PROST: Physical Reasoning about Objects through Space and Time

Stéphane Aroca-Ouellette*, **Cory Paik**, **Alessandro Roncone** and **Katharina Kann**
University of Colorado Boulder
firstname.lastname@colorado.edu

Abstract

We present a new probing dataset named *PROST: Physical Reasoning about Objects Through Space and Time*. This dataset contains 18,736 multiple-choice questions made from 14 manually curated templates, covering 10 physical reasoning concepts. All questions are designed to probe both causal and masked language models in a zero-shot setting. We conduct an extensive analysis which demonstrates that state-of-the-art pretrained models are inadequate at physical reasoning: they are influenced by the order in which answer options are presented to them, they struggle when the superlative in a question is inverted (e.g., *most* ↔ *least*), and increasing the amount of pretraining data and parameters only yields minimal improvements. These results provide support for the hypothesis that current pretrained models' ability to reason about physical interactions is inherently limited by a lack of real world experience. By highlighting these limitations, we hope to motivate the development of models with a human-like understanding of the physical world.

1 Introduction

In the context of natural language processing (NLP), Bender and Koller (2020) provides a working definition of "understanding" as the ability to recover the communicative intent from an utterance. To achieve this, one must be able to query a set of concepts that is aligned with the speaker's own understanding. An example of such alignment is our interaction with the physical world. This experience, shared by all humans, provides a common set of concepts to rely on in communication. For example, the reader can map the phrase *I dropped my pint glass* to a set of relevant experiences and generate a mental depiction of the scene. Further, the reader can also use their knowledge of gravity and the properties of a pint glass to reason about potential outcomes: the pint glass will fall toward the ground and will likely break on impact.

A person drops a *glass*, a *pillow*, a *coin*, and a *pen* from a balcony. The [MASK] is most likely to break.
A) glass **B)** pillow **C)** coin **D)** pen

Figure 1: An example question from PROST.

Children grab, push, and play with the objects around them to form concepts about the world they live in even before learning to talk (Hespos and Spelke, 2004). These concepts are then linked with words to enable communication, eventually providing the necessary grounds for concepts and language to co-develop (Bloom, 2002; Gelman, 2009). In contrast, current language models (LMs) are not exposed to real-world experiences, making them incapable of grounding language (Bisk et al., 2020a). We hypothesize that this lack of experience impedes their ability to both understand an utterance relating to the physical world and their ability to reason about its implications.

In order to investigate our hypothesis, we create *PROST: Physical Reasoning of Objects Through Space and Time*, a probing dataset to evaluate the ability of pretrained LMs to understand and reason about the physical world. PROST consists of multiple-choice cloze-style questions covering 10 basic concepts: direction, mass, height, circumference, stackable, rollable, graspable, breakable, slideable, and bounceable. Importantly, PROST is designed to avoid models succeeding in unintended

*Email has no accent, but includes the hyphen

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4597–4608
August 1–6, 2021. ©2021 Association for Computational Linguistics

ways. First, PROST provides no training data, so as to probe models in a zero-shot fashion. This prevents models from succeeding through spurious correlations between training and test data and encourages success through a true understanding of and reasoning about the concepts at hand. Second, we manually write templates for all questions in an effort to prevent models from having seen the exact same sentences in their training data. Finally, it focuses on a small set of well defined, objective concepts that only require a small vocabulary. This allows researchers to focus more on the quality of training data rather than the size of it.

Contributions We make two contributions: 1) We introduce PROST, a dataset with 18,736 cloze-style questions created from 14 manually written templates, covering 10 physical reasoning tasks. 2) We conduct an extensive analysis which demonstrates that state-of-the-art pretrained models are inadequate at physical reasoning. More specifically, they are influenced by the order in which answer options are presented to them, they struggle when the superlative in a question is inverted (e.g., *most* ↔ *least*), and increasing the amount of pretraining data and parameters only yields minimal improvements. The dataset and code is available at github.com/nala-cub/prost.

2 Related Work

Evaluation of Reasoning Abilities As pretrained models are excelling on many NLP tasks, more work is being done on understanding their abilities. A subset of this work focuses on physical reasoning. PIQA (Bisk et al., 2020b) tests physical commonsense, with concepts ranging from hard shell tacos to separating egg yolks. In order to succeed on PIQA through reasoning, a model would need to be able to understand thousands of human experiences. In contrast, PROST provides a first step towards grounded understanding and reasoning by focusing on a few simple concepts. Bakhtin et al. (2019) provides a set of 2D puzzles that involve placing a new object in a scene to accomplish a goal. This research also focuses on simple physics, however there is no language component. Clark et al. (2018) and Kembhavi et al. (2017) both provide a large set of grade school multiple-choice questions, including some that could be solved with reasoning. However both provide corresponding material where the solution can be found, relying more on information retrieval than a general understanding and reasoning about the world.

Another set of reasoning-based benchmarks focuses on common sense reasoning. SWAG and its extension hellaSWAG evaluate commonsense natural language inference (Zellers et al., 2018, 2019). Sap et al. (2019) tests commonsense reasoning about social situations. However, commonsense reasoning is often subjective and requires understanding of complex human–human interactions involving social and societal norms. In contrast, physical reasoning is based on objective and well defined constructs.

Other datasets (Forbes and Choi, 2017; Elazar et al., 2019; Goel et al., 2019) focus on object–attribute comparison. However, they compare concepts at a word level rather than sentence level and use a large training set to create an engineered object–attribute comparison model. It is difficult to see how these models could generalize to other forms of reasoning.

Moreover, all the above datasets follow a pretraining-agnostic identically distributed (PAID) paradigm (Linzen, 2020), making them susceptible to models that can leverage unintended correlations between the training and test sets.

Zero-Shot LM Probes Similar to PROST, several recent benchmarks have circumvented the concern of identically distributed training and test sets by probing models in a zero-shot manner. Petroni et al. (2019) queries masked LMs (MLMs) for factual knowledge using templates in the format of *Dante was born in [MASK]*. Talmor et al. (2020) use a similar format to probe six concepts ranging from age comparison to taxonomy conjunction. Ettinger (2020) uses this format to show that BERT robustly retrieves hypernyms, but fails to understand negation. Lin et al. (2020) probe numerical commensense in both MLMs and traditional LMs. Warstadt et al. (2020) measures traditional LMs' sense of grammatical acceptability by comparing sentence probabilities.

Grounded Language Environments PROST investigates if pretrained models show a lack of understanding of the physical world which could result from learning language without grounding. While not used for pretraining, a number of multi-modal environments have been developed to ground language. Shridhar et al. (2020)'s ALFRED builds on other vision-and-language navigation environments (Gordon et al., 2018; Reg-

neri et al., 2013; Zhu et al., 2017; Anderson et al., 2018), and enables grounding of language instruction to actions, behaviours, and objects. BABYAI (Chevalier-Boisvert et al., 2019) and BABYAI++ (Cao et al., 2020) provide an environment to ground simple language in a gridworld. Additionally, other work has explored grounding language in simulations or the real world (Hill et al., 2020; Lynch and Sermanet, 2020). While they provide important resources to ground language, little emphasis is placed on the language modules themselves. They are often trained tabulae rasae, learning language for a singular purpose and missing out on the syntax and coverage learnt during pretraining;[1] language is only ever an input, and no analysis has been done on how language understanding evolves as the agent learns to succeed on different tasks.

3 PROST

PROST consists of $18,736$ cloze-style multiple-choice questions designed for probing a LM's physical reasoning ability. They cover 10 basic concepts: direction, mass, height, circumference, stackable, rollable, graspable, breakable, slideable, and bounceable. We choose these concepts because they are well defined, easily learned by interacting with the world, and are useful concepts for any embodied agent. The questions are constructed from 14 manually written templates. Each template follows one of three different formats: the first format is specific to the set of questions pertaining to directions; the second format is used to gauge the relative attributes—specifically mass, height, and circumference—of objects; and the third format targets the affordances of objects—specifically whether an object is stackable, rollable, graspable, or breakable, and whether a surfaces is slideable or bounceable[2]. We use CheckList (Ribeiro et al., 2020) to obtain the questions from our templates. We show all templates in Table 1 and explain them in detail below. We end this section by describing the objects featured in PROST.

Direction Templates We use two templates to generate questions which probe understanding of direction. The first focuses on cardinal directions. The second uses a set of four manually crafted questions to probe understanding of how gravity affects the directions of a ball throughout its trajectory. Due to their similarity, we count these four questions as a single template. The direction templates create a total of 16 questions.

Attribute Templates The second set of templates probe the models' ability to reason about relative mass, height, and circumference of common objects. For each of these three concepts we create a set of six objects that are easily ordered by their respective attributes. A context is first presented with up to four of the six objects to prime the models with the range of possible choices. This is followed by a prompt that probes the model to select one of the objects based on the object's mass, height, or circumference. By inverting the superlative in the prompt (e.g., *longest* $\leftrightarrow$ *shortest*), we can probe the model's ability to identify both the object with the highest attribute value and the object with the lowest attribute value from the set of choices. We permute through all objects and all orders. Each of the three attributes are tested using two templates that share the same set of objects. Each template produces ${}_6P_4 * 2 = 720$ questions, meaning each attribute is probed using 1440 questions.

Affordance templates The remaining templates target an understanding of object affordances. For each affordance—stackable, rollable, graspable, breakable, slideable, and bounceable— we collect a set of five objects with and five objects without that affordance. Again, we first provide a short context that contains each of the four possible objects. We then provide a prompt that requires the model to select the only object either with or without the affordance. We include all permutations of objects where there is exactly one correct answer. These templates produce ${}_5P_1 * {}_5P_3 * 4 * 2 = 2400$ questions for each of the six affordances.

Objects in PROST All possible values for the placeholders in our templates are shown in Table 3. For affordances, we display objects in two groups: those with and without each affordance. For attributes, objects are sorted by increasing order, e.g., for mass, *leaf* is the lightest object and *microwave* is the heaviest object. Each object in PROST is selected to be single-token compatible for a wide range of vocabularies to enable easy probing of MLMs. We validate the order of our attribute objects and the group membership for our affordance objects by collecting judgments from 9 human val-

[1] An exception is Lynch and Sermanet (2020), which incorporates modern LMs and provides impressive generalizability. However, they too only use language as an input and do not analyze how language understanding evolves.

[2] *Bounceable* here refers to providing an elastic collision.

Category	Qs.	Template
Directs. 1	12	**C:** A person is walking *{north/east/south/west}*. They turn *{left/right/around}*. **Q:** They are now walking [MASK]. **O:** **A)** north **B)** east **C)** south **D)** west
Directs. 2a	1	**C:** A person drops a ball. **Q:** Immediately after leaving the person's hand, the ball is moving toward the [MASK].
Directs. 2b	1	**C:** A person throws a ball straight into the air. **Q:** Immediately after leaving the person's hand, the ball is moving toward the [MASK].
Directs. 2c	1	**C:** A person throws a ball straight into the air. **Q:** Immediately after reaching the highest point in it's trajectory, the ball is moving toward the [MASK].
Directs. 2d	1	**C:** A person drops a ball. The ball then bounces off the ground. **Q:** Immediately after bouncing off the ground, the ball is moving toward the [MASK]. **O:** **A)** ground **B)** sky **C)** left **D)** right
Mass 1	720	**C:** A(n) *{mass_obj1}*, a(n) *{mass_obj2}*, a(n) *{mass_obj3}*, and a(n) *{mass_obj4}* moving at identical speeds each collide with a static hockey puck. **Q:** The puck hit by the [MASK] slides the *{shortest/longest}* distance.
Mass 2	720	**C:** A(n) *{mass_obj1}* and a(n) *{mass_obj2}* are placed on either end of a perfectly balanced seesaw. **Q:** The side of the seesaw with the [MASK] moves *{up/down}*. **O:** **A)** *{mass_obj1}* **B)** *{mass_obj2}* **C)** *{mass_obj3}* **D)** *{mass_obj4}*
Height 1	720	**C:** Four balls are dropped. The first is dropped from the height equivalent of a *{height_obj1}*, the second is dropped from the height equivalent of a *{height_obj2}*, the third is dropped from the height equivalent of a *{height_obj3}*, and the fourth is dropped from the height equivalent of a *{height_obj4}*. **Q:** The ball dropped from the height of the [MASK] takes the *{longest/shortest}* amount of time to fall.
Height 2	720	**C:** There are four staircases. The first staircase leads to the top of a *{height_obj1.}*, the second staircase leads to the top of a *{height_obj2.}*, the third staircase leads to the top of a *{height_obj3.}*, and the fourth staircase leads to the top of a *{height_obj4.}*. **Q:** The staircase leading to the top of the [MASK] is the easiest/hardest to walk up. **O:** **A)** *{height_obj1}* **B)** *{height_obj2}* **C)** *{height_obj3}* **D)** *{height_obj4}*
Circumf. 1	720	**C:** Four people are walking at identical speeds. The first walks around a *{circ_obj1}*, the second walks around a *{circ_obj2}*, the third walks around a *{circ_obj3}*, and the fourth walks around a *{circ_obj4}*. **Q:** The [MASK] takes the *{longest/shortest}* amount of time to walk around.
Circumf. 2	720	**C:** A person paints a circle around a *{circ_obj1}*, a *{circ_obj1}*, a *{circ_obj1}*, and a *{circ_obj1}*. **Q:** The circle around the [MASK] takes the *{most/least}* amount of paint. **O:** **A)** *{circ_obj1}* **B)** *{circ_obj2}* **C)** *{circ_obj3}* **D)** *{circ_obj4}*
Stackable	2400	**C:** A person is trying to stack *{stack}*, *{no stack1}*, *{no stack2}*, and *{no stack3}*. **Q:** The [MASK] are the *{easiest/hardest}* to stack. **O:** **A)** *{stack}* **B)** *{no stack1}* **C)** *{no stack2}* **D)** *{no stack3}*
Rollable	2400	**C:** A person is trying to roll a(n) *{roll}*, a(n) *{no roll1}*, a(n) *{no roll2}*, and a(n) *{no roll3}*. **Q:** The [MASK] is the *{easiest/hardest}* to roll. **O:** **A)** *{roll}* **B)** *{no roll1}* **C)** *{no roll2}* **D)** *{no roll3}*
Graspable	2400	**C:** A person is trying to move a pile of *{break}*, a pile of *{no break1}*, a pile of *{no break2}*, and a pile of *{no break3}* from one side of a room to the other using only one hand. **Q:** The [MASK] is the *{most/least}* likely to break. **O:** **A)** *{break}* **B)** *{no break1}* **C)** *{no break2}* **D)** *{no break3}*
Breakable	2400	**C:** A person drops a *{break}*, a *{no break1}*, a *{no break2}*, and a *{no break3}* from a balcony. **Q:** The [MASK] is the *{most/least}* likely to break. **O:** **A)** *{grasp}* **B)** *{no grasp1}* **C)** *{no grasp2}* **D)** *{no grasp3}*
Slideable	2400	**C:** A person is sliding four bricks across four hard surfaces. The first surface is covered with *{slide}*, the second surface is covered with *{no slide1}*, the third surface is covered with *{no slide2}*, and the fourth surface is covered with *{no slide3}*. **Q:** The surface covered with [MASK] is the *{hardest/easiest}* for the brick to slide across. **O:** **A)** *{slide}* **B)** *{no slide1}* **C)** *{no slide2}* **D)** *{no slide3}*
Bounceable	2400	**C:** A person is trying to bounce a rubber ball. They drop a first ball onto *{bounce}*, a second ball onto *{no bounce1}*, a third ball onto *{no bounce2}*, and a fourth ball onto *{no bounce3}*. **Q:** The ball dropped onto[MASK] bounces the *{most/fewest}* times. **O:** **A)** *{bounce}* **B)** *{no bounce1}* **C)** *{no bounce2}* **D)** *{no bounce3}*

Table 1: All templates in PROST. **C:** = Context, **Q:** = Question, **O:** = Options. {} indicate placeholders. The objects can be found in Table 3. The rest of the placeholders show their possibilities in the braces themselves. [MASK] indicates the position of the blank that the models need to fill. See Section 3 for more information.
NOTE: The number of objects with and without the affordances are swapped when the superlative is inverted.

Model Type	Input	Target
Decoder	*(context)* They are now walking ⟨**O**⟩.	Max probability for each sentence input to the model
Encoder	*(context)* They are now walking ⟨**M**⟩.	Max probability for the masked token.
T5	*(context)* They are now walking ⟨**X**⟩.	⟨**X**⟩ south ⟨**Y**⟩
UnifiedQA	Which way are they walking now? \n (A) north (B) south (C) east (D) west \n *(context)*	south

Table 2: Overview of the task preprocessing for different architectures evaluated. In all methods, the context remains unchanged and is "A person is walking west. They turn left."

idators. The validators obtained a 100% agreement on the object ordering, and 94.6% agreement on the object group membership.

Attributes	
Attribute	*Objects*
mass	leaf, coin, egg, apple, brick, microwave
height	book, microwave, table, car, house, mountain
circ	book, microwave, table, car, house, mountain
Affordances	
Affordance	*Objects*
stack	books, blocks, boxes, coins, plates
no stack	balls, bottles, eggs, flowers, lamps
roll	apple, ball, bottle, egg, can
no roll	book, box, block, mirror, microwave
grasp	balls, blocks, books, bottles, flowers
no grasp	flour, rice, salt, snow, sugar
break	bottle, egg, glass, mirror, plate
no break	ball, coin, pen, pillow, shirt
slide	ice, frost, grease, oil, soap
no slide	carpet, concrete, grass, gravel, rubber
bounce	asphalt, brick, concrete, rubber, steel
no bounce	carpet, foam, grass, leaves, snow

Table 3: Objects used in the templates.

4 Models

Using PROST, we probe three types of transformer-based models (Vaswani et al., 2017): decoder models, encoder models, and encoder-decoder models. Each model has slightly different formatting requirements, which we show in Table 2. For each model type, we probe a range of different sizes to investigate the effects of scaling. We use Huggingface's (Wolf et al., 2020) pretrained models, see Table 4 for the full set.

Decoder Models We analyze OpenAI's GPT-1 (Radford et al., 2018) and GPT-2 (Radford et al., 2019). Both are based on a transformer decoder architecture and trained on a traditional language modeling objective. We run these models over

Model		**Params (M)**	**Data (GB)**
GPT		116.5	2
GPT-2	B	124.4	40
	M	354.8	40
	L	774.0	40
	XL	1557.6	40
BERT	B	110.1	13
	L	336.2	13
RoBERTa	B	124.7	160
	L	355.4	160
ALBERT V2	B	11.8	160
	L	17.8	160
	XL	59.0	160
	XXL	223.2	160
T5	S	60.5	170
	B	222.9	170
	L	737.7	170
	3B	2851.6	170

Table 4: Summary of models evaluated on PROST. We list the amount of pretraining data as the size of the uncompressed text corpus used.

each question four times, each time with a different choice replacing the [MASK] token. Following Warstadt et al. (2020), we select the sentence with the highest probability.

Encoder Models We analyze BERT (uncased) (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), which are all based on transformer encoders. BERT is trained on MLM and next sentence prediction and uses static masking, RoBERTa is trained on MLM with dynamic masking, and ALBERT uses whole-word n-gram masking. For probing, we filter out all but the four answer choices from the output vocabulary and select the token with the highest probability as the model's decision.

Encoder-decoder Models We also include results for T5 (Raffel et al., 2020). T5 is trained using a span corruption objective, in which spans of the input sequence are randomly replaced with

a single unique mask token. During pretraining, span lengths are chosen randomly with an average length of three. To keep our results consistent with the other models, we restrict the span length to one token. We find that two of the options for sliding surfaces, namely *ice* and *frost*, violate our single-token constraint. To avoid any unfair comparison between answers that differ in token lengths and following previous work (Goldberg, 2019), we chose to omit presenting the results for T5 on the sliding concept.

Finetuned Conditional LMs To better understand the limitations of text-only training, we additionally evaluate UnifiedQA (Khashabi et al., 2020). UnifiedQA is a pretrained QA model, built off T5, and finetuned on SQuad 1.1, SQuaD 2.0, NarrativeQA, RACE, ARC, OpenBookQA, MCTest, and BoolQ (Rajpurkar et al., 2016, 2018; Kočiský et al., 2018; Lai et al., 2017; Clark et al., 2018; Mihaylov et al., 2018; Richardson et al., 2013; Clark et al., 2019). We format all of our templates to fit their multiple-choice question answering format and use their provided scoring metrics to select the models' answers.[3]

5 Results

The per model and per concept results are shown in Table 5. For concepts with more than one template—direction, mass, height, and circumference—we average across templates to get the concept's score. We can see that, on average, ALBERT-V2-XL performs best, with an accuracy of 31.8%[4], and GPT-2 performs worst, with an accuracy of 23.6%. We note that random guessing would yield an accuracy of 25%. Furthermore, every model underperforms random guessing on at least one concept. Since PROST is trivially solvable for humans, this supports our hypothesis that pretrained models are unable to perform physical reasoning anywhere close to human performance.

Comparing across all concepts, we see that direction obtains the highest average accuracy with 46.8%. The second best accuracy is observed for the mass attribute with 36.5%. The concepts models struggle the most with are the slideable and bounceable affordances, both with an average accuracy of 19.9%.

[3]https://github.com/allenai/unifiedqa

[4]Note: as detailed in Section 4, T5 and UnifiedQA are not being evaluated on sliding. We therefore disregard their average accuracy.

6 Analysis

Object Order in Context For the concepts that use objects, all four choices are listed in each question's context. PROST contains all permutations with regards to their ordering. This enables us to directly look at the effect of the correct answer's position within the context on the models' accuracy. These results shown in Table 6.

We see that models have a strong tendency to select either the first or the last item seen in the context. The largest difference is found for T5, with an accuracy of 52.4% for objects at position 1 and an accuracy of only 1.9% for objects at position 3. We note that a proper understanding of the questions, as most humans would have, would be robust to the order in which the choices are presented. This further underlines that state-of-the-art models do not perform human-like physical reasoning.

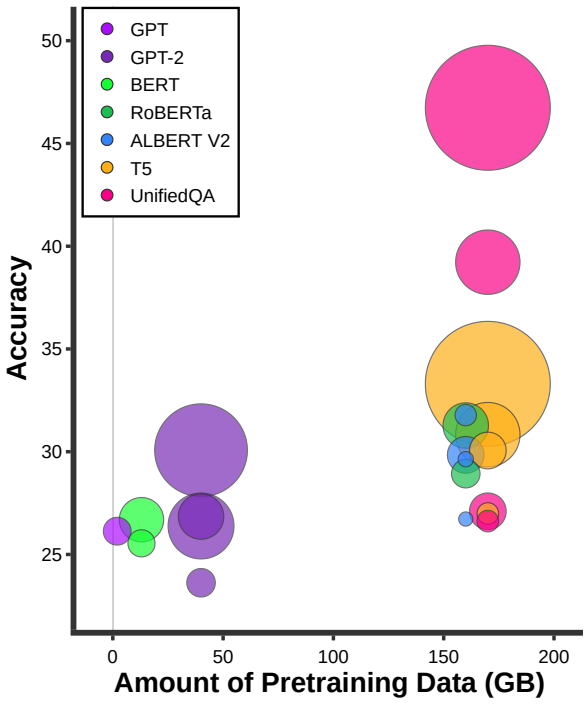

Figure 2: Scaling effect of models on accuracy. Circles size represents number of parameters.

Superlative Inverses By inverting the superlative in a question, we are able to probe a mirrored version of the question. For example, for attributes, this would require the model to identify the lightest object instead of the heaviest object, or, for affordances, it would require the model to identify the not stackable object instead of the stackable object. We call these mirrored versions *superlative inverses*. A true understanding of the questions in

Model		Direction	Mass	Height	Circum.	Stack	Roll	Grasp	Break	Slide	Bounce	Macro Average
GPT		46.7	40.1	24.3	22.8	28.2	27.9	19.6	22.7	14.6	14.4	26.1
GPT2		43.3	31.4	22.0	18.8	26.2	20.3	17.9	22.5	16.9	17.0	23.6
	M	48.3	34.1	21.6	21.6	25.5	23.7	24.9	27.8	22.5	18.5	26.8
	L	46.7	33.1	25.4	27.0	25.5	26.9	20.6	21.8	21.3	15.6	26.4
	XL	46.7	34.2	25.8	26.3	31.1	36.3	29.4	26.7	23.7	20.5	30.1
BERT	B	40.0	32.9	27.5	25.6	20.9	26.1	23.3	28.0	18.2	13.0	25.5
	L	**70.0**	38.8	19.4	17.5	21.3	19.2	26.7	19.5	15.9	18.6	26.7
RoBERTa	B	46.7	36.9	25.8	23.5	34.5	19.3	25.4	**45.0**	20.9	11.4	28.9
	L	66.7	**43.4**	33.8	22.7	22.7	22.2	29.4	23.8	22.7	25.5	31.3
ALBERT V2	B	21.7	35.4	30.2	26.0	25.2	32.5	35.3	22.8	15.3	22.9	26.7
	L	41.7	38.2	31.9	27.5	23.3	29.7	34.0	24.5	23.4	22.1	29.6
	XL	46.7	38.7	**42.0**	**37.4**	30.2	28.2	**37.1**	17.8	**25.3**	14.3	**31.8**
	XXL	68.3	33.8	28.1	24.5	29.4	23.4	21.2	30.2	17.5	22.1	29.8
T5	S	20.0	36.5	29.8	25.2	25.0	25.9	25.4	25.0	—	30.2	27.0*
	B	40.0	37.0	32.6	23.8	25.0	23.4	25.2	25.6	—	**37.8**	30.1*
	L	46.7	35.7	30.7	27.6	31.8	23.0	34.0	25.2	—	22.7	30.8*
	3B	46.7	39.6	35.6	29.8	**34.7**	31.5	35.6	33.8	—	12.5	33.3*
UnifiedQA	S	0.0	34.2	34.8	30.3	24.4	29.0	28.8	27.1	—	31.0	26.6*
	B	0.0	17.8	33.3	22.3	25.5	34.9	27.9	36.5	—	45.7	27.1*
	L	83.3	17.2	49.5	47.3	23.5	28.4	27.5	43.6	—	32.6	39.2*
	3B	63.3	37.8	55.2	66.9	31.2	35.2	24.8	81.4	—	24.8	46.7*
Task Average		46.3	36.5	28.6	25.2	27.1	25.9	27.4	26.0	19.9	19.9	28.5

Table 5: Macro average for each concept and overall for each model on PROST. The best accruacy for general pretrained-only models is displayed in bold. Note that the task average does not include UnifiedQA.

Model	Position Accuracy			
	1	2	3	4
GPT	27.0	24.3	7.6	38.6
GPT-2	29.9	23.1	8.1	42.0
BERT	28.4	24.3	5.7	38.2
RoBERTa	39.0	28.7	11.2	30.0
ALBERT V2	32.5	25.8	9.7	44.2
T5	52.4	21.1	1.9	35.2
UnifiedQA	41.0	27.7	18.8	51.9
Position Average	35.7	25.0	9.0	40.0

Table 6: Accuracy across the correct answer's position in the context.

PROST should be robust to this kind of inversion. However, Table 7 shows all models perform better on one of the two versions. Of the probed models, GPT-2 is the most unbalanced, averaging 30.6% higher for one version over the other.

Data and Model Scaling Figure 2 shows each model's accuracy as a function of the number of its parameters. Unlike for many modern benchmarks, where increasing the number of parameters or training data provides significant benefits (Talmor et al., 2020; Wang et al., 2018), PROST does not see much improvement from such scaling. We observe some improvements with T5-3B outperforming T5-small, but this 6.6% increase requires a 48x increase in parameters and T5-small still outperforms T5-3B on one task. Moreover, some models break this trend: ALBERT's XL version outpeforms its XXL counterpart and GPT-2 M outperforms GPT-2 L. While previous work has revealed the impressive scaling laws of transformer-based architectures (Kaplan et al., 2020), PROST highlights the importance of relevant and informative training. As physical reasoning is not an ability that humans acquire via text, even substantially more open domain textual data is unlikely to lead to more than marginal improvements.

The Limits of Text-based Training To our knowledge, UnifiedQA is the most qualified model to succeed on our task, having been finetuned on a significant amount of relevant text data. While this additional data does provide benefits on PROST, it still falls short, with the best performing model we tested only achieving a 46.7% accuracy. Additionally, from Tables 6 and 7, it still lacks the robustness of proper understanding. This emphasizes that models are unlikely to obtain human-like reasoning from text-based training alone. Rather, PROST motivates exposing models to concepts through multiple modalities that mirror a human's experience.

Model		Mass	Height	Circum.	Stack	Roll	Grasp	Break	Slide	Bounce	Macro Average
GPT		9.4	2.2	14.0	35.9	43.0	22.7	13.3	9.8	10.1	17.8
GPT-2		19.2	24.3	1.1	16.1	5.1	12.1	31.9	24.9	15.8	16.7
	M	31.2	12.9	21.5	12.7	20.2	7.8	49.9	33.3	9.2	22.1
	L	25.8	24.2	25.4	5.6	16.7	18.4	24.6	28.2	23.7	21.4
	XL	43.5	6.7	1.5	56.1	51.5	36.3	31.5	15.8	32.4	30.6
BERT	B	5.0	40.0	2.5	12.3	15.4	3.1	44.9	12.2	11.2	16.3
	L	19.2	21.5	5.8	1.8	3.4	4.2	17.2	9.5	30.4	12.6
RoBERTa	B	4.7	4.0	6.5	55.0	13.8	27.8	89.6	21.8	15.8	26.5
	L	31.0	24.7	26.8	9.7	21.1	33.2	31.5	33.9	33.2	27.2
ALBERT V2	B	4.7	31.0	7.9	14.8	14.4	66.4	**2.3**	**7.2**	1.5	16.7
	L	**0.6**	11.9	23.6	30.0	36.8	52.2	6.9	28.7	13.2	22.7
	XL	9.4	**0.7**	8.5	19.5	8.1	31.0	10.9	8.2	20.6	13.0
	XXL	18.1	2.9	18.5	4.2	12.2	2.0	37.2	16.3	11.3	13.6
T5	S	8.3	12.9	13.5	**0.0**	3.9	0.8	3.1	—	1.8	**5.5**
	B	8.7	26.9	5.8	0.0	**0.2**	**0.5**	3.1	—	26.2	8.9
	L	5.0	20.3	1.7	7.5	22.6	10.5	37.7	—	**0.5**	13.2
	3B	16.1	12.2	**0.7**	9.2	8.8	5.1	34.7	—	24.9	14.0
UnifiedQA	S	46.9	5.4	19.7	2.8	34.5	31.0	39.0	—	9.3	23.6
	B	19.2	19.2	4.3	24.6	30.4	10.4	4.9	—	49.5	20.3
	L	18.5	28.5	26.2	18.6	28.1	41.6	7.9	—	48.1	27.2
	3B	8.2	46.2	36.1	6.8	1.8	0.7	13.4	—	26.4	17.5
Task Average		15.3	16.4	10.9	17.1	17.5	19.7	27.7	19.2	16.6	17.6

Table 7: Absolute difference in accuracy between a question and its superlative inverse.

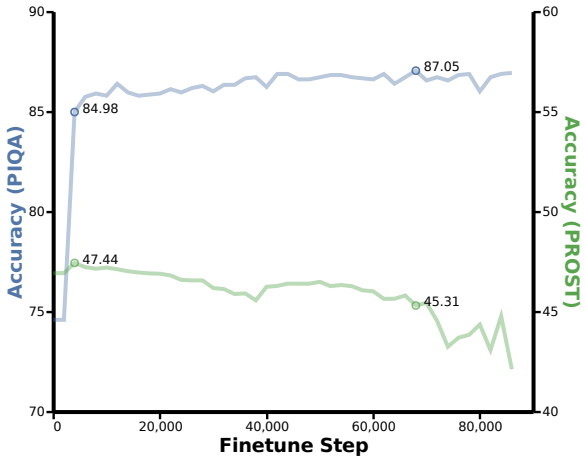

Figure 3: Analysis of the performance of UnifiedQA 3B on PROST throughout PIQA finetuning. The left and right Y axis represent Accuracy on the PIQA dev set and Macro accuracy on PROST respectively. We finetune for 100K steps, and compute metrics every 2k steps. Annotations correspond to the checkpoints with the best performance on PIQA and PROST. Note that PIQA has two answer choices, while PROST has 4.

Comparing PROST and PIQA Due to their shared focus on text-based physical reasoning, PROST and PIQA share similarities. To test if models trained on PIQA are able to carry over any concepts to PROST, we further finetune a UnifiedQA model on PIQA and evaluate it on PROST. The results, shown in Figure 3, indicate that training a model on PIQA is detrimental to its performance on PROST. While PIQA and PROST share a few conceptual similarities, they differ in terms of format, style, and vocabulary. We thus hypothesize that current models learn more about these surface-level differences than the conceptual similarities underpinning the questions. We further highlight two key differences between the two datasets:

- PROST probes models in a zero-shot fashion, whereas PIQA provides training and test sets of identically distributed examples. This makes it possible for models on PIQA to answer successfully using spurious correlations rather than physical reasoning.

- PIQA (Bisk et al., 2020b) covers an extensive range of objects and challenging physical concepts. Bisk et al. (2020a) argues that experience is a prerequisite for understanding. It is hard to imagine how to expose a model to experiences ranging from egg yolk separation to making a pillow out of a garbage bag. In contrast, PROST provides a clear set of well defined concepts and objects that a model could potentially experience.

7 Discussion

Our experiments show that all the models we analysed fail to demonstrate a robust understanding of physical reasoning. Beyond performing poorly across every concept, they are easily influenced by changing the order of the objects in a question's context and by superlative inverses. Moreover, our analysis indicates that these issues are not likely to be solved simply by increasing the amount of model parameters or training data. All this evidence supports Bender and Koller (2020)'s and Bisk et al. (2020a)'s theory that experience is a prerequisite of understanding.

A number of other reasoning benchmarks have been solved to some extent by a large finetuned model. UnifiedQA (11B parameters), based on T5 (Raffel et al., 2020), achieved 81.4% on ARC (Clark et al., 2018); and UNICORN[5] (11B parameters), also based on T5, achieved a 93.9% accuracy on hellaSWAG (Zellers et al., 2019). While all these models are larger and are trained on more data, our results force us to ask the question whether they perform well because these additional parameters and data have imbued the models with an ability to reason, or if they succeed by finding subtle unintended correlations in the data. This forces us to look more closely at how models succeed, and not just the accuracy they achieve. Tools like CheckList (Ribeiro et al., 2020) can aid in this endeavor by demonstrating how robust models are to changes in the distribution of the data.

How to Use this Probe PROST is intended to help analyze any model that can be deployed in a text-only setting. However, we maintain that multi-modal data is necessary to experience the concepts in PROST, and that these experiences are likely a crucial step in succeeding on this dataset. One way that multi-modal models could prepare for this type of text-only evaluation is through multi-task training, where one of the tasks is only conditioned on text. Such an approach has already been considered: Brown et al. (2020) propose an extension to their CLIP model which is trained on multiple modalities in a multi-task fashion. Because of the templated nature of PROST, its exact format can be adapted to match specific styles of language training, as we do for T5 and UnifiedQA.

PROST's language-only approach is motivated by two reasons. First, we believe that true multi-modal models should be able to function on any subset of their modalities. We note that humans can easily interact with text-only inputs (e.g., a text message) while still learning from and interacting with other modalities. Second, it enables the comparison of models trained using different modalities or domains. For example, we believe comparing how language understanding modules evolve when trained on vision-and-language navigation compared to visual question answering would provide invaluable insights.

Limitations We caution that achieving a high accuracy on PROST does not necessarily guarantee that a model is able of physical reasoning. It is likely easy to succeed on this benchmark if one were to intentionally train models on similar enough sentences or a subset of PROST itself. We hope that the community will use this dataset in the intended way: in a zero-shot setting to probe models which have been trained on data not specifically collected to succeed on PROST.

8 Conclusion

We present a probing dataset called PROST, which is designed to test a model's ability to reason about the physical world. Our experiments show that current state-of-the-art pretrained models lack the ability to reason about physical interactions. Further, all models struggle when the order of options is changed and when questions are inverted, both things that would not confuse humans. Lastly, our analysis shows that these issues are unlikely to be solved by simply scaling models. Our results highlight the need to look beyond text-based pretraining and to provide models with the necessary experiences for human-like understanding of the physical world.

Acknowledgments

We would like to thank the members of CU Boulder's NALA and HIRO Groups for their feedback on this work.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

[5] leaderboard.allenai.org/hellaswag/submissions/public

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, volume 32, pages 5082–5093. Curran Associates, Inc.

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, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tianshi Cao, Jingkang Wang, Yining Zhang, and Sivabalan Manivasagam. 2020. Babyai++ : Towards grounded-language learning beyond memorization. In *ICLR*.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Susan A. Gelman. 2009. Learning from others: Children's construction of concepts. *Annual Review of Psychology*, 60:115–140.

Pranav Goel, Shi Feng, and Jordan Boyd-Graber. 2019. How pre-trained word representations capture commonsense physical comparisons. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 130–135, Hong Kong, China. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual question answering in interactive environments. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE.

Susan J Hespos and Elizabeth S Spelke. 2004. Conceptual precursors to language. *Nature*, 430(6998):453–456.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Corey Lynch and Pierre Sermanet. 2020. Grounding language in play. *arXiv preprint arXiv:2005.07648*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, and A. Farhadi. 2017. Visual semantic planning using deep successor representations. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 483–492.