

TEXT2AFFORD: Probing Object Affordance Prediction abilities of Language Models solely from Text

Sayantana Adak, Daivik Agrawal, Animesh Mukherjee* and Somak Aditya*

IIT, Kharagpur
West Bengal – 721302

Abstract

We investigate the knowledge of object affordances in pre-trained language models (LMs) and pre-trained Vision-Language models (VLMs). A growing body of literature shows that PTLMs fail inconsistently and non-intuitively, demonstrating a lack of reasoning and *grounding*. To take a first step toward quantifying the effect of grounding (or lack thereof), we curate a novel and comprehensive dataset of object affordances – TEXT2AFFORD, characterized by 15 affordance classes. Unlike affordance datasets collected in vision and language domains, we annotate *in-the-wild* sentences with objects and affordances. Experimental results reveal that PTLMs exhibit limited reasoning abilities when it comes to uncommon object affordances. We also observe that pre-trained VLMs do not necessarily capture object affordances effectively. Through few-shot fine-tuning, we demonstrate improvement in affordance knowledge in PTLMs and VLMs. Our research contributes a novel dataset for language grounding tasks, and presents insights into LM capabilities, advancing the understanding of object affordances. ¹

1 Introduction

Object affordance refers to the properties of an object that determine what actions a human can perform on them (Gibson, 1979). Gaining the knowledge of object affordances while learning textual representation from large corpora maybe hard; as in NLP, we lack corresponding images (or videos) which provides necessary visual cues such as shape, color, and texture to predict affordances. This lack of mapping or rather *grounding* ability has been noted by many researchers in the context of large pretrained language models (PTLMs). Authors in Bender and Koller (2020) have pointed the

*Equal advising

¹Code and Data are available at <https://github.com/sayantana11995/Text2Afford>

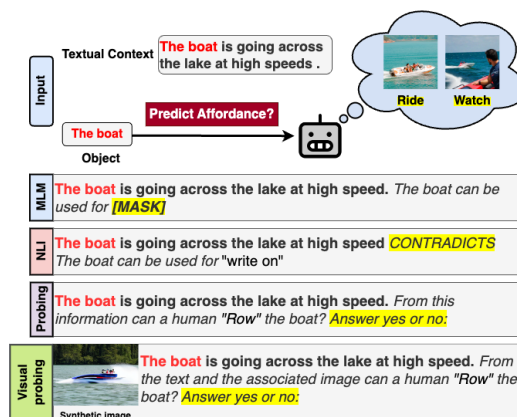


Figure 1: Overview of TEXT2AFFORD with its derived tasks.

lack of symbol grounding to be a fundamental factor behind PTLMs failing to grasp *meaning* from *form* (surface form text). The authors argue that language models which are exposed to only text (surface form) may never truly understand *meaning*, as PTLMs are unaware of possible *groundings* of the surface text. Most current NLP datasets and tasks are not designed to evaluate *grounding*, as it is hard to evaluate *grounding* without any visual context. Here, we aim to *quantify* the ability of pretrained models to learn *affordances* – which in turn requires the ability to *ground* symbols in text to real-world objects. In other words, *grounding* ability from text can enable understanding and reasoning about the physical properties of an object, which may help predict affordances.

As another example, for the sentence “an apple in the tree”, we should infer that the “apple” can be eaten, and is rollable. However we cannot roll an “apple logo”. In computer vision and robotics efforts, an accompanying image (or video) often provides necessary information about shape and physical properties of the entity, which can be used to predict affordances (Zhu et al., 2014). However such information is absent in NLP tasks. To capture this nuance, we annotate crowdsourced text intended for other tasks (such as NLI) with the

Dataset	Train size	Dev size	Test size	Reasoning type	Source	Image-dependent	Targeted affordance	Publicly available
PaCo (Qasemi et al., 2022a)	5,580	1,860	4,960	Preconditioned commonsense	Crowd-sourced	✗	✗	✓
WINOVENTI (Do and Pavlick, 2021)	-	-	4,352	Commonsense with exceptions	Crowd-sourced	✗	✗	✓
PVLIR (Qasemi et al., 2023)	-	-	34,000	Preconditioned visual commonsense	Other dataset	✓	✗	✓
Normlense (Han et al., 2023)	-	-	10,000	Defeasible visual commonsense	Crowd-sourced	✓	✗	✓
WinoViz (Jin et al., 2024)	-	-	1,380	Reasoning object’s visual property	Crowd-sourced	✗	✗	✗
PROST (Aroca-Ouellette et al., 2021)	-	-	18,736	Reasoning object’s physical property	Other dataset	✗	✗	✓
NEWTON (Wang et al., 2023)	-	-	2,800	Reasoning object’s physical property	Crowd-sourced	✗	✗	✓
Persiani and Hellström (2019)	734,002	-	314,572	Object affordance without context	Synthetic	✗	✓	✗
TEXT2AFFORD (Ours)	-	-	35,520 (2368 * 15)	Contextual object affordance	<u>Crowd-sourced</u>	✗	✓	✓

Table 1: Comparison of TEXT2AFFORD with other reasoning datasets. A larger version is in Appendix A.3 Table 9.

objects and affordances. We use 15 affordance classes from Zhu et al. (2014). Through extensive pilot studies, we train a set of annotators using the toloka.ai platform. We choose 25 highly-skilled annotators who annotated a total of 2368 sentence-object pairs with 15 affordance classes, on a 0-3 Likert-like scale. For each sentence-object pair and each affordance class we ensure annotations from three annotators to enable majority votings. We name this novel dataset TEXT2AFFORD. We use the dataset for zero-shot evaluations of small LMs, open-source LLMs and also some VLMs by forming different task setups. Figure 1 presents an example from TEXT2AFFORD and the derived tasks (detailed in Section 4). We evaluate the effect of few-shot fine-tuning on few PTLMs and VLM. Our contributions can be summarized as follows.

- We curate a novel large scale crowdsourced text to affordance dataset – TEXT2AFFORD, consisting of 35,520 test data points (2368 sentence-object pairs with 15 unique affordance classes per pair). We ensure at least three annotations for each sentence-object pair for each class.
- Using TEXT2AFFORD, we perform zero-shot evaluation of several state-of-the-art PTLMs along with a few VLMs in different settings to identify the extent to which they gain the knowledge of affordance during pretraining. We further ensemble the VLM and the PTLM predictions to examine whether pre-training with images can enrich affordance prediction from text. Overall, we observe that the SOTA LLMs face difficulties predicting contextual object affordances solely from text (accuracy < 55%) and the performance gets slightly enhanced when using powerful VLMs in presence of synthetic images.
- We also fine-tune few PTLMs on a small subset of our data as well as on some commonsense reasoning tasks to understand how quickly the affordance knowledge get scaled up and how far the affordances are related to commonsense knowl-

edge. In addition, we examine the in-context learning (ICL) ability of few of the SOTA generative LLMs and VLMs in affordance prediction task. We find that the pre-trained encoder based models gain some knowledge about object affordance during fine-tuning using the commonsense reasoning dataset.

- Additionally through finetuning on our dataset, we show that knowledge of affordance can improve model’s physical reasoning capability.

2 Related work

Reasoning about object affordances. Object affordances has been extensively studied in Computer Vision and Robotics (Sun et al. (2014); Zhu et al. (2014)). Recent methods employ deep learning approaches to detect object affordance. Nguyen et al. (2017) applies an object detector, CNN and dense conditional random fields to detect object affordance from RGB images. Persiani and Hellström (2019) extracts object-action pairs from web corpora using semantic role labelling. In contrast, we propose a crowd-sourced text only affordance dataset to audit the strength of SOTA LLMs and VLMs to reason about contextualized object affordance.

Probing methods. Talmor et al. (2020) utilizes probing and employs Multi-choice MLM (Masked Language Modelling) and Multi-choice QA (Question Answering) setup to capture reasoning capabilities of pre-trained Language Models. Yang et al. (2022) examines zero-shot prediction performances on different tasks by LLM through novel visual imagination. Aroca-Ouellette et al. (2021) highlights the shortcomings of state-of-the-art pre-trained models in physical reasoning, with a further performance decline observed when incorporating option shuffling and superlatives in reasoning questions. Liu et al. (2022) proposes a novel spatial commonsense probing framework to investigate object scales and positional relationship knowledge

in text-based pre-trained models and models with visual signals. Joshi et al. (2020) uses probing methods to investigate a more fine-grained logical reasoning capabilities of pre-trained models.

Reasoning tasks and dataset. Reasoning about object affordance require a sort of commonsense reasoning. A series of works (Singh et al. (2021), et al. (2023), Bisk et al. (2019), Huang et al. (2019), Talmor et al. (2019), Talmor et al. (2021), Zellers et al. (2018)) study the text based commonsense knowledge of language models. Dataset such as δ -NLI (Rudinger et al., 2020) focuses on defeasible inference of commonsense knowledge; PaCo (Qasemi et al., 2022a) and PInKS (Qasemi et al., 2022b) deal with preconditioned commonsense inference of language models. PVLIR (Qasemi et al., 2023), Normlense (Han et al., 2023) use images as precoditions to reason about defeasible commonsense norms. However, none of these specifically focus on reasoning of object affordance. Wang et al. (2023) proposes a benchmark of object-attribute pairs plus a diverse set of questions to reason object’s physical properties. Aroca-Ouellette et al. (2021) tackles physical and affordance reasoning from an object-centric approach. Persiani and Hellström (2019) attempts to extract common object-action pairs from web corpora. In Table 1, we demonstrate the comparison of TEXT2AFFORD with other datasets which perform different kind of reasoning tasks. We emphasize that, TEXT2AFFORD is the largest crowdsourced publicly available text based contextualized affordance dataset with a test size of 35,520 (2368 sentence-object pairs and 15 affordance classes).

Present work. Although a substantial number of work study the reasoning capabilities of language models and propose commonsense reasoning datasets, however, none of these work concentrate specifically on evaluating the knowledge of affordance and contextual affordance prediction *solely* from text. To bridge this gap, we present a reliable crowdsourced test dataset for identifying the contextualized affordance prediction capability of LLMs as well as VLMs. Our results show that the advanced large language models fail to understand an object’s physical properties aka the affordances from texts, and there is significant room for improvement which may further motivate researchers to explore models that explicitly learns to *ground* objects in text to predict its physical properties and affordances.

3 TEXT2AFFORD dataset construction

We select 20,000 sentences from a crowdsourced English dataset (XNLI English) (Conneau et al., 2018)² and extract the noun phrases using the Stanford CoreNLP tool. As we restrict to the affordances that humans can directly perform, we filter the phrases which do not represent a tangible object (using ConceptNet). We manually filter out objects that cannot be acted upon directly by humans (such as school, building). After this preprocessing, we obtain a set of sentence-object pairs ($\langle x_i, o_i \rangle$), where the sentence acts as the context for the corresponding object. Each sentence on average has 2-3 such objects. We use the 15 predefined affordance classes from Zhu et al. (2014) to label each sentence-object pair for annotation.

We utilize the Toloka platform³ for conducting the data annotation. We design an interface on this platform, containing clear instructions and examples for annotating the data. We conduct two rounds of pilot studies along with additional AMA (Ask Me Anything) sessions to analyze the subjective understanding of the annotators and, thereby, only select the high quality, serious annotators. A total of 114 annotators participated in the pilot study, and out of that we finally engage 25 skilled annotators to annotate a total of 2,368 sentence-object pairs each containing 15 affordance classes. Each datapoint (i.e., sentence-object pair along with an affordance class) has been annotated by *three* different annotators. We provide the details of the *pilot studies & annotator training* in Appendix A.1. By evaluating the complexity of the task for the annotators from the pilot studies, we intentionally consider a relatively small number of datapoints at a point for the annotation. This leads us to a total of 10 phases to complete the final annotation. We carefully reviewed each annotation and provided feedback with guidance in case of mistakes. For instance, annotators initially got confused with the affordance ‘Watch’ as human can *watch* any visual objects. In another instance, some annotators asked whether ‘Throw’ can be valid affordance for the object ‘Kittens’ as humans can perform ‘Lift’, ‘Throw’ to the object ‘Kitten’. We discussed these types of ambiguities with the annotators after each phase. Throughout each of the data annotation phases, we put scrupulous attention to quality con-

²We choose XNLI as a source to facilitate multilingual extensions of our dataset.

³<https://toloka.ai/>

Agreement category	Affordance classes	Objects	Object-affordance pair
High agreement (>0.75)	Row, Feed, Ride, Fix	the horse, striped white shirts, a brown paper sack, Chinese lanterns, Adrin’s sword, The movie	breakfast-Feed, a horse-Watch, crops-Fix, sports-Grasp, sports-Lift, sports-Push, the phone-Feed, football-Ride
Medium agreement (>0.4 & <0.75)	Throw, PourFrom, WriteWith, LookThrough, Lift, Grasp, Play, Push	A red flag, An arrow, Art, Automatic weapons, Babies, Black-and-white TV	computers-WriteWith, cats-Feed, football-Play, book-WriteWith, the door-Push
Low agreement (<0.4)	Watch, SitOn, TypeOn	Brandy from Spain, stone circles, iron, batteries, his fist, historical artifact, gift, olive oil, outdoor tables, bumper sticker on a car	weapon-Push, The table-Lift, boat-Fix, paintings-LookThrough, cats-Throw

Table 2: Agreement based on difficulty in disambiguating different affordance classes, objects and object-affordance pairs.

# of sentence-object pairs annotated	2368
# of affordance class	15
# of instances annotated	106560 (2368 × 15 × 3)
Avg # of objects / class	333
Most prominent class	Lift (851 objects)
Least prominent class	WriteWith (3 objects)
Total skilled annotators used	25
Avg agreement (Krippendorff’s α)	0.68

Table 3: TEXT2AFFORD dataset statistics. # of instances annotated: (# of <s-o> pairs) * (# of classes) * (# of annotations per class).

trol, including iterative annotation refinement, and manual evaluation. The overall statistics for this *currently* constructed dataset – TEXT2AFFORD is in Table 3. The TEXT2AFFORD dataset consists of 2368 sentence-object pairs having $\sim 100k$ annotations (2368 × 15 × 3). For further details of the dataset construction, and our method of handling ambiguous scenarios, we refer the reader to Appendix A.1.

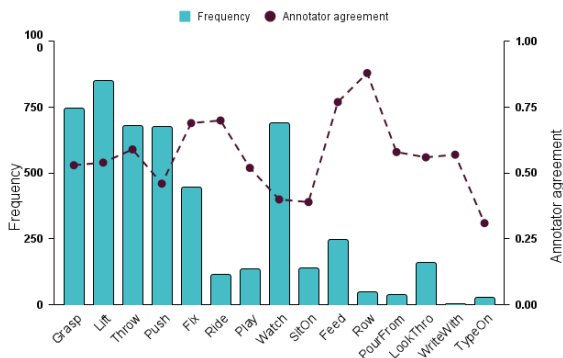


Figure 2: Classwise distribution of the number of objects and the annotator agreement.

TEXT2AFFORD data exploration. We observe that classes such as ‘Grasp’, ‘Lift’, ‘Throw’, ‘Push’, and ‘Watch’ are the most common affordances for the objects present in the dataset (see Figure 2). Most frequent objects and their corresponding agreement scores are shown in Appendix A.10 Fig. 8. We observe, agreement scores are fairly uniform (0.5-0.6) for frequent objects, with high agreement for some frequent objects (0.8 for “the movie”). In Figure 9 (see Appendix A.9), we also see that ‘Grasp’, ‘Lift’, and ‘Throw’ are highly cor-

related classes. There is similar positive correlation between the class ‘SitOn’ and ‘Ride’, and some correlation between ‘Watch’ and ‘LookThrough’. In Table 2, we list down the affordance classes based on the annotator agreement score, and divide it into *three* categories to understand which of the affordance classes pose the most and least difficulties for the human annotators. We observe that the classes - ‘Watch’, ‘SitOn’, and ‘TypeOn’ are the most difficult to disambiguate. Further, to explore the difficulty of understanding contextual object affordance, we employ three *naïve annotators* to annotate some samples of the TEXT2AFFORD, and we observe that on an average in 88% cases the humans are able to predict affordance correctly, and in some cases the *context* introducing inherent difficulty for predicting affordance. Details of the study is provided in Appendix A.2.

4 Task description

Our first objective is to *audit* the strength of Large Language Models in identifying the pre-defined affordance classes of objects from text in zero-shot settings. Given a textual context, and the object, the task is to predict whether a particular affordance class is applicable to that object conditioned on the context. We majorly leverage 4 types of task setup for the experiments. For the encoder based models (e.g., RoBERTa, BERT) we choose Masked Language Modelling (MLM) and Natural Language Inference (NLI) based setup, and for the generative models we adopt 2 types of probing setup (text only, text+image) to formalize the task. Table 4 demonstrates different types of tasks that we engage for conducting the experiments from the TEXT2AFFORD dataset.

5 Experiments

We explore various state-of-the-art baselines using pre-trained language models (RoBERTa-large, BART-large), instruction-fine-tuned large language models (e.g., FLAN-T5, Falcon, ChatGPT, Llama-3), pre-trained multi-modal vision and language architectures (CLIP-ViT, ViLT, InstructBLIP,


Model architecture	Tasks	Input instance	Output
Encoder based	MLM	All the women in India wear bangles [SEP_TOKEN] bangles can be used for [MASK_TOKEN] by human	Probabilities of each affordance classes as the [MASK_TOKEN]
	NLI	premise: All the women in India wear bangles hypothesis: bangles can be used for <Affordance> by human	Entailment scores for each affordance classes
Generation based	Probing with text	Consider the sentence - 'All the women in India wear bangles'. Now, from this information can a human <Affordance> bangles? Answer YES or NO:	YES\NO
	Probing with text and image	 Consider the sentence - 'All the women in India wear bangles'. Now, from this information can a human <Affordance> bangles? Accompanying this query is an image of the bangles. Answer YES or NO:	YES\NO

Table 4: Overview of the tasks using TEXT2AFFORD. For detailed prompting see Appendix 10.

IDEFICS, LLaVA). We observe whether these models gain the knowledge of affordances through their pre-training, fine-tuning on commonsense tasks (NLI, PIQA), or few-shot fine-tuning scenarios.

5.1 Zero-shot affordance prediction

5.1.1 Pre-trained language models

We frame the zero-shot prediction task in different ways.

MLM based approach. Here, we pose the zero-shot task as masked word prediction problem. We choose BERT-large-uncased, RoBERTa-large (Zhuang et al., 2021), and BART-large (Lewis et al., 2020) models for the experiment. We pass the sentence and prompt separated by a [SEP] token as an input to the model. We use the prompt “<Object> can be used for <MASK_TOKEN> by human” and obtain the probability of each affordance label using the logit corresponding to the <MASK_TOKEN>.

Predictions from generative LLMs. We pose the task as ‘YES\NO’ questions-answering format and apply autoregressive language models such as FLAN-T5 (Chung et al., 2022) (large, xl, and xxl), Falcon (Almazrouei et al., 2023) (7b and 40b), Llama-3⁴, ChatGPT to get the predictions. We provide with a ‘YES\NO’ question-answer based prompt to the LLMs to predict whether a particular affordance can be performed on the given object. Based on rigorous prompt engineering we choose specific prompts for the different models as shown in the Appendix Table 10. We map ‘YES\NO’ predictions to 1\0 labels respectively.

5.1.2 Commonsense reasoning tasks

To understand whether the injection of the common sense knowledge in the pre-trained models can enhance the performance of the affordance prediction, we first fine-tune the pre-trained models on common sense reasoning dataset such as PIQA (Bisk et al., 2019). Then we run the fine-tuned models on our dataset using the MLM setup. We

use BERT-base, BERT-large, RoBERTa-large, and BART-large models.

Apart from this, we leverage RoBERTa-large and BART-large fine-tuned on the Multi-genre NLI (MNLI) corpus (Williams et al., 2018) to evaluate on NLI setup. We utilize the sentence as premise and use the hypothesis as “<object> can be used for <affordance> by human” for each object-affordance pair, and use the entailment score to rank the affordance classes and report mAP and accuracy. Details of the experiment can be found in Appendix B.2.1.

5.1.3 Multimodal models

We explore both unimodal and multi-modal task setup for pre-trained vision and language models.

Text-only MLM setup

VLMs are pre-trained on large datasets having both image and text. The main goal of their pre-training is to capture some visual knowledge corresponding to the text while pre-training on multi-modal dataset such as image-caption pairs. To examine this, we first use the vision-language model CLIP, by providing only text prompt as the input and predict the affordance in an MLM setup.

Multimodal task setup

Images contain necessary information about shape, texture, and size of objects that can be utilized to effectively predict an object affordance (such as the handle of the bucket can be used to grasp and lift). Hence, we also convert the problem into a multi-modal task by synthesizing corresponding images from the context sentence, and predict the affordance of an object (mentioned in the sentence) based on the input.

Synthesizing images. In this setup, we use two different techniques to synthesize *semantically close* images to corresponding context sentences using 1) retrieval and 2) generation. We further use top five images for both, to get an accurate estimation. *Image retrieval:* We use the CLIP (Radford et al.,

⁴<https://github.com/meta-llama/llama3>

2021) based sentence-transformers architecture to search for top five semantically similar images for each of the contexts from the Visualgenome (Krishna et al., 2017) dataset.

Image generation: We adopt the generative *StableDiffusion* (Rombach et al., 2022) model to generate top five images based on the sentence as a text prompt. Details can be found in the Appendix B.4.1.

We use the top five retrieved images by using retrieval and generation methods each. We use *CLIP* (Radford et al., 2021) and *ViLT* (Kim et al., 2021) as our vision-text models. CLIP has a text encoder f_T and a visual encoder f_V , which can project text and image into the shared latent space. We aggregate the k ($=5$) corresponding images and use CLIP to compute the relevance score of (x, y) : $Score_{VI}(x, y) = \frac{1}{k} \sum_{i=1}^K \cos(f_T(x), f_v(I_y^k))$, where I_y^k is the k^{th} image for the input text y . In the ViLT model, we provide the text prompt along with the representative images as input to predict the masked token. We use the same prompt as the previous MLM task (i.e., “<Object> can be used for <MASK_TOKEN> by human.”) and get the probability of each affordance class as the logit corresponding to the <MASK_TOKEN>.

Text generation based. Similar to section 5.1.1, we utilize state-of-the-art VLMs to make predictions regarding object affordances. We provide with a ‘YES\NO’ question answering based text prompt along with the aligned images as input to the VLMs, and the model should generate an answer whether a particular affordance can be performed on the given object. We use state-of-the-art VLMs such as IDEFICS (Laurençon et al., 2023), LLaVA (Liu et al., 2023b), InstructBLIP (Dai et al., 2023) for this task. The text prompt used for the models can be found in the Appendix D, Table 10.

Ensemble language and vision prediction. Following Yang et al. (2022), we use the weighted sum as the late fusion over the final output probabilities of each affordance class from the language and multi-modal models. Experimental details can be found in Appendix B.3.

5.2 Few-shot prediction

We conduct few-shot experiments by 1) fine-tuning the encoder based models, 2) randomly selecting 5 demonstration examples for the generative models to perform few-shot in-context learning (ICL). We consider the 62 annotated objects and correspond-

ing 15 affordance classes by Zhu et al. (2014) for the few-shot based experiments.

Training data To create few-shot training examples for fine-tuning encoder based PTLMs, we take all the 62 objects, and for each object we randomly select exactly 1 positive affordance class (i.e., the class label annotated as 1) and 1 negative affordance class (i.e., the class label annotated as 0) for generating the training prompt. Overall they constitute 124 training examples (62 sentence-object pairs and 2 selected classes for each) for the few-shot experiment. For more details of the training data curation and the selection of examples for in-context learning, refer to Appendix B.4

Experimental setup. We fine-tune the encoder based language models using the training data, and for the generative LLMs and the VLMs, we utilize the training data to select in-context demonstration examples.

Fine-tuning PTLM: We fine-tune the encoder based PTLMs in NLI based setup having the context sentence as premise and use same hypothesis (i.e., “<object> can be used for <affordance> by human”) which we use in the zero-shot settings. We use BERT-large-uncased, RoBERTa-large and BART-large for fine-tuning in this setup. For implementation details refer to Appendix B.4

In-context learning for generative models: We employ the same generative LLMs as well as VLMs to perform affordance prediction using five demonstration examples from the training data. We use the same text prompt as zero-shot setting and concatenate the five demonstration examples along with corresponding label (i.e., ‘NO’ for positive class, and ‘NO’ for the negative class) to the prompt and ask the LLMs and VLMs to predict the affordance. In case of the VLMs, we do not provide any additional image example here.

6 Benchmarking TEXT2AFFORD prediction

Evaluation metric. To assess the performance of the zero-shot affordance prediction, we calculate accuracy in the following way. Each affordance class is treated as a binary classification problem, where a value of 1 represents a positive class indicating that the affordance can be performed on the object, and a value of 0 represents a negative class indicating that the affordance cannot be performed. For each positive class $\in \{P_1, P_2, ..P_n\}$, we compare the predicted scores of that affordance class

with the predicted scores of the negative classes $\in \{N_1, N_2, \dots, N_m\}$. If the predicted score of the positive class is higher than the predicted score of all the negative classes (i.e., $p(P_i) > p(N_j)_{\forall j}$), we increment the correct count by 1⁵. Conversely, if the predicted score of the negative class is higher, we increment the wrong count by 1. The final accuracy is calculated by dividing the total number of correct counts by the total number of the instances. To rank the affordance classes based on the predicted score, we also report the Mean Average Precision (mAP@K, where K is the number of affordance classes).

Encoder based								
NLI based								
Model	Actual		Fine-tuned		LM + VI (CLIP)		LM + VI (ViLT)	
	Acc	mAP	Acc	mAP	Acc	mAP	Acc	mAP
RoBERTa-large-mnli	0.64	0.43	0.72	0.49	0.79	0.52	0.79	0.54
BART-large-mnli	0.65	0.38	0.69	0.48	0.62	0.4	0.64	0.43
MLM based								
BERT-large-uncased	0.46	0.26	0.58	0.33	0.55	0.38	0.53	0.37
RoBERTa-large	0.55	0.36	0.77	0.49	0.61	0.41	0.62	0.43
BART-large	0.47	0.28	0.65	0.38	0.56	0.35	0.52	0.34
Multi-modal models (zero-shot)								
CLIP-ViT (text-only)	0.47	0.34	-	-	-	-	-	-
CLIP-ViT (retrieval)	0.56	0.35	-	-	-	-	-	-
CLIP-ViT (generation)	0.61	0.4	-	-	-	-	-	-
ViLT (retrieval)	0.41	0.31	-	-	-	-	-	-
ViLT (generation)	0.44	0.32	-	-	-	-	-	-

Table 5: Performance for affordance prediction using encoder based models. Acc: Accuracy, LM: Language model, VI: Vision. Only LMs are ensembled with VI. The best results are in bold.

Generation based				
Predictions from generative LLM				
Model	Acc (zero-shot)		Acc (ICL)	
Random baseline	0.18		-	
FLAN-T5-large	0.06		0.13±0.04	
FLAN-T5-xl	0.07		0.21±0.03	
FLAN-T5-xxl	0.33		0.39±0.04	
Falcon-7b-instruct	0.19		0.24±0.03	
Falcon-40b-instruct	0.43		0.47±0.06	
Llama-3-8b-instruct	0.36		0.43±0.05	
ChatGPT (GPT-3.5 turbo)	0.41		0.44±0.05	
Multi-modal models				
Model	Acc (zero-shot)		Acc (ICL)	
	IR based	IG based	IR based	IG based
Idefics-9b-instruct	0.26	0.25	0.36±0.02	0.37±0.03
Llava-1.5-7b	0.32	0.34	0.36±0.03	0.40±0.04
InstructBlip-vicuna-13b	0.37	0.39	0.43±0.03	0.45±0.03
InstructBlip-flan-t5-xl	0.12	0.16	0.15±0.02	0.18±0.02
InstructBlip-flan-t5-xxl	0.39	0.45	0.48±0.04	0.53±0.05

Table 6: Zero-shot and in-context learning (ICL) performance for affordance prediction using generative models. IR: Image Retrieval; IG: Image Generation. Number of demonstration examples used for ICL = 5. We also mention the variance over different selections of examples. The best results are in bold.

Zero-shot performance. Table 5 shows the results of the zero-shot affordance predictions from the mentioned models. The second column (i.e., Ac-

⁵During calculation we discard the cases when there is no positive class for a sentence-object pair in the ground truth. We do not find any instance where no negative class is present.

MLM based			
Model	Accuracy mAP		
BERT-base-uncased-finetuned-piqa	0.45	0.26	
BERT-large-uncased-finetuned-piqa	0.56	0.29	
RoBERTa-large-finetuned-piqa	0.64	0.45	
BART-large-finetuned-piqa	0.59	0.35	

Table 7: Affordance prediction using models trained on commonsense data. Best results are marked in bold.

tual) indicates the values from the original LM and multi-modal models. The third and fourth columns (i.e., LM + VI) indicate the performances of ensembling language models with two of the multi-modal models we used. We observe that, the PTLMs have some knowledge about object affordances, but they still lack the comprehensive reasoning ability about these affordances, which is reflected in the low mAP values. Further, the performances vary across different settings. In case of NLI based setup, the fine-tuned RoBERTa and BART models show improvement in the performance, which indicates that *during fine-tuning on MNLI dataset, those models gain some reasoning ability*. In Table 6 we show the generation based results in a zero-shot setting. In case of FLAN-T5-large model, where we use it to predict a binary label (YES\NO) for an affordance class, the performance drops significantly (the accuracy is less than 7%). This shows that there are still some challenges for the text-to-text models in general reasoning ability about the object affordances. In addition, we find that, the multi-modal models do not perform well in text-only settings, despite being pretrained on text and image data. The performances of the language models get boosted when ensembling with the multi-modal models, which indicates that the prediction of object affordance from sentence is a difficult task, and can be enhanced in presence of images. In addition to evaluating generative models, we establish a random baseline (Detailed in Appendix B.1). Interestingly, we find that models like Flan-T5-large and Flan-T5-XL underperform compared to this random baseline in zero-shot settings.

Finetuning on commonsense datasets. We observe that the fine-tuned model on commonsense reasoning task (Table 7) show improved performance for the affordance prediction task. This indicates that the pre-trained models lack the reasoning of object affordance. Interestingly, we find that the smallest BERT-base model fine-tuned on PIQA, performs almost similar to that of the BERT-large or BART-large models (see Table 5).

Few-shot performance. We find that, in presence of few examples from our affordance dataset, the

reasoning capability about object affordances can be enhanced for the PTLMs. The results with 124 shots (62 pairs as discussed earlier) are noted in Table 5. In Table 6, we note the results for the in-context learning performance of the generative LLMs and VLMs. We observe a significant performance gain over zero-shot settings. Having said that, we also observe that, even with the in-context learning, the performance of the generative models (with more than 7b parameters) do not reach even close to the performance of the fine-tuned BERT-large model (340M parameters). This suggests that, for the specific affordance prediction tasks from text, finetuning is absolutely essential even for the state-of-the-art LLMs and VLMs.

Error analysis

Encoder based models. We conducted a qualitative analysis of the erroneous cases for the two models (BART-large and RoBERTa-large) in MNLi settings to understand what are the typical causes of errors. We take examples where accuracy is below 0.3. Consider the representative example below.

Sentence: The salt from La Mata is often used as table salt. *Object:* table salt
 Top 5 predicted affordances (according to the probability score) - ['sitOn', 'pourFrom', 'grasp', 'fix', 'lookThrough']

The model predicts 'SitOn' as the top affordance for table salt, implying that the model misinterprets "table salt" with "table". Similarly, for the object "the window sill", the model predicts 'lookThrough', 'watch' as top affordances, which again suggests that the model is confused between "the window sill" and a "window". In another case, the model predicts ['grasp', 'writing', 'typing', 'lookThrough', 'throw'] as the top affordance labels for the object "any rock concerts".

Analysis of generative models. In Appendix Figure 6a, we plot the correlation between error rate made by chatGPT for each affordance classes and the classwise annotator agreement. We observe a moderately negative correlation ($\rho = -0.29$) which suggests that there is a chance that the model is making higher mispredictions where the agreement is low. Similarly we observe that the mispredictions made by chatGPT for the most frequent objects has a moderately negative correlation ($\rho = -0.58$) with the annotator agreement. The correlation is shown in Figure 6b. The trends are similar for the other LLMs. These results together indicate that those objects and affordance classes

which are hard to disambiguate by humans also pose a challenge to the most sophisticated GenAI models in predicting the correct answer.

7 TEXT2AFFORD for physical reasoning

Apart from benchmarking LLMs and VLMs, we observe whether Text2Afford can be used as a source of affordance knowledge. We choose the physical commonsense reasoning as a target as the 'Object affordance' represents an innate physical property of an object, and we believe that any language model having strong affordance reasoning capability can enhance the physical reasoning capability. To explore this, we perform an 'instruction fine-tuning' on the TEXT2AFFORD dataset (although it is not meant for training) using few open-source LLMs (llama-3-8b-instruct, flan-t5), and test on two physical reasoning dataset - (1) PROST (Aroca-Ouellette et al., 2021), which contains 10 types of different physical properties of an object (including 6 affordance properties - rolling, breaking, stacking, grasping, sliding, bouncing) along with complex reasoning questions, and (2) PIQA (Bisk et al., 2019) which, focuses on selecting appropriate option given a situation that requires physical commonsense.

For PROST, using llama-3 the accuracy boosts from 0.36 to 0.42 after instruct fine-tuning with TEXT2AFFORD. Moreover, out of the 6 affordance properties from PROST, the accuracy got boosted for the reasoning of 5 affordance properties. For the PIQA, the same LLM gives a maximum of 4% accuracy boost. The full result is shown in Table 8. This suggest the generalizability of TEXT2AFFORD in physical reasoning tasks.

Model	Dataset			
	PROST		PIQA	
	Zero-shot	+TEXT2AFFORD	Zero-shot	+TEXT2AFFORD
Llama-3-8b	0.36	0.42(+.06)*	0.74	0.78(+.04)*
FLAN-T5-x1	0.13	0.16(+.03)	0.57	0.59(+.02)
FLAN-T5-xx1	0.34	0.38(+.04)*	0.72	0.75(+.03)

Table 8: Text-only physical reasoning dataset evaluation using different LLMs fine-tuned on TEXT2AFFORD. +TEXT2AFFORD: instruction fine-tuned on TEXT2AFFORD. * indicates p -value (< 0.05) using Mann-Whitney U-Test.

8 Additional details

Reason for choosing XNLI. We select XNLI to incorporate object references from less conventional and commonly explored scenarios. Unlike typical object identification datasets, XNLI offers sentences derived from novels, thus presenting a more in-the-wild textual context, which adds complexity and diversity to our dataset. Specifically,

we choose the hypothesis portion of the XNLI sentences due to its shorter context length. This choice intentionally poses a challenge to LLMs, allowing us to better evaluate their reasoning capabilities, especially when dealing with minimal contextual information.

Non-explicit mention about contextual object affordance in the instruction. The instructions shown in Appendix Figure 3 represent the initial guidance provided to annotators as an introduction to the task. Since understanding contextual object affordance can be challenging for non-expert annotators, this initial step was designed to give a basic idea of the task. However, we follow this up with a comprehensive training process and conduct two AMA (Ask Me Anything) sessions to ensure that annotators fully understood the need to base their judgments on the provided context. These efforts are key in ensuring high-quality annotations throughout the dataset creation.

Reason for choosing 0-3 Likert scale in data annotation. We opt for a 0-3 Likert scale (4-point) to minimize the potential for neutral or non-committal responses, which can often arise when a midpoint option is available. Our initial observations indicated that some annotators tended to select an “average” value without fully considering the contextual affordance of the objects, which diminished the depth of their evaluations and limited the discussion around ambiguities. By adopting a 4-point scale, we aim to encourage more decisive judgments. In addition, we provide a textbox (see Appendix Figure 3) for annotators to express any uncertainties or ambiguities they encountered, which has helped us in capturing more nuanced feedback.

Reason for choosing visual genome. We chose visual genome as a primary source for real images due to its rich, complex scenes, which are widely used in visual reasoning tasks. The complexity of the images in visual genome provides diverse contexts that align well with the goals of our study, which focuses on contextual object affordances. While other methods, such as using search engines like Bing, have been employed in prior work to retrieve images, we opt for visual genome to ensure that the images contain sufficient contextual and visual detail to support affordance prediction, even if there are minor limitations in reasoning.

Reason for choosing stable diffusion. Regarding the use of stable diffusion, we have been inspired by its demonstrated capability to generate high-quality, realistic images, particularly in prior studies where it was effective in reasoning tasks. While CLIP is primarily trained on real-world images, we hypothesize that stable diffusion could generate contextual images with sufficient accuracy to complement the real images. The generated images provide additional diversity, which helps us explore the affordance prediction task from a different angle. The benefit of using stable diffusion lies in its ability to create controlled, context-specific images that may not always be available in existing datasets, providing a broader range of testing scenarios for our models.

Reason for framing generative tasks as a binary decision problem. In the generative setting, we opt for a binary yes/no classification to evaluate the affordance of individual context-object-affordance triples. We decide this based on the observation of the tendency of smaller LLMs to hallucinate, which can make direct affordance prediction challenging, particularly in zero-shot scenarios. By framing it as a binary classification task, we aim to simplify the evaluation and obtain more reliable results. In addition, our approach allows for a comprehensive evaluation of both positive and negative affordances. This is critical for our dataset, as it is designed to assess affordances that are applicable, as well as those that are not, in a given context.

9 Conclusion

In this paper we introduced a novel text-based affordance dataset TEXT2AFFORD to investigate the affordance knowledge of PTLMs and pre-trained VLMs in different zero-shot settings. Our findings suggest that, the state-of-the-art language models, particularly text-to-text models, still exhibit limitations in their ability to reason about object affordances. In this seemingly easy task, we observe how context can introduce various levels of ambiguity and difficulty. We also observe, that even in the presence of such difficulty, human performance is superior and LLMs/VLMs still face difficulty in gaining such knowledge during their pretraining. Additionally, we observe how our dataset provides some additional knowledge that can be useful for physical commonsense reasoning – stressing its orthogonality more with respect to the pretraining knowledge LLMs and VLMs possess.

Acknowledgments

We would like to express our sincere gratitude to our co-authors for their invaluable contributions throughout this work. We also extend our thanks to the reviewers for their constructive feedback, which significantly helped improve the quality of the paper. Additionally, we gratefully acknowledge the support of the Toloka Research Grant program, which partially funded the data annotation process.

Limitations

All of our experiments were conducted for English language. The models may act differently in multilingual settings. Our dataset is curated based on a specific set of affordance classes, which may introduce bias in terms of affordance representation. This could limit the generalizability of our findings to other domains or contexts. Despite efforts to train annotators and ensure agreement, subjective interpretations of affordance classes, can introduce noise. Our study primarily relies on textual information for affordance prediction. The absence of grounded visual information may limit the model's ability to accurately predict affordances, as some affordances may be more visually dependent.

Ethics Statement

We used the publicly available XNLI corpus to curate our TEXT2AFFORD dataset. Our dataset does not contain any harmful or offensive contents. Any personal or sensitive information is anonymized and treated with utmost confidentiality. We ensure the protection of participants' privacy and obtain informed consent for data collection, annotation, and analysis. We incentivized all the annotators uniformly throughout the annotation process.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. [The falcon series of open language models](#).

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *Findings ACL-IJCNLP 2021*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *ACL*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#).

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

James J. Gibson. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. In *EMNLP*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming

- Xiong, and Dragomir Radev. 2022. [Folio: Natural language reasoning with first-order logic](#).
- Weinan He, Canming Huang, Yongmei Liu, and Xiaodan Zhu. 2021. WinoLogic: A zero-shot logic-based diagnostic dataset for Winograd Schema Challenge. In *Proceedings of the 2021 conference on empirical methods in natural language processing*.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. Hong Kong, China.
- Woojeong Jin, Tejas Srinivasan, Jesse Thomason, and Xiang Ren. 2024. [Winoviz: Probing visual properties of objects under different states](#).
- Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. Taxinli: Taking a ride up the nlu hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. [Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, IJCAI’20*.
- Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*.
- Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding.
- Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments.
- Michele Persiani and Thomas Hellström. 2019. Un-supervised inference of object affordance from text corpora.
- Ehsan Qasemi, Filip Ilievski, Muhao Chen, and Pedro Szekely. 2022a. PaCo: Preconditions attributed to commonsense knowledge.
- Ehsan Qasemi, Piyush Khanna, Qiang Ning, and Muhao Chen. 2022b. PInKS: Preconditioned commonsense inference with minimal supervision. Online only.
- Ehsan Qasemi, Amani R. Maina-Kilaas, Devadutta Dash, Khalid Alsaggaf, and Muhao Chen. 2023. [Preconditioned visual language inference with weak supervision](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. [Hierarchical text-conditional image generation with clip latents](#).
- 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.
- Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings EMNLP 2020*.

- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. [Photorealistic text-to-image diffusion models with deep language understanding](#).
- Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. Online.
- Yu Sun, Shaogang Ren, and Yun Lin. 2014. Object-object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496.
- Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olympics-on what language model pre-training captures. *TACL*, 8:743–758.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge.
- Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification.
- Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. NEWTON: Are large language models capable of physical reasoning? In *Findings EMNLP 2023*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on EMNLP*.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on EMNLP*.
- Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about object affordances in a knowledge base representation. In *ECCV*, pages 408–424. Springer.
- Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*.

Appendices

A Data annotation

A.1 Details of the TEXT2AFFORD dataset construction

Preprocessing. We select 20,000 sentences from a crowdsourced English dataset (XNLI English) (Conneau et al., 2018)⁶ and extract the noun phrases using the Stanford CoreNLP tool. As we restrict to the affordances that humans can directly perform, we filter the phrases which do not represent a tangible object (using ConceptNet). We manually filter out objects that cannot be acted upon directly by humans (such as school, building). After this preprocessing, we obtain a set of sentence-object pairs ($\langle x_i, o_i \rangle$), where the sentence acts as the context for the corresponding object. Each sentence on average has 2-3 such objects. We use the 15 predefined affordance classes from Zhu et al. (2014) to label each sentence-object pair for annotation.

We further expand our dataset with the labeled dataset provided by Zhu et al. (2014). Authors present 62 common objects and their corresponding 15 affordance labels. Given that our task is *context-based affordance* prediction, we require to have sentence-object pairs for labelling. To generate diverse context for this dataset, we utilize the ChatGPT UI⁷ model to generate synthetic sentences for each of the objects, followed by careful manual correction.

Pilot studies & annotator training. We annotate the dataset using the Toloka platform⁹. We design an interface on this platform, which contained clear instructions and examples for annotating the data. We conduct two rounds of pilot studies to analyze the subjective understanding of the annotators and, thereby, filter out the high quality, serious annotators. For the first pilot study, we present the annotators with the smaller 62 sentence-object pairs and ask them to label the instance with each affordance class on a scale of 0 to 3, indicating whether or not the affordance can be performed on the object. Here, 0-1 indicates that the affordance cannot be performed (high-low) and 2-3 indicates that the affordance can be performed low-high). We will

further use these 62 synthetic sentence-object pairs for few-shot training. For quality control, we select the top 90% of the available annotators in the platform, who are proficient in English, and use computers to complete the tasks¹⁰. A total of 15 annotators labelled the data, and all of them were incentivized uniformly. After the first pilot, we find that there is an extremely poor agreement among the annotators, and the overall precision is around 28%. Therefore, we moved on to a second pilot study. Here, we use all the 62 sentence-object pairs from the previous study, along with 32 randomly selected sentence-object pairs from the XNLI data. We use the top 30% of the annotators (based on the quality determined by the platform) available on the platform, while other criteria remained the same. We annotate 32 sentence-object pairs ourselves, and use all the labelled examples as *control* data points to guide the annotators while labelling. A total of 114 annotators (including the 14 annotators from the first pilot study) participated in this version of the pilot study. We assign a specific skill to the annotators who attained more than 30% precision and 30% recall. In total, 48 annotators passed this criteria. Through initial pilot studies, we learnt that without grounded images, the task appears quite subjective to annotators. The main goal of the pilot studies have been to understand the annotators' quality, their comprehension of the task, and their preferences for incentives per task. We have also conducted two additional AMA (Ask Me Anything) sessions with interested annotators to further clarify the task.

Final annotation. In the final phase, we conduct the annotation on a larger set of sentence-object pairs, carefully selecting a total of 2,368 pairs. To ensure diverse perspectives and minimize bias, we engage 25 skilled annotators in this phase. Three annotators independently annotated each of the sentence-object pairs. Each annotator meticulously evaluated the affordance classes for every pair, contributing to a comprehensive annotation of the dataset. We perform the annotations in phases and complete the full task over 10 phases.

Reason for multiple annotation phases. We intentionally consider relatively small number of data points for annotation in a single phase to make the review process easier. We carefully reviewed each annotation and provided feedback with guidance

⁶We choose XNLI as a source to facilitate multilingual extensions of our dataset.

⁷<https://chat.openai.com>

⁸Prompt used: Can you make realistic sentences with the following objects? Followed by the list of object names.

⁹<https://toloka.ai/>

¹⁰We exclude mobile-users as we believe the instructions may not appear clearly on mobile devices.

in case of mistakes. For instance, annotators initially got confused with the affordance ‘Watch’ as human can *watch* any visual objects. In another instance, some annotators asked whether ‘Throw’ can be valid affordance for the object ‘Kittens’ as humans can perform ‘Lift’, ‘Throw’ to the object ‘Kitten’. We discussed these types of ambiguities with the annotators after each phase. We measured class-wise agreement and average agreement across all classes after each annotation phase to ensure the quality of the annotations. The overall statistics for this *currently* constructed dataset – TEXT2AFFORD is in Table 3. Throughout the data processing pipeline, we put scrupulous attention to the quality control, including the use of pilot studies, iterative annotation refinement, and manual filtering. These measures ensure that the dataset is comprehensive, accurate, aligned with the objectives of the study and can be reliably reused in future. Overall, our TEXT2AFFORD dataset consists of 2368 sentence-object pairs having $\sim 100k$ annotations ($2368 \times 15 \times 3$).

A.2 Additional analysis on the datapoints by human

To further interpret the difficulty (or ambiguity) of the datapoints, we filter out the “sentence-object-affordance” triples based on the percentage annotator agreement. We categorize the triples into 3 sections:

Agreement > 0.75 : Total 26,411 triples
 $0.4 < \text{Agreement} < 0.75$: Total 7,084 triples
 Agreement < 0.4 : Total 2,025 triples

In general, the average agreement is higher for negative affordance classes than that of positive classes, which implies that it is easier for humans to tell which ‘affordance’ is not applicable to a particular object.

We employ three postgraduate students and provide them with the same set of instructions. We randomly sample 200 datapoints from the high agreement category (>0.75), and 200 samples from the low agreement category (<0.4) and ask to annotate independently. For the high agreement category scenario, we observe that in 86%, 87%, 91% of the cases their answers aligned with the majority voted answers. For the low agreement category, in most of the cases they feel there is not enough information in the context to answer about affordance. In some cases, it was easier to tell the affordance of

the object alone, but the context made it difficult to answer. For example:

Context: “SCR systems are primarily made from tree branches, lime and sawdust.” Can a human “Sit On” tree branches?

Without the context, it is easier to say “Yes”.

A.3 Comparison with other reasoning dataset

A.4 Instruction page on the Toloka platform

Figure 3 shows the guidelines/instructions, that the annotators had to follow for labelling.

A.5 Interface for labelling

A sample task interface is shown in Figure 4.

A.6 Annotators demographics

Figure 5 provides the demographic information about the annotators. We can observe that a large number of annotators (36%) are from Russia and most of the annotators having the age in between 20-35.

A.7 Phasewise annotator agreement

We plot the soft agreement¹¹, hard agreement¹² in Figure 7, which shows gradual increase in agreement scores.

A.8 Incentive details

During the pilot study, we provided USD 0.05 per task-suite where in each task-suite, there were 10 examples (15 affordance labels for each example) to be answered. We attempted to take feedback from the tolokors who had answered randomly (e.g., mark all the values as 0), to understand their requirements properly. Most of them suggested that a wage of \$0.1 to \$0.15 would be ideal for the survey.

During the main study we provided USD 0.25 per task-suite, where in each task-suite there were 5 examples to be answered. Some of them were consistently providing good answers and few of them also suggested improvement on the objects. We awarded them with an additional bonus of USD 0.5. Overall, we spent USD 777 for the annotation process.

¹¹Soft agreement: Mapping Likert scale ratings to binary labels for measuring agreement by applying a threshold value.

¹²Hard agreement: Treating each Likert scale rating as a distinct label.

Dataset	Train size	Dev size	Test size	Reasoning type	Source	Image-dependent	Targeted affordance	Publicly available
α NLI (Bhagavatula et al., 2020)	169,654	-	1,532	Abductive logical reasoning	Crowd-sourced	✗	✗	✓
α ARCT (Niven and Kao, 2019)	2420	632	888	Abductive logical reasoning	Crowd-sourced	✗	✗	✓
FOLIO (Han et al., 2022)	1004	204	227	Deductive logical reasoning	Expert written	✗	✗	✓
ANLI (Nie et al., 2020)	162,865	3,200	3,200	Deductive logical reasoning	Synthetic	✗	✗	✓
WinoLogic (He et al., 2021)	-	-	562	Deductive logical reasoning	Crowd-sourced	✗	✗	✓
LogiQA (Liu et al., 2021)	7,376	651	651	Mixed logical reasoning	Crowd-sourced	✗	✗	✓
LogiQA 2.0 (Liu et al., 2023a)	-	-	3,238	Mixed logical reasoning	Crowd-sourced	✗	✗	✓
PaCo (Qasemi et al., 2022a)	5,580	1,860	4,960	Preconditioned commonsense	Crowd-sourced	✗	✗	✓
δ -NLI(Rudinger et al., 2020)	36,999	3,329	3,512	Defeasible commonsense	Other dataset	✗	✗	✓
WINOVENTI (Do and Pavlick, 2021)	-	-	4,352	Commonsense with exceptions	Crowd-sourced	✗	✗	✓
PVLIR (Qasemi et al., 2023)	-	-	34,000	Preconditioned visual commonsense	Other dataset	✓	✗	✗
Normlense (Han et al., 2023)	-	-	10,000	Defeasible visual commonsense	Crowd-sourced	✓	✗	✓
WinoViz (Jin et al., 2024)	-	-	1,380	Reasoning object’s visual property	Crowd-sourced	✗	✗	✗
PROST (Aroca-Ouellette et al., 2021)	-	-	18,736	Reasoning object’s physical property	Other dataset	✗	✗	✓
NEWTON (Wang et al., 2023)	-	-	2,800	Reasoning object’s physical property	Crowd-sourced	✗	✗	✓
Persiani and Hellström (2019)	734,002	-	314,572	Object affordance without context	Synthetic	✗	✓	✗
TEXT2AFFORD (Ours)	-	-	35,520 (2368 * 15)	Contextual object affordance	Crowd-sourced	✗	✓	✓

Table 9: Comparison of TEXT2AFFORD with other reasoning datasets.

A.9 Correlation of affordances

In Figure 9 we show the correlation between the different affordance classes.

A.10 Most frequent objects

Figure 8a shows the most frequent 15 objects in the TEXT2AFFORD dataset.

B Experimental setup

B.1 Random baseline

In addition to evaluating generative models, we establish a random baseline. For this baseline, we randomly assign "yes" to the 15 affordance classes for each sentence-object pair, with random selections made from 0 to 9 (based on the observation that the maximum number of positive affordances per pair is 9). Interestingly, we find that models like Flan-T5-large and Flan-T5-XL underperform compared to this random baseline in zero-shot settings, highlighting the inherent difficulty of the task in such scenarios.

B.2 Zero-shot experiments

B.2.1 Commonsense reasoning tasks

To understand whether the injection of the common sense knowledge in the pre-trained models can enhance the performance of the affordance prediction, we first fine-tune the pre-trained models on common sense reasoning dataset such as PIQA (Bisk et al., 2019). Then we run the fine-tuned models on our dataset using the MLM setup. We use BERT-base, BERT-large, RoBERTa-large, and BART-large finetuned on MNLI.

NLI based approach. The NLI task considers a premise and a hypothesis as input pair $\langle p, h \rangle$, and

the models are trained to predict the probability whether the hypothesis is entailed by, contradicts or neutral with respect to the premise. Here we use the entailment probability from the models: $p_{L_a}(h|p) = p(l = \text{"ENTAILMENT"} | (p, h))$. This approach requires language models to be fine-tuned on premise-hypothesis pairs with the corresponding labels. Here we use RoBERTa-large and BART-large fine-tuned on the Multi-genre NLI (MNLI) corpus (Williams et al., 2018) consisting of 433k sentence pairs. For each sentence-object pair in our dataset as the premise, and use the hypothesis as " $\langle object \rangle$ can be used for $\langle affordance \rangle$ by human" for each object present in the sentence and 15 affordance classes. Using the NLI setting, we predict the entailment score for each affordance class for the given sentence-object pair. We use these scores for ranking the affordance classes and report mAP scores as well as accuracy.

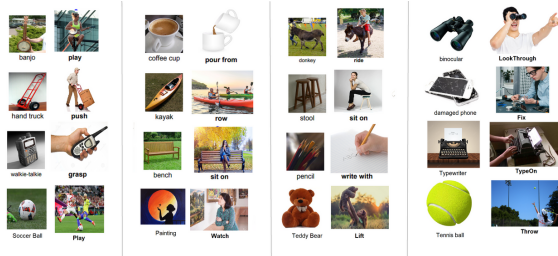
B.3 Ensemble language and vision prediction

Following Yang et al. (2022), we use the weighted sum as the late fusion over the final output probabilities of each affordance class from the language and multi-modal models. Before late fusion, we normalize the output probability scores from different models. We calculate the score as: $P_{ens}(y|x) = (1 - w)p_{L_a}(y|x) + wp_{V_I}(y|x)$ where w is the relative size of the vision-text model and the language model (following Yang et al. (2022)): $w = Sigmoid\left(\frac{\rho_{V_I}}{\rho_{L_a}}\right)$. Here ρ_{V_I} and ρ_{L_a} denote the number of parameters of the multi-modal and language models respectively.

Introduction

Mike has bought a Robot to do simple household tasks such as writing on a paper, playing a guitar, throwing garbage outside based on what Mike says to the Robot. However, the Robot is not accustomed with the Mike's household objects, so it does not know **which thing** can be used for **which of the tasks**. For example, the Robot is not aware that a pen or a pencil can be used for writing on a paper, but can not be played. A guitar or a banjo can be played, but not used for writing. This is important for the Robot to know before acting on instructions such as "clean the dishes for me". However, the good news is that the Robot can be taught about any object and its corresponding action. You, as a trainer, have been asked to teach the Robot about the household objects. Your task is simple -- there are few common objects (or things) in the house and you need to tell the Robot what actions (i.e. **tasks**) can be performed with each of those from a set of selected actions (tasks). This will help the Robot learn about what action can be performed on what type of objects.

See the below figure to understand which kind of action can be performed on which objects.



Task Description

You are given a **sentence** and the **object name** present in the sentence. You are required to mark the actions that can be performed from a given list of 15 actions.

For example:

Sentence: The tennis shoes have a range of prices.

Object: The tennis shoes

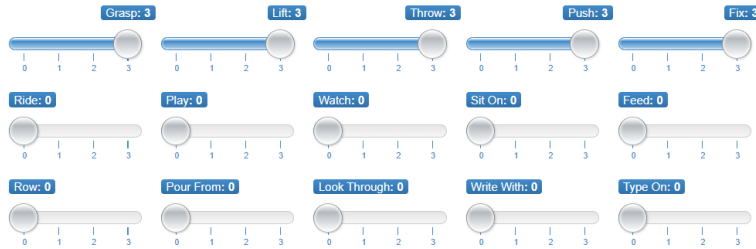
Out of the 15 given actions: **Grasp, Lift, Throw, Push, Fix, Ride, Play, Watch, SitOn, Feed, Row, PourFrom, LookThrough, WriteWith, TypeOn**. Select: **Grasp, Lift, Throw, Push, Fix** as that is something we typically do/is done/can be done with "The tennis shoe".

For each of the given actions, you are given a scale ranging from 0 to 3. The selection of a score of "0" means you strongly believe the action cannot be done, while a score of "3" means you strongly believe the action can be done. Scores of "1" and "2" are for cases where you are less sure about whether or not the action can be done. One example of selections is given below for the object "The tennis shoes"

Object:

The tennis shoes

Select the below actions:



Additional Examples:

- Objects that can be **grasped**: Pencil, tennis ball
- Objects that can be **Lift**: a book, a box, a chair
- Objects that can be **Thrown**: a baseball, a frisbee, a rock
- Objects that can be **Pushed**: table, brakes of a car
- Objects that can be **Fixed**: machines, vehicles, electronics
- Objects that can be **Ride**: bicycles, motorcycles, horses, roller coasters
- Objects that can be **Play**: musical instruments (guitar, piano, violin), sports equipment (tennis racket, soccer ball), electronic devices (video game console)
- Objects that can be **Watch**: televisions, computer screens, movie screens
- Objects that can be **SitOn**: chairs, benches, sofas
- Objects that can be **Feed**: animals such as dogs and cats, as well as birds
- Objects that can be **Row**: boats, canoes, kayaks, and rowboats
- Objects that can be **PourFrom**: a pitcher, a bottle, a jug, a teapot
- Objects that can be **looked through**: windows, telescopes, binoculars
- Objects that can be **WriteWith**: pens, pencils, markers
- Objects that can be **TypeOn**: computers, laptops, tablets, smartphones

Figure 3: The instruction used for annotators in the Toloka platform

B.4 Few-shot experiments

Training data To create few-shot training examples for fine-tuning encoder based PTLMs, we take all the 62 objects, and for each object we randomly select exactly 1 positive affordance class

(i.e., the class label annotated as 1) and 1 negative affordance class (i.e., the class label annotated as 0) for generating the training prompt. As this dataset does not contain any context sentences for a corresponding object, we use ChatGPT UI

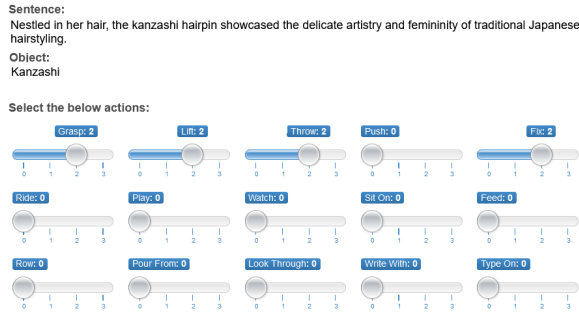


Figure 4: The sample task interface used for the annotators in the Toloka platform

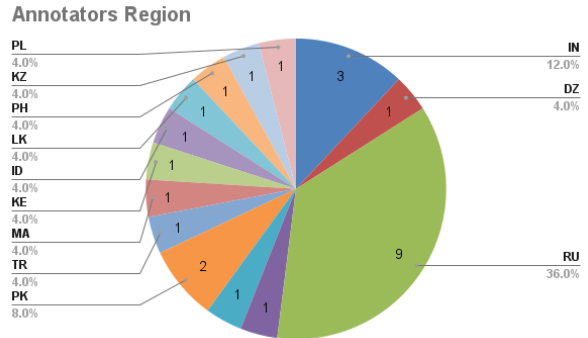
to generate the sentences for the corresponding objects and manually verify the sentences, so that it does not contain any invalid information. Finally, we have 62 sentence-object pairs and 2 classes (one positive and one negative) per pair, which we use to generate training examples. Each training example consists of a prompt and a label. They constitute 124 training examples (62 sentence-object pairs and 2 selected classes for each) for the few-shot experiment.

Selecting examples for in-context learning

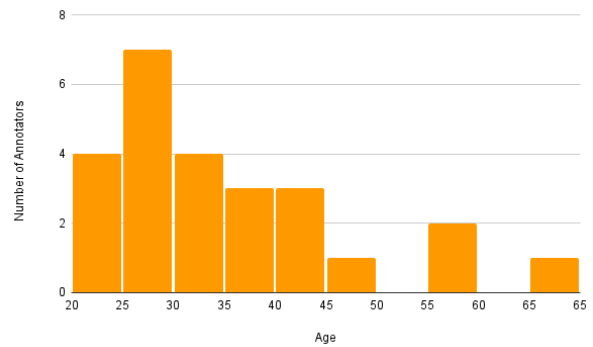
: We randomly sample five sentence-object-affordance triples from the above training data as the incontext demonstration examples in such a way that there should be k positive affordance classes. We vary the number of positive affordance classes $k \in \{1, 2, 3\}$ and report the average accuracy.

Experimental setup. We fine-tune the encoder based language models using the training data, and for the generative LLMs and the VLMs, we utilize the training data to select in-context demonstration examples.

Fine-tuning PTLM: We fine-tune the PTLMs in two different setups - NLI based and prompt based. For the NLI based setup we have the context sentence as premise and use same prompt (i.e., “<object> can be used for <affordance> by human”) which we use in the zero-shot settings as hypothesis. We use label as 1 for the positive affordance and label as 0 for the negative affordance. We use BERT-large-uncased, RoBERTa-large and BART-large for fine-tuning in this setup. We reuse these fine-tuned models for few-shot predictions in MLM setup. We use Adam optimizer with a learning rate of 2×10^{-5} . We fine-tune the model for 5 epochs for



(a) Country distribution of the annotators



(b) Age distributions of the annotators

Figure 5: The Annotators Demographics

each case.

In-context learning for generative models: We employ the same generative LLMs as well as VLMs to perform affordance prediction using *five* demonstration examples from the training data. We use the same text prompt as zero-shot setting and concatenate the five demonstration examples along with corresponding label (i.e., ‘YES’ for positive class, and ‘NO’ for the negative class) to the prompt and ask the LLMs and VLMs to predict the affordance. In case of the VLMs, we do not provide any additional image example here.

B.4.1 Multimodal task setup

Images contain necessary information about shape, texture, and size of objects that can be utilized to effectively predict an object affordance (such as the handle of the bucket can be used to grasp and lift). Hence, we also convert the problem into a multi-modal task by retrieving (or generating) a corresponding image from the context sentence, and predict the affordance of an object (mentioned in the sentence) based on the input.

Synthesizing images. In this setup, we use two different techniques to synthesize *semantically*

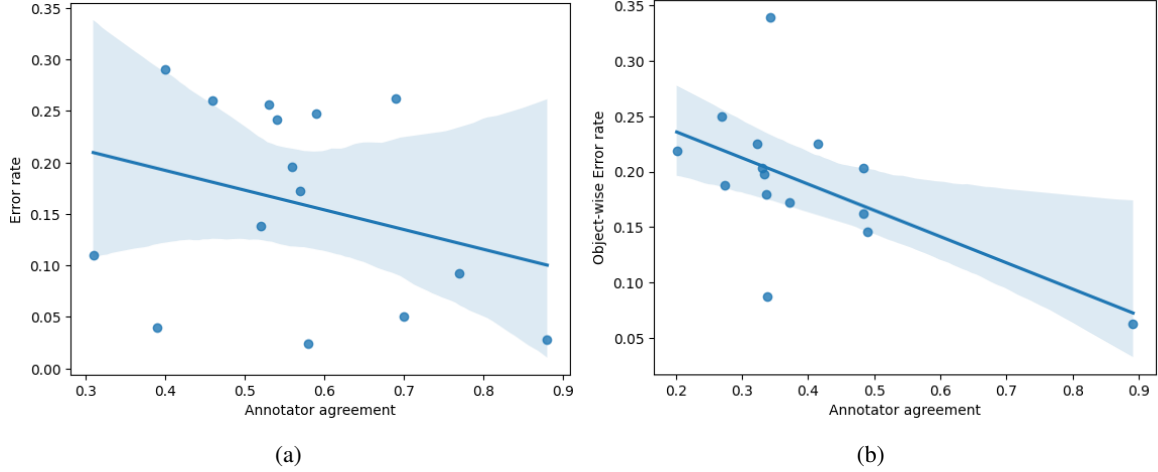


Figure 6: (a) Correlation between average classwise error rate made by chatGPT and the annotator agreement. ($\rho = -0.29$) (b) Correlation between frequent object wise error rate made by chatGPT and the annotator agreement. ($\rho = -0.58^*$). *indicates a p -value < 0.05 .

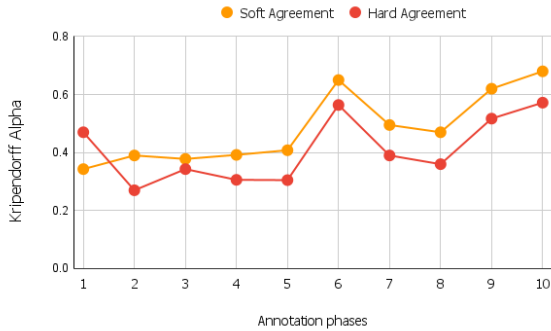


Figure 7: Phase-wise annotator agreement.

close images to corresponding context sentences using 1) retrieval and 2) generation. We further use top five images for both, to get an accurate estimation.

Retrieval based: We employ Visualgenome (Krishna et al., 2017) dataset, consisting of 108,077 images and 3.8 million object instances as the image database. We first encode the images using multi-modal CLIP (Radford et al., 2021) based sentence-transformers architecture, and index those image embeddings using Approximate Nearest Neighbour search (ANN)¹³, for making the search efficient. Now, for each sentence, we search for top five images from the database to be used further.

Generation based: Recently, the multi-modal generative models (Ramesh et al., 2022; Saharia et al., 2022) have shown incredibly good performance for text based image generation tasks. We adopt

the recent *StableDiffusion* (Rombach et al., 2022) model to generate top five images based on the sentence as a text prompt.

We use the top five retrieved images by using retrieval and generation methods each. We use *CLIP* (Radford et al., 2021) and *ViLT* (Kim et al., 2021) as our vision-text models. CLIP is pre-trained on 400M image-caption pairs with the contrastive learning strategy. CLIP has a text encoder f_T and a visual encoder f_V , which can project text and image into the shared latent space. We aggregate the k ($=5$) corresponding images and use CLIP to compute the relevance score of (x, y) : $Score_{VI}(x, y) = \frac{1}{k} \sum_{i=1}^k \cos(f_T(x), f_V(I_y^k))$, where I_y^k is the k^{th} image for the input text y . In the ViLT model we provide the text prompt along with the representative images as input to predict the masked token. We use the same prompt as the previous MLM task (i.e., “<Object> can be used for <MASK_TOKEN> by human.”) and get the probability of each affordance class as the logit corresponding to the <MASK_TOKEN>.

Text generation based. Similar to section 5.1.1, we utilize state-of-the-art VLMs to make predictions regarding object affordances. We provide with a ‘YES\NO’ question answering based text prompt along with the aligned images as input to the VLMs, and the model should generate an answer whether a particular affordance can be performed on the given object. We use state-of-the-art VLMs such as IDEFICS (Laurençon et al., 2023), LLaVA (Liu et al., 2023b), InstructBLIP (Dai et al.,

¹³<https://pypi.org/project/annoy/>

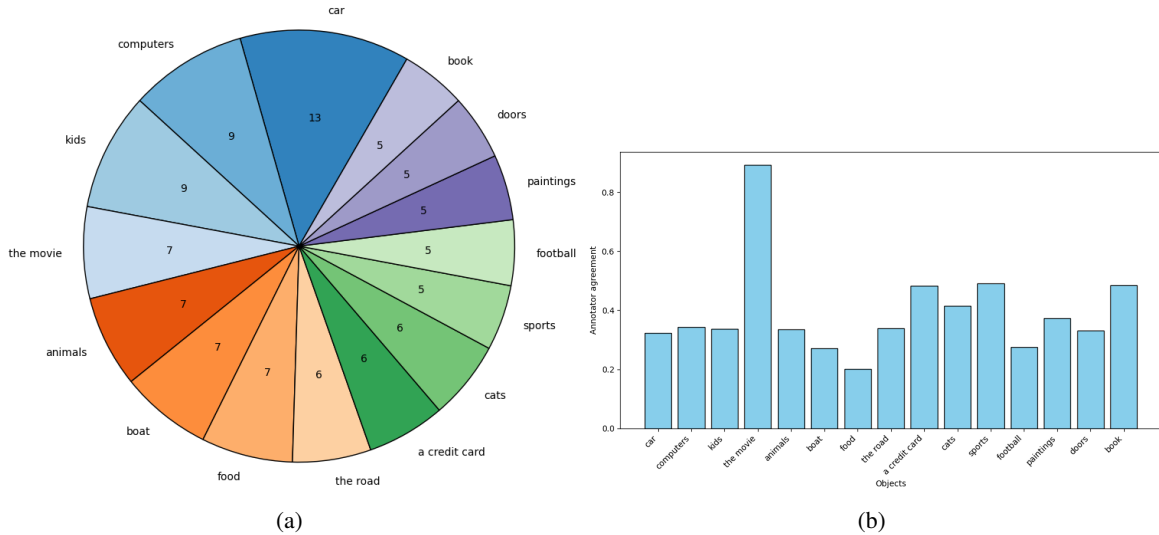


Figure 8: (a) Most frequent 15 objects and their corresponding frequency in the TEXT2AFFORD dataset. (b) Annotator agreement for the most frequent 15 objects.

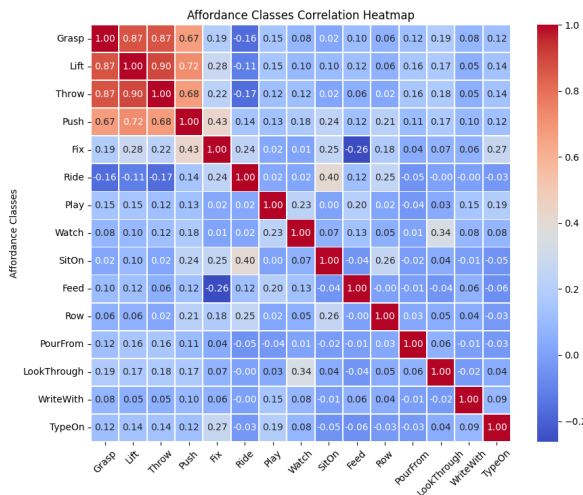


Figure 9: Correlation between each of the affordance classes.

2023) for this task. The text prompt used for the models can be found in the Appendix D, Table 10.

C Additional (mis)prediction analysis

C.1 Affordance classwise mis-prediction

We analyze the mis-prediction rates for each class using the best LLMs (chatGPT, llama-3-8b). We observe that, the classwise mis-prediction rate is similar to the distribution of each class in the original data, i.e., the classes such as ‘grasp’, ‘lift’ having higher mis-predictions compared to ‘typeOn’, ‘row’.

C.2 Objects with multiple positive affordances

We conduct an analysis to determine whether the frequency of positive affordances for an object impacts model accuracy. Our findings indicate that the accuracy is highest when an object has a single positive affordance. Beyond this point, the number of positive affordances does not significantly influence the model’s performance. Specifically, we observe that as the number of positive affordances increases, the accuracy fluctuates without a clear pattern, suggesting that additional positive affordances do not contribute to a consistent improvement or decline in model accuracy.

C.3 Correlation of ChatGPT accuracy and average human agreement

We provide the figures corresponding to the generative model analysis in Figure 6.

D Prompt selection

We use intuitive prompts for each of the setups, which are suitable for affordance related to object.

E Instruction fine-tuning setup

Data sample selection. We select sentence-object pairs from the TEXT2AFFORD dataset where at least one positive affordance is present. For each selected sentence-object pair, we randomly assign one positive affordance and one negative affordance, yielding a balanced dataset of 1819 training instances (positive and negative classes). To incorporate additional domain knowledge and

Model	Prompt used
FLAN-T5	consider {sentence}. Now, from this information can human {affordance} the {object_name}? Answer YES or NO:
Falcon	""You are a helpful AI assistant. Answer only "YES" or "NO" for the question based on the given context. Context:sentence \n »QUESTION« Can human {affordance} the {object_name}? \n »ANSWER«"" <code>.strip()</code>
I-BLIP, IDEFICS, LLaVA	consider the sentence {sentence}. Now from this information, can human {affordance} the {object_name}? Accompanying this query is an image of the object_name. Note that the image may contain noise or variations in appearance. Given the textual description and the image, answer YES or NO whether the human can {affordance} the {object_name}. Answer: "

Table 10: Prompt format used by different models for the prediction. I-BLIP: InstructBLIP.

reduce the likelihood of generating hallucinated answers, we include 500 randomly sampled instances from the training set of the target task (i.e., PIQA). For the PROST task, as the training set is not explicitly available, we sample from the test set and ensure these samples are removed from the evaluation set during testing. The training instances are framed in a multiple-choice question answering format.

Fine-tuning setup. We utilize Alpaca-formatted prompts (shown in Table 11, Table 12 and Table 13 for the TEXT2AFFORD, PIQA and PROST tasks, respectively). We fine-tune 4-bit quantized models with PEFT, focusing on the adapter layers. We perform the fine-tuning over 5 epochs with a batch size of 8, a learning rate of $2e-10$, weight decay, and a maximum sequence length of 256.

F Model implementation details

The language models and the ViLT are built on top of the huggingface API¹⁴. For NLI based zero-shot prediction, we use the zero-shot classification pipeline¹⁵. We adapted the CLIP model from the OpenAI’s public repo¹⁶, and we select the ViT/B32 as the image encoder. For ViLT, we select the vilt-b32-mlm¹⁷ model. For generative LLMs and VLMs we apply the models available on huggingface¹⁸. All the experiments were conducted on 2x NVIDIA RTX 4090 GPU server.

¹⁴<https://huggingface.co/>

¹⁵https://huggingface.co/docs/transformers/main_classes/pipelines

¹⁶<https://github.com/openai/CLIP>

¹⁷[dandelin/vilt-b32-mlm](https://github.com/dandelin/vilt-b32-mlm)

¹⁸<https://huggingface.co/models>

G Details of evaluation metric

For a ‘Sentence-Object’ pair we calculate accuracy in the following way. In the ground-truth, each affordance class is treated as a binary value, where a value of 1 represents a ‘positive affordance’ indicating that the affordance can be performed on the object, and a value of 0 represents a ‘negative affordance’ indicating that the affordance cannot be performed. Now, for a particular ‘Sentence-Object’ pair, let’s assume there are two positive affordances (P1, P2) in the ground truth; then there will be 13 negative affordances (as we have a total 15 affordance classes). In case of encoder-based models, for each positive affordance, we compare its prediction score against each negative affordance’s score. If a positive affordance’s score is higher, we increase the Correct count; otherwise, the Wrong count. Accuracy is calculated as $\text{Correct} / (\text{Correct} + \text{Wrong})$.

In case of encoder-decoder or decoder-only models, Due to the inherent difficulty in automatic evaluation, we predict ‘YES\NO’ for each affordance class, mapping ‘YES’ to 1 and ‘NO’ to 0. Accuracy is then measured in the same way as for encoder-based models (assuming 1 or 0 as the score for each affordance class).

H Dataset creation time

Annotating affordances about the object from a text itself is a difficult and very subjective task. It took approximately 5 months for completing the extraction of noun-phrases from xnli data, filtering objects, selecting skillful tolokors and training, and then final phase-wise annotation after rigorous review process.

I Sample dataset

Figure 10 shows a sample of TEXT2AFFORD dataset

J Additional experiments

J.1 Qualitative analysis of generated images

We conducted a qualitative analysis on 50 randomly sampled objects and their corresponding generated images. Two annotators (one Phd student and one undergrad student) marked each of the 5 generated images as 1 or 0 according to their relevance and non-relevance to the object respectively. We considered the image as relevant if both of the annotators marked that image as 1. We achieved an

Instruction to fine-tune TEXT2AFFORD

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:

You are an AI assistant that has strong reasoning capability. You are given a context containing an object, and you are asked to answer a question about the object based on the context. Just response 'Yes' or 'No'.

Context:

{context}

Object:

{object}

Question:

Can human {affordance} the {object}?

Answer:

{answer}

Table 11: Instruction to fine-tune TEXT2AFFORD.

Instruction to fine-tune PIQA

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:

You are an AI assistant that has strong reasoning capability. You are given a situation and asked to choose the most appropriate option from given two options.

Situation:

{situation}

Options:

[0] {option0}

[1] {option1}

Only response the 'answer id'. For example if the answer is [0] then response 0. DO NOT respond anything other than <0, 1>.

Answer:

{answer}

Table 12: Instruction to fine-tune PIQA.

Instruction to fine-tune PROST

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:

You are an AI assistant that has strong reasoning capability. You are given a question with 4 options and you have to choose the right option.

Question:

{question}

Options:

[0] {option_A}

[1] {option_B}

[2] {option_C}

[3] {option_D}

Only response the 'answer id'. For example if the answer is [0] then response 0. DO NOT respond anything other than <0, 1, 2, 3>.

Answer:

{answer}

Table 13: Instruction to fine-tune PROST.

Sentence	Object	Grasp	Lift	Throw	Push	Fix	Ride	Play	Watch	SitOn	Feed	Row	PourFrom	LookThrough	WriteWith	TypeOn
This diablo only comes out to slaughter the cattle.	cattle	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
Delivery points should include at least a bench and a locked storage compartment.	bench	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0
There are four fences, and you can only go past the second one if you are a member of the imperial family, or a high-ranking priest.	fences	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0
Users are excited about being able to share their own events on the calendar page.	calendar page	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0
While ran towards where the people were hitting each other with swords.	swords	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
The cat ate every kind of fish except tuna.	fish	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
The snake was hissing underneath the deck.	deck	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
On the higher levels of the town hall, Umbrian and Tuscan paintings are on show.	the town hall	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
He couldn't follow up because his mouth was gagged by a group of mercenaries.	mercenaries	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
A gristle gun is featured.	gristle gun	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10: Example snapshot of TEXT2AFFORD dataset.

Acc@1 of 0.2, Acc@5 of 0.88 and an MAP@5 of 0.36. Which suggests that in most of the cases there are relevant images in the top-5 generated images. In our pursuit of assessing the statistical significance of our sampled data (i.e., the 50 examples), we embarked upon a rigorous hypothesis testing procedure utilizing the binomial distribution. Within our specific context, we accorded greater significance to the top-5 accuracy metric, which demonstrated an impressive achievement of 0.88. This signifies that among the 50 selected examples, in 44 instances, at least one of the five generated images displayed relevance to the object under consideration.

Guided by this success rate, we proceeded to conduct a meticulous hypothesis test employing the binomial distribution. We assumed an expectation of success at 0.75. The outcome of this statistical analysis revealed a p-value of less than 0.02, thereby underscoring the statistical significance of our success rate.