present OpenEQA, for embodied question answering in an open environment, focused on locational and action questions. Williams and Huckle (2024) present a benchmark of "easy problems that LLMs get wrong," including spatial reasoning. They astutely observe that the prevalence and proliferation of large scale benchmarks encourages optimization toward the benchmark, rather than a focus on holistic evaluation. In the spirit of holistic evaluation, we note that their spatial reasoning benchmark does not focus on habitats and affordances, as ours does.

The above works indicate how many previous approaches address problems that are adjacent to habitat-based reasoning in language models. However, there remains a gap in directly evaluating the problem of habitats themselves, as we do here, on a wide variety of models.

3 Data Collection

Our data, termed HabitatQA, consists of multiple-choice questions, each with associated images that depict the scene or scenario described in the question. There are a total of 210 questions and 617 image/question pair samples (consisting of a question paired with an associated image, on which multimodal models are evaluated). Each question has at least 2 associated images, and some images may be associated with more than one question, because they appropriately depict multiple scenarios described. The data was collected/constructed using the methods below.

3.1 Question Construction

We began by creating a set of multiple-choice affordance-centering and habitat-centering questions. An "affordance-centering" question is considered to be one that directly addresses the behaviors that a specific object can participate in, while a "habitat-centering" question asks about configurations that may be required to execute a given behavior or created by executing one. Examples of each are given below (correct answers bolded).

AFFORDANCE-CENTERING QUESTION

Which one of the following objects can contain something?
a) solid rectangular prism
b) **cardboard box**

HABITAT-CENTERING QUESTION

A knife is inserted into a glass. What is the likely direction in which the handle points?
a) knife handle touches the bottom of the glass
b) **knife handle faces towards the opening of the glass**

The questions are intentionally simple, but also require selecting for specific properties of the object to answer correctly, such as, in these examples, the containing nature of a cardboard box or a glass, or a knife's inherent directedness and mereotopological relations between it and a container.

We constructed an initial set of questions which were validated by a subject matter expert in habitats and affordances for the types of object properties addressed and to minimize ambiguity in the interpretation of the questions. This initial set was then given to ChatGPT and Claude 3 to expand the question set by replacing objects in the initial questions with objects that have similar physical properties but different canonical uses (for example, replacing "mug" with "ramekin"). These substitutions then underwent a further human-in-the-loop correction and validation step. This process resulted in 116 multiple-choice questions about object habitats (concerning 54 different objects as either part of the question or answer options) and 94 about affordances (concerning 102 different objects as either part of the question or answer options). Each question had between 2 and 6 answer options, and required reasoning about object properties such as concavity, size, rotation, direction of orientation, contact surfaces, as well as causal factors[1], to answer correctly.

All questions were answered by two annotators. Each double checked their work after completion. We computed inter-annotator agreement and arrived at a kappa score of 1.0, meaning full agreement, resulting in the gold standard. This indicates not only how the generated questions are easy for humans to answer correctly, but also how humans have strongly concurrent notions of spatial relations and afforded behaviors for everyday objects.

3.2 Image Collection

We follow the intuitions that linguistic input alone does not allow a language model to truly ground its reasoning to anything external to the language, as

[1]Causal factors being primarily concerned with what the resultant state or configuration would be if a specified action were enacted over the object.

Figure 1: Example image corresponding to the example habitat-centering question given above.

humans do (Bender and Koller, 2020), and that a single underspecified linguistic description may describe any number of real-world situations (Krishnaswamy, 2017), prompting the addition of images to our data. This allows an evaluation of habitat- and affordance-based reasoning in text-only and multimodal language models, to evaluate the contribution of visual information and additional reasoning capability (i.e. visual reasoning) in multimodal models.

Natural images We conducted a crowd-sourcing of images corresponding to the generated questions. Image collectors were asked to take pictures of the scene described in the question, including all the objects mentioned in either the question or the answer options. For instance, an image corresponding to the example affordance-centering question above would include both a solid rectangular prism and a cardboard box, while one corresponding to the example habitat-centering question would depict a knife inserted into a glass (Fig. 1). Collectors were asked to take multiple pictures of the situation, including from different angles or on different backgrounds. Following Barbu et al. (2019), we asked collectors to include cluttered backgrounds and non-traditional lighting or angles. Images were collected with the assistance of an app that ingested a spreadsheet with the questions and prompted collectors with the objects mentioned in the question, so they could gather the objects and construct the scene before taking pictures. A total of 7 people participated in image collection, resulting in a total of 478 (129 affordance and 349 habitat) natural images. All images were resized to 1,000p resolution and converted to PNG format.

Figure 2: Example image generated with Stable Diffusion.

Generated images Some questions presented scenarios that were infeasible to gather images for in an everyday context (e.g., due to lack of access to certain objects, like a *decanter*, or described counterfactual situations).

Therefore, we turned to generative AI, specifically generative text-to-image models, to generate images representative of the scenarios described in our questions, with a goal of at least 2 images per question.[2] We used Stable Diffusion 2.1 (Rombach et al., 2022), and prompted the model with the scene described in the question, with augmentation to make explicit things that may be implicit in the question (e.g., "There are two objects on the table: One cone and one glass sitting *next to each other*."). The prompt was run multiple times using the default temperature value until an image was generated that adequately represented the question according to human judgment. In certain cases, Stable Diffusion's image-to-image generation process was used, where a previously-generated image that captured some but not all of the correct properties was fed back into the model along with a text prompt based on the question. Fig. 2 shows an example image generated with Stable Diffusion. A total of 91 (62 affordance and 29 habitat) images (saved in JPEG format) were generated through this process.

Simulated images Finally, for a small number of some questions that represented scenarios governed by physical dynamics, images that accurately depicted the scene could only be effectively captured as still frames from a video. For these, scenes were constructed in the Unity game engine and populated with objects described in the question,

[2]Works such as Nath et al. (2024) have convincingly argued for the utility of AI-generated images in multimodal NLP tasks.

Category	# Questions	# Images
put-9.1-2	25	35
contain-15.4	25	65
cut-21.1-1	1	2
shake-22.3-1-1	4	11
turn-26.6.1	3	6
knead-26.5	1	2
bend-45.2	26	33
break-45.1	3	8
roll-51.3.1	54	99
Total	94	231

Table 1: Distribution of questions and associated images into different affordance categories. Images or questions may belong to multiple categories and questions have multiple associated images.

Category	# Questions	# Images
Intrinsic	26	163
Extrinsic	27	125
Hot Spot	42	45
Resultant State	21	67
Concave/Convex	57	143
Constraints	4	7
Subcomponents	5	27
Habitat Chain	15	35
Total	116	386

Table 2: Distribution of questions and associated images into different habitat categories. Images or questions may belong to multiple categories and questions have multiple associated images.

with all relevant physical properties (weights, material, density, etc.) encoded in the scene. The scene was then run and environmental physics allowed to apply. We used the Unity API to save JPEG images from the scene at the moment the scenario in the question was best represented. A total of 48 (40 affordance and 8 habitat) simulated images were collected, representing questions about material properties and their effects on motion in space.

3.3 Question and Image Categorization

We categorized our questions in the habitat domain into groups reflecting different features drawn from Pustejovsky and Krishnaswamy (2016)'s VoxML, a modeling language with a formalism for habitats. In the affordance domain, categories reflect classes of the VerbNet hierarchy (Kipper et al., 2000), which is compatible with Pustejovsky (1995)'s Generative Lexicon (GL), and hence VoxML and habitats. Note that questions, but more so the associated images, may belong to multiple categories because they select for or display multifunctional properties of the focus object (Pustejovsky, 2001), and therefore the sum over all categories may exceed the value given in "Total". Table 1 shows the distribution of affordance-centering questions and associated images into the different affordance categories, mapped to VerbNet classes. The most dense categories involve translation motions (sliding, rolling, bouncing—VerbNet class `roll-51.3.1`), flexibility (`bend-45.2`), placement/stacking (`put-9.1-2`), and containment (`contain-15.4`).

Following definitions in Pustejovsky (2012) and Pustejovsky and Krishnaswamy (2016), the **habitat** domain in our data concerns questions about objects' *intrinsic* or *extrinsic* habitats, their functional regions (or "hot spots"—such as the tines of a fork, also discussed in Nagarajan et al. (2019)), resultant states under transformation (these typically take the form of counterfactual questions, such as "what would be the result if..."; cf. Pustejovsky and Batiukova (2019)), concavity/convexity (also includes openings of containers, which is shared with the "hot spot" category), constraints (usually concerning size), and subcomponents of objects (usually articulated as questions involving object with sub-parts made of different materials). Following Krishnaswamy and Pustejovsky (2022) where the enactment of a program π over an object may result in state $\mathcal{R}$ that is itself a new habitat, we also have questions pertaining to compositions of habitats (such as objects stacked on top of each other in different configurations). We will call these **habitat chains**. Table 2 shows the distribution of questions and images into the different categories.

4 Evaluation

We performed zero-shot evaluations of habitat-centered and affordance-centered reasoning on 16 LLMs, including 11 text-only and 5 multimodal models, and a random guessing baseline. The models we test include members of the LLaMA 2 (Touvron et al., 2023) and LLaMA 3 (AI@Meta, 2024) families, the FLAN family (Chung et al., 2024; Chia et al., 2023), and the LLaVA 1.5 and 1.6 families (Liu et al., 2023), with different parameter sizes, as well as UnifiedQA-v2-large (Khashabi et al., 2022) and BLIP (Li et al., 2022). The default evaluation settings were used for all models, and results come from a single evaluation run. Experiments were run on 2 NVIDIA RTX A6000 48 GB device. Input to the text-only models included the question as written with the answer

choices. Input to the multimodal models (BLIP and LLaVA) included the question plus an image resized to 512×512 pixels. Because multiple images were associated with each question, we input the question with each image, and the accuracy of the model on that question was considered to be the proportion of times the model got the question correct.[3]

5 Results

Fig. 3 shows accuracy of each evaluated model on the habitat-centering and affordance-centering questions. Text-only models are shown in blue and multimodal models are shown in orange. Results shown are from a single evaluation run using each model's default settings. We also present values for what random guessing would achieve (in green). This was computed by performing 1,000 iterations of randomly selecting an answer for each question, and averaging accuracy over all iterations. Random guessing on affordance-centering questions results in 32% accuracy, and on habitat-centering questions, 48%. In the habitat scenario particularly, all LLMs are performing around the level of random guessing.

Performance on affordances is widely variable, ranging from 22% accuracy (BLIP) to 77% accuracy (LLaMA 3-70B). On affordances, larger parameter sizes within the same family of models displays at least a weak correlation with performance. At 77% accuracy, LLaMA 3-70B almost reaches performance that is likely to be sufficient for many reasoning tasks involving affordances. There does not exist a direct comparison on affordance reasoning, but one can consider reported human performance on foundational attribute comprehension by Wang et al. (2023), which hovers around 80% agreement with majority human response. Meanwhile the best performing multimodal model was LLaVA 1.6-34B at 69% accuracy, which is a small improvement on contemporaneous text-only models, like LLaMA 2-70B, implying that images concerning affordances provide a modest performance boost to a sufficiently large model.

By contrast, we actually see very consistent, if mediocre, performance on habitats. All models, even the newest ones, hit a performance plateau with an upper bound of 57%. The best-performing models, UnifiedQA-v2 and LLaVA 1.6-34B, achieve only 57% and 55% accuracy, respectively, which is only about 7–9% better than random chance. We also see that the LLaVA family of models, which, as multimodal models, are evaluated using image inputs, performs consistently worse on habitats than affordances.

Performance by Image Type When evaluating LLaVA 1.6-34B, the best-performing multimodal model, according to the type of image used in the input (natural, generated, or simulated), we observe that images of different provenance may perform differently. Table 3 shows mean performance (% correct) and failure rate (defined as the percent of images that the model fails to answer the question correctly for). Although we have a small number, simulated habitat images may be cleaner with fewer potential distractor objects, and show higher performance. Generated habitat images tend to underperform because many of them are associated with counterfactual/resultant state questions, which is one of the most challenging categories. Meanwhile, generated affordance images tend to perform well. The only AI-generated images that accurately reflected the situation described in the question came from questions that were simpler and involved affordances from common VerbNet classes like `roll-51.3.1`. We hypothesize that the language encoders of diffusion models share biases with the models under evaluation here (see Sec. 6).

	Habitats	
Image Type	**Mean performance**	**Failure rate**
Natural	54% (.46)	38%
Generated	57% (.49)	38%
Simulated	75% (.50)	25%
	Affordances	
Image Type	**Mean performance**	**Failure rate**
Natural	63% (.42)	26%
Generated	82% (.33)	20%
Simulated	62% (.48)	35%

Table 3: Mean performance (stdev in brackets) and failure rates of different image types in LLaVA 1.6-34B.

6 Discussion

Affordance questions that LLaMA 3-70B (the best multimodal model, at 69% accuracy) failed on typically concern object multifunctionality. E.g.,[4]

[3] That is, if a question had 4 associated images and a model answered correctly when given 3 of those images and incorrectly with 1, the model was considered to have been 75% correct.

[4] In this and all following examples, the correct answer is bolded, as in Sec. 3.

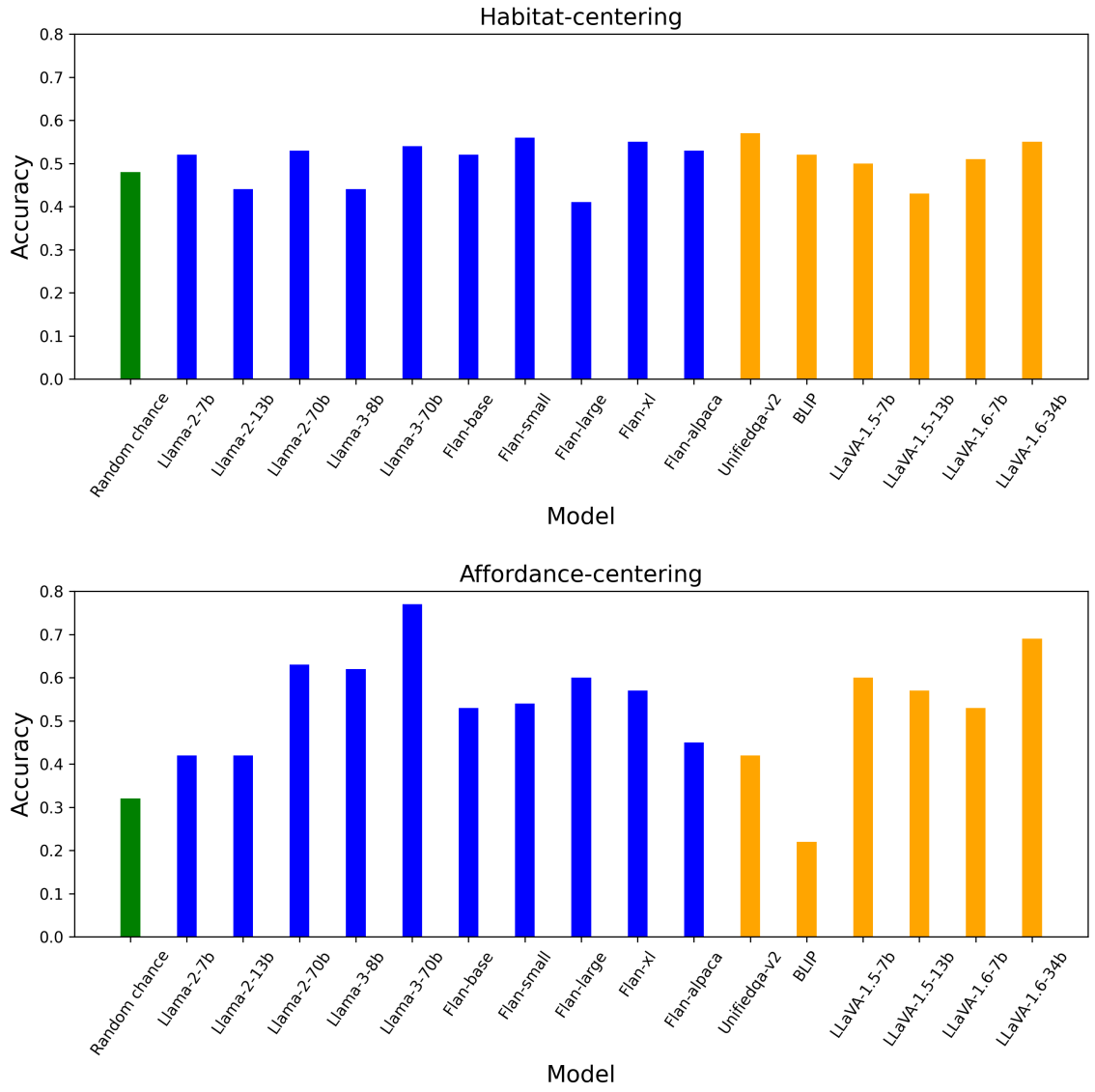

Figure 3: Accuracy of all evaluated models on habitat-centering (top) and affordance-centering (bottom) questions. Text-only models shown in blue and multimodal models shown in orange. Random chance shown in green. Within the same model family, models are ordered left-to-right by parameter size. "Flan-alpaca" denotes the `Flan-Alpaca-GPT4-XL model`.

What is a glass capable of?
a) stacking
b) sliding
c) containing
d) rolling
e) **all options are correct answers**

LLaMA 3-70B misses the *rolling* affordance of the multifunctional *glass* object. If we consider the afforded behavior *roll* in the context of its VerbNet class, *roll* alternates with *slide*, which can happen to a glass in its default orientational configuration (habitat), and so the collocation "glass" + "slide" is likely more common in the training data of even a very large model in contexts in which "roll" might also occur. Similarly, LLaMA 3-70B also failed on questions like the following:

Which of the following objects has a flat surface that would allow it to slide?
a) a beach ball
b) a book
c) a sphere
d) a cylinder
e) **b and d are correct choices**

Knowledge is needed of the flat or round parts of objects to accommodate the corresponding action. In general we observe that models have difficulty selecting innovative or non-traditional uses

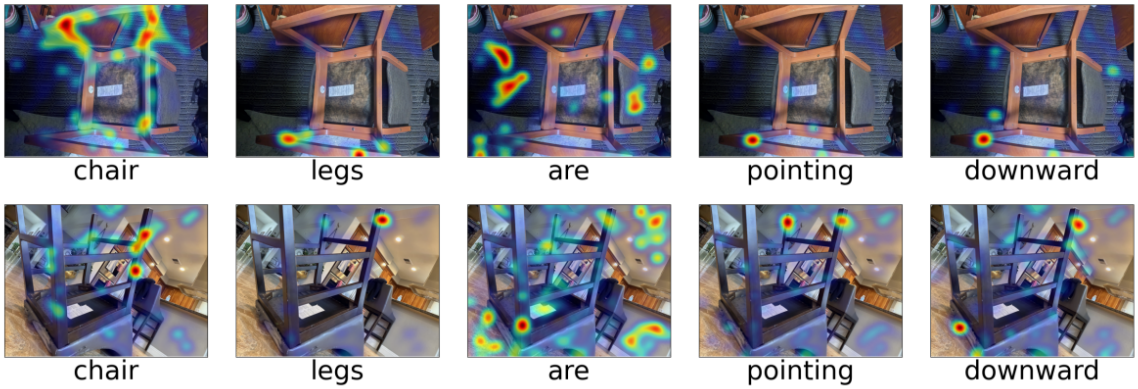

Figure 4: BLIP habitat failure case.

of objects, like using a bowl or glass to scoop. This indicates that challenging affordance reasoning typically involves some knowledge of object orientation, configuration, and habitat.

The best performing **habitat** models, UnifiedQA-v2 and LLaVA 1.6-34B, fail most frequently on questions involving resultant states under transformations, including counterfactuals:

> There is an upright coffee pot on a table. How would its bottom be positioned if the coffee pot were rotated 180 degrees?
> a) **coffee pot bottom would face upward**
> b) coffee pot bottom would touch the table

UnifiedQA-v2 fails on questions that isolate a counterfactual or resultant state and habitat chains, and appears to confuse concavity and convexity:

> A cup is placed mouth-down in the pot. How is the cup's concave side oriented?
> a) **facing toward the opening of the pot**
> b) facing toward the bottom of the pot

When we examine questions that LLaVA 1.6-34B never answered correctly with any corresponding image, we see patterns relative to the categories in Table 2. 12/57 Concave/Convex question always fail (21%), as do 7/21 Resultant State questions (33%), 10/26 Intrinsic questions (38%), 2/5 Subcomponent questions (40%), and markedly, 10/15 (67%) of the "Habitat Chain" questions, such as:

> A bowl is placed mouth-down on a table. An inverted mug is placed upside-down on top of the bowl and an upright can is placed on top of the mug. Is the opening of the mug obstructed?
> a) **mug opening is obstructed**
> b) mug opening is not obstructed

Vision-Language Grounding BLIP is notably the *worst*-performing model on affordances, but generally on par with all other on habitats. Given this discrepancy, an examination of some outputs of BLIP's habitat-centered reasoning is illustrative of where multimodal model failures may be occurring. Fig. 4 shows two different images of upside-down chairs paired with the same question:

> A chair is flipped-over. As the chair is positioned right now, in which direction are its legs pointing?
> a) **chair legs are pointing upward**
> b) chair legs are pointing downward

In both cases, the visualization of BLIP attention over "legs" and "downward" have a very similar distribution (over parts of the legs). This highlights the strong bias toward "downward" that is introduced by "legs," even when the image obviously depicts the opposite. The image input introduces noise given the biases inherent in the model.

That is, a *telic role*, as would be expressed in, e.g., Qualia Structure (Pustejovsky, 1995):

$$\lambda x \exists y \left[\begin{array}{l} \textit{chair} \\ \text{QS} = \left[\begin{array}{l} \text{F} = \textbf{phys(x)} \\ \text{T} = \lambda z, e[sit\_in(e, z, x)] \end{array} \right] \end{array} \right]$$

leaves a trace of sufficient density in unstructured free text of the kind trained into LLMs, whereas the habitat that perhaps conditions whether that telic role can be exploited may not be as present. Namely, sentences like "chairs are for sitting in" are expected to some degree, meaning the collocation between *chair* and *sit in* allows the model to effectively learn this property, but sentences expressing semantics like "chairs must be upright to be sat in" and its corollary "a chair's legs point up if its seat points down" are substantially more rare, leading to a bias toward affordance-centering

semantics like "chairs are for sitting in" and against corresponding habitat-centering semantics.

Fig. 5 shows further evidence that large VLMs such as BLIP are biased toward canonical object orientations and their relative positioning. The question associated with these 2 images, taken from different perspectives, is:

> A jar is placed mouth-down on the carpet. An upright glass is on the jar. What touches the bottom of the glass?
> a) bottom of glass is in touch with jar bottom
> b) **bottom of glass is in touch with jar lid**

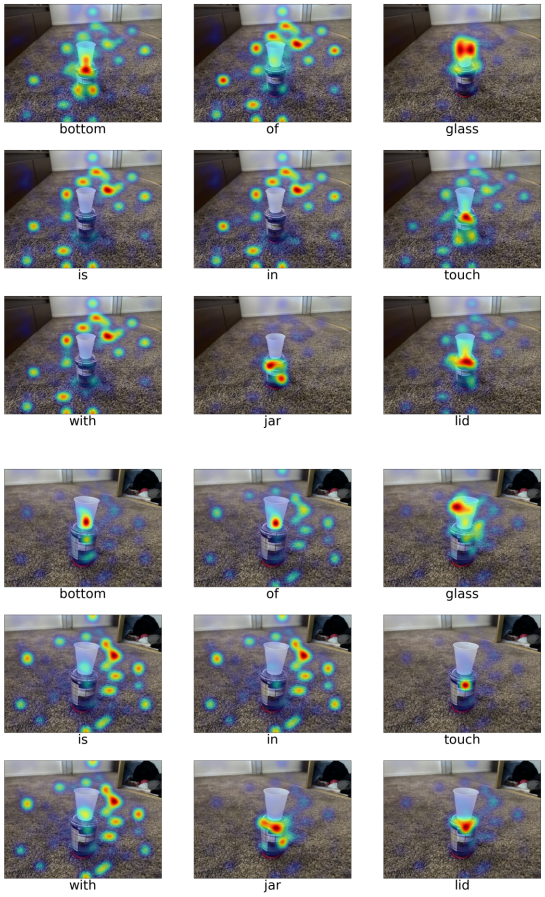

Figure 5: Visual grounding of "bottom, "glass", "touch", "jar", and "lid" tokens.

Both images show that the tokens "bottom, "glass", "touch", and "jar" are grounded to the correct region of the images, but "lid," while grounded to a portion of the jar, selects the top region, even though the jar is overturned. This indicates that the model's level of world knowledge about jars does not extend to reasoning about transformations over them, and that the collocation "jar" + "lid" induces a strong bias toward the top region of the jar that is stronger than the visual features of the lid itself.

7 Conclusion

In this paper, we presented a thorough evaluation of LLMs' and MLLMs' reasoning capabilities focused specifically on the concept of "habitats", compared to affordances. We also developed a novel challenge dataset for this task. Our data collection and collation was theoretically-grounded in the lexical semantics underlying habitats and affordances. To our knowledge, this is the first dataset and evaluation designed specifically for the under-addressed problem of habitat-centered reasoning and for both text-only and multimodal evaluation.

We found that LLMs'/MLLMs' performance on affordance-centered reasoning was highly variable but in the best cases approached 70-80% accuracy, which previous research has indicated is roughly aligned with human performance (Wang et al., 2023). By contrast, performance on habitat-centered reasoning consistently plateaued no higher than 57%, across all models. This indicates that habitat-centered reasoning remains a challenge for LLMs and their multimodal variants.

In many cases, we found that image information actually adversely impacted performance on habitats. This can be attributed to biases toward canonical orientations that previous research (Barbu et al., 2019) has found to be pervasive in image training datasets. Further, (Pustejovsky, 1995)'s formalism of the *telic role* suggests why artifacts are overwhelmingly likely to be discussed in terms of their canonical uses and associated orientations, such that when habitats are exploited to perturb that canonical alignment, it poses a unique challenge to modern models that future work must addressed to achieve true common-sense multimodal reasoning. Specific approaches may include augmenting LLMs with object and counterfactual reasoning and making them more common-sense oriented towards these types of questions using knowledge distillation (Li et al., 2023; West et al., 2021). While the consistent low performance of habitat reasoning highlights this challenge, the variability of the better-explored affordance reasoning performance, and sensitivity to small differences, highlight a need for better methods of guaranteeing or predicting model performance.

Limitations and Ethical Statement

We performed a thorough evaluation of a large number of models, but only performed interpretive probing on a small subset of them. Results

are displayed for BLIP to illustrate the biases in vision-language grounding that we observe. Due to differences in the architecture and training of each model, each requires different methods of interpretation (or at least different pipeline engineering to arrive at the same interpretable features), rendering an exploration of all models at that level out of scope due to space limitations.

Our dataset is on the smaller side, and therefore we approach it from the perspective of a challenge dataset for a specific problem rather than a benchmark. To our knowledge, this is the first organized dataset addressing habitat-centered reasoning specifically (not to be confused with the Habitat platform for embodied AI reasoning (Savva et al., 2019; Ramakrishnan et al., 2021) though the formal notion of a habitat is certainly relevant there also). Krishnaswamy and Pustejovsky (2022) note the challenge in scaling up a library of habitats, and we also find that acquiring a broad scope of habitats at present still requires time-consuming human collection to collect guaranteed, hallucination-free images and questions for. Due to the shortcomings of LLMs in habitats, it is not as easy or well-explored how to use them directly to scale up habitat vocabulary the way that they might currently be used to rapidly source affordance knowledge (Rai et al., 2024).

Affordance-based reasoning is at its core a kind of "stereotype"-based reasoning (viz. "chairs are for sitting in"). In the domain of common everyday objects, the risks of such stereotype-based reasoning are probably minimal, although if other objects with potentially harmful affordances (e.g., firearms or other weapons) are used without noting that those behaviors are harmful and should be avoided, this presents potential for abuse. Most LLMs have guardrails built-in against this type of problem, but this is not a given for all, as it is a design choice made by the developers.

Acknowledgments

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and by the National Science Foundation (NSF) under a subcontract to Colorado State University on award DRL 2019805 (Institute for Student-AI Teaming). Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of, the Department of Defense, the National Science Foundation, or the U.S. Government. All errors are the responsibility of the authors. Our thanks also go out to the anonymous reviewers whose feedback helped improve the final copy of this paper, and to Avyakta Chelle, Jade Collins, August Garibay, Olivia Jones, Rohit Sandadi, and Victoria Yang for their data collection efforts.

References

AI@Meta. 2024. Llama 3 model card.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Katherine M Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Josh Tenenbaum. 2022. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Jiafei Duan, Samson Yu, and Cheston Tan. 2021. Space: A simulator for physical interactions and causal learning in 3d environments. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 2058–2063.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2023. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2023. Grounding and distinguishing conceptual vocabulary through similarity learning in embodied simulations. *IWCS 2023*, 305:305.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2024. Exploring failure cases in multimodal reasoning about physical dynamics. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 105–114.

James J Gibson. 1977. The theory of affordances. *Hilldale, USA*, 1(2):67–82.

Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. 2023. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE.

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. 2021. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. *Advances in Neural Information Processing Systems*, 34:17427–17440.

Siyuan Huang, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. 2024. Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models. *CoRR*.

Cameron R Jones, Tyler A Chang, Seana Coulson, James A Michaelov, Sean Trott, and Benjamin Bergen. 2022. Distrubutional semantics still can't account for affordances. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 44.

Cameron R. Jones and Sean Trott. 2024. Multimodal language models show evidence of embodied simulation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11928–11933, Torino, Italy. ELRA and ICCL.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. 2000. Class-based construction of a verb lexicon. *AAAI/IAAI*, 691:696.

Kazushi Kondo, Saku Sugawara, and Akiko Aizawa. 2023. Probing physical reasoning with counter-commonsense context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 603–612.

Nikhil Krishnaswamy. 2017. *Monte Carlo Simulation Generation Through Operationalization of Spatial Primitives*. Brandeis University.

Nikhil Krishnaswamy and James Pustejovsky. 2022. Affordance embeddings for situated language understanding. *Frontiers in artificial intelligence*, 5:774752.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. 2019. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697.

Abhijnan Nath, Huma Jamil, Shafiuddin Rehan Ahmed, George Arthur Baker, Rahul Ghosh, James H Martin, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Multimodal cross-document event coreference resolution using linear semantic transfer and mixed-modality ensembles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11901–11916.

James Pustejovsky. 1995. *The generative lexicon*. MIT press.

James Pustejovsky. 2001. Type construction and the logic of concepts. *The language of word meaning*, 91123.

James Pustejovsky. 2012. The semantics of functional spaces. *Practical Theories and Empirical Practice: A linguistic perspective. Philadelphia, John Benjamins*, pages 307–325.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10.

James Pustejovsky and Olga Batiukova. 2019. *The lexicon*. Cambridge University Press.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613.

James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actions. In *International Conference on Human-Computer Interaction*, pages 137–160. Springer.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. 2024. Affordancellm: Grounding affordance from vision language models. *arXiv preprint arXiv:2401.06341*.

Arushi Rai, Kyle Buettner, and Adriana Kovashka. 2024. Strategies to leverage foundational model knowledge in object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1723.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. 2024. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. Newton: Are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9743–9758.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Sean Williams and James Huckle. 2024. Easy problems that llms get wrong. *arXiv preprint arXiv:2405.19616*.

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

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. 2024. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. 2024. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Full Vocabulary of Objects

Section 3.3 in the main body enumerates the categories of behaviors and properties singled out in the different questions.

The list of objects mentioned in the habitat-centering questions is: beaker, bench, bottle, bowl,

bucket, can, ceramic pitcher, chair, chalice, coffee pot, colander, cone, cup, decanter, desk, fork, footstool, glass, goblet, gravy boat, hammer, jar, knife, ladle, laundry basket, measuring cup, mug, nightstand, pitcher, pizza pan, plastic container, plate, pliers, pot, ramekin, roasting pan, sack, saucepan, scale, scissors, sphere, screwdriver, seesaw, sofa, soup tureen, spatula, spoon, stool, strainer basket, table, tumbler, umbrella, urn, vase, wrench.

The list of objects mentioned in the affordance-centering questions is: apple, axe, ball, barrel, basin, basket, basketball, beach ball, belt, bench, block, book, bowling ball, bottle, bowl, box, brick, briefcase, cable, canister, canteen, capsule, carafe, cardboard box, CD, chopsticks, cloth, Coca-Cola can, coin, colander, cone, cube, cutting board, cylinder, decanter, deck of cards, disk, egg, ellipsoid, envelope, flashlight, flask, fork, frying pan, glass, half of a sphere, hand saw, hockey puck, jar, jug, kettle, knife, ladle, leather strap, lunchbox, marble, measuring cup, mop, mug, pail, paper, pea seed, pitcher, pipe, plastic bag, plate, pyramid, rectangular prism, ring, rod, rolling pin, rubber band, ruler, safe, saw, screwdriver, sieve, slab, soda can, sphere, sponge, spool of thread, spoon, spray bottle, spring, stump, table, teapot, thermos, tissue box, tissue roll, toroid, torus, tray, triangular prism, tube, vase, vial, water bottle, wheel, whisk, wrench, Ziploc bag.

B Details on Image Collection

Image collectors were recruited through the authors' research lab and through student groups at a university. All volunteered their time. This activity was determined to be Not Human Subjects Research by the Institutional Review Board.

Among the image collectors, 4 were female, 2 were male, and 1 was non-binary. They were given the following instructions:

The included spreadsheet contains a set of questions about object properties in different configurations. We are trying to source different images that depict the scene in question, including the mentioned objects in the relevant configuration.

Please read the questions carefully. Consider taking multiple images for each question. Ideally, images should be somewhat cluttered, taken from non-traditional angles, with distinct lighting (everything needs to be visible, but think about creative placement of shadows, etc.). Also please consider using different backgrounds or objects with different colors.

They were instructed to not show themselves in their images, or any information that might identify themselves, their institutional affiliation, or location. Any images that inadvertently disclosed this were removed from the dataset before evaluation.

C Choice of Open-Weight LLMs

The state of the art is such that the most powerful current models, such as GPT-4, are closed and proprietary. Many recent papers present zero-shot evaluations of such models on their task, to demonstrate what performance of a strong LLM looks like. While reasonable people may disagree, in our view, the closed and proprietary nature of such model makes them invalid for rigorous comparison. There are two primary reasons for this, both pertaining to lack of guarantees provided by a closed model:

1) There is no guarantee that input to the model (such as test questions) is not saved for later training, thus artificially inflating later performance.

2) There is no guarantee that if one logs out and logs back in to continue an experiment, that the input is routed through exactly the same model weights that were accessed before, precluding an apples-to-apples comparison.

Unfortunately, the way that major industrial players handle proprietary models at this time does not lend them to robust comparison. Beyond budgetary restrictions precluding extensive GPT-4 evaluations and inability to access the model weights, too much is unknown about the behind-the-scenes functionality to make such models evaluable on a level playing field. This motivated our focus on open-weight models. Many of these open-weight models, like Flan-Alpaca-GPT4-XL, advertise performance that is competitive with models like ChatGPT or GPT-4, and we believe it is not reasonable to expect that, after evaluating over a dozen other models that show similar result, an arbitrary proprietary model would suddenly be able to address all the shortcomings of the other evaluated models.

D Data Availability and Use Statement

Our data (linked in Sec. 1) is available under a Creative Commons Attribution Non Commercial Share Alike 4.0 International (CC BY-NC-SA 4.0) or similar license for intended research use.