MIND'S EYE: GROUNDED LANGUAGE MODEL REASONING THROUGH SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful and effective communication between humans and AI relies on a shared experience of the world. By training solely on written text, current language models (LMs) miss the grounded experience of humans in the real-world—their failure to relate language to the physical world causes knowledge to be misrepresented and obvious mistakes in their reasoning. We present Mind's Eye, a paradigm to ground language model reasoning in the physical world. Given a physical reasoning question, we use a computational physics engine (DeepMind's MuJoCo) to simulate the possible outcomes, and then use the simulation results as part of the input, which enables language models to perform reasoning. Experiments on 39 tasks in a physics alignment benchmark demonstrate that Mind's Eye can improve reasoning ability by a large margin (27.9% zero-shot, and 46.0% few-shot absolute accuracy improvement on average). Smaller language models armed with Mind's Eye can obtain similar performance to models that are 100× larger. Finally, we confirm the robustness of Mind's Eye through ablation studies.

1 INTRODUCTION

"In questions of science, the authority of a thousand
is not worth the humble reasoning of a single individual."

——Galileo Galilei, 1632

"Do objects fall proportionately to their weight?" This famous question was once controversial until Galileo's Leaning Tower of Pisa experiment[1]—Galileo dropped two balls of different masses from the same height (i.e., *experiment*) and concluded that their time of descent was independent of their mass (i.e., *inductive reasoning*). Such an experiment-reasoning paradigm has been used by humans for centuries to ground reasoning on complicated problems (Newell, 1980) and transfer learned knowledge to unfamiliar domains (Novak & Gowin, 1984).

Current language models (LMs) follow a different path—by training on natural language, they attempt to reverse engineer the physical world, so that they are able to reason about it. Large-scale pre-trained LMs have achieved revolutionary performance on many tasks, such as solving math word problems (Roy & Roth, 2015; Ling et al., 2017; Cobbe et al., 2021) and commonsense reasoning (Talmor et al., 2022; Geva et al., 2021). However, these models do not experience firsthand the situations that are described by the language (McClelland et al., 2020), and lack the ability to find the correct answers by performing experiments like humans. As a consequence, when asked the same free fall question, one of the most widely-used LMs, GPT-3[2] (Brown et al., 2020)—though achieving superhuman performance in many reasoning tasks—will generate the wrong answer: *"The heavier object will fall faster."* (as shown in Figure 1). Due to the lack of grounded reasoning, current LMs also have issues in truthfulness (Lin et al., 2021) and factuality (Petroni et al., 2020).

To tackle these problems, existing remedies include using improved prompting techniques, such as inserting hand-written decomposed reasoning steps in few-shot demonstrations (Wei et al., 2022;

[1] In *Physics*, Aristotle (384–322 BC) claims that the speed at which two identically shaped objects fall is directly proportional to their weights, which was later challenged by Aristotelian commentator John Philoponus.

[2] Specifically, we use text-davinci-002, which is the "most capable GPT-3 model" at the time of writing from OpenAI: https://beta.openai.com/docs/models/overview.

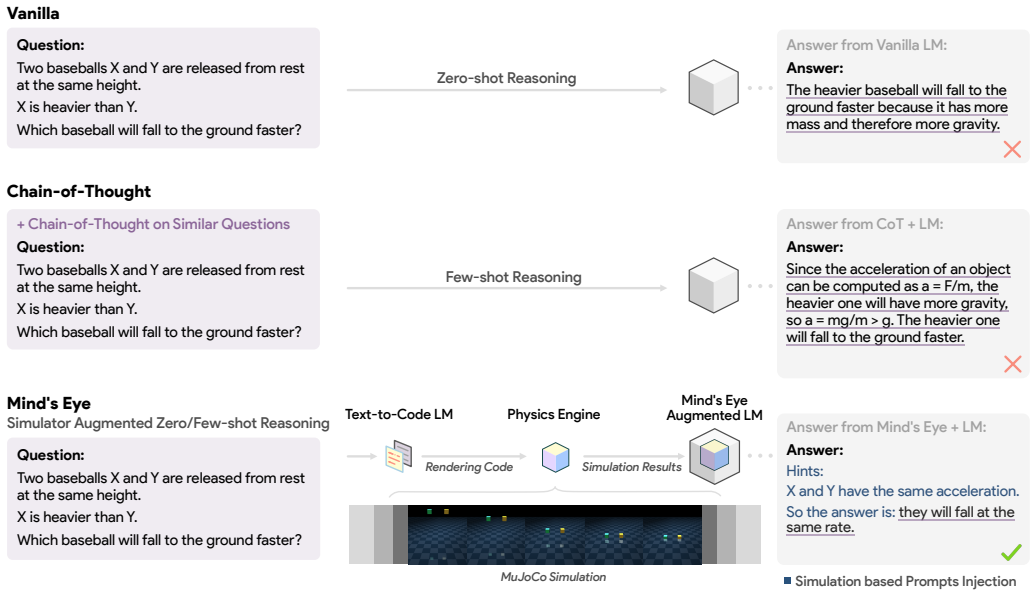

Figure 1: Current language models are still challenged by simple questions that require a good understanding of the physical world. The answer elicited by Chain-of-Thought can still be wrong if the required knowledge is missing or misrepresented in LMs. Mind's Eye, instead, enables grounded LM reasoning by directly simulating the scene in the given question. Then the LM can reason over the injected ground-truth evidence to generate the correct answers.

Zhou et al., 2022). These methods are inherently limited as their reasoning ability completely relies on the knowledge perpetuated in the LM—their performance could suffer if the knowledge learnt by the LM is incorrect (Petroni et al., 2019) or outdated (Dhingra et al., 2022). To incorporate external knowledge, retrieval-augmented LMs such as REALM (Guu et al., 2020), RAG (Lewis et al., 2020) and RETRO (Borgeaud et al., 2022), retrieve relevant documents as additional evidence for given questions, and may also fine-tune the LM on the question-document-answer triplets. However, the knowledge presented in written language is known to have *reporting bias* (Bisk et al., 2020), whereby some everyday unspoken facts or rarely seen (but practically possible) compositions are commonly missing in text (Paik et al., 2021).

Correct and complete understanding of properties and interactions in the physical world is not only essential to achieve human-level reasoning (Lake et al., 2017), but also fundamental to build a general-purpose embodied intelligence (Huang et al., 2022). In this work, we investigate to *what* extent current LMs understand the basic rules and principles of the physical world, and describe *how* to ground their reasoning with the aid of simulation. Our contributions are three-fold:

- We propose a new multi-task physics alignment dataset, UTOPIA, whose aim is to benchmark how well current LMs can understand and reason over some basic laws of physics (§2). The dataset contains 39 sub-tasks covering six common scenes that involve understanding basic principles of physics (e.g., conservation of momentum in elastic collisions), and all the ground-truth answers are automatically generated by a physics engine. We find that current large-scale LMs are still quite limited on many basic physics-related questions (24% accuracy of GPT-3 175B in zero-shot, and 38.2% in few-shot).

- We explore a paradigm that adds physics simulation to the LM reasoning pipeline (§3) to make the reasoning grounded within the physical world. Specifically, we first use a model to transform the given text-form question into rendering code, and then run the corresponding simulation on a physics engine (i.e., MuJoCo (Todorov et al., 2012)). Finally we append the simulation results to the input prompts of LMs during inference. Our method can serve as a

plug-and-play framework that works with any LM and requires neither handcrafted prompts nor costly fine-tuning.

- We systematically evaluate the performance of popular LMs in different sizes on UTOPIA before and after augmentation by Mind's Eye, and compare the augmented performance with many existing approaches (§4.2). We find Mind's Eye outperforms other methods by a large margin in both zero-shot and few-shot settings. More importantly, Mind's Eye is also effective for small LMs, and the performance with small LMs can be on par or even outperform that of 100× larger vanilla LMs.

2 UTOPIA BENCHMARKING

Humans are able to understand their physical environment and intuit rules about the world from embodied experience. The rules and principles behind the real world have been discovered as scientific laws—we humans have ingrained them as knowledge or intuition (Kaiser et al., 1986) to make reliable predictions on how observed events will unfold in day-to-day life (Kubricht et al., 2017). For example, when driving, we can anticipate when to brake when approaching a stop sign, using intuition or knowledge from Newton's second law of motion. We also know it would be a disaster to collide with a heavy truck, both from embodied experience of colliding objects and because the conservation of momentum tells us that the lighter object will have a greater velocity after collision.

We are thus inspired to design a physics alignment dataset that covers this knowledge, aiming to benchmark to what extent current LMs understand basic physical concepts and rules. As shown in Table 1, we choose six representative scenes, mainly from textbooks (e.g., high-school Physics). The sub-tasks are defined based on the composition of observed and queried concepts. For example, one task in a motion scene could be: given the observed *acceleration* of the two objects with the same mass, please answer what is the relationship of *forces* applied on them. In total we have 39 sub-tasks across different scenes, and each sub-task contains various hand-designed questions whose language style is similar to that of textbooks as well.

Table 1: We propose UTOPIA, a multi-task physics alignment dataset, investigating the grounded reasoning ability of LMs on 39 sub-tasks. Unlike many other datasets, UTOPIA deliberately describes the questions in relative relations (e.g., *greater than*) instead of absolute numbers (e.g., 3.5 m/s), to approximate human's perceptional sensing ability in real world. The ground-truth answers to the questions are generated by the physics engine, which makes it easy to scale UTOPIA to larger sizes.

Scenes	(Simplified) Sample Questions	Concepts	# Tasks
Motion	Amy pulls two sleds X and Y with the same force. X has a greater mass than Y. Friction can be ignored. Which one has a greater acceleration after the same period of time?	mass force velocity	6
Friction	Two boxes X and Y move at the same velocity. We only consider kinetic frictions, and X undergoes a smaller friction than Y. Which one has a greater velocity after the same period of time (before stop)?	mass velocity friction	6
Free fall	Two balls are dropped from the same height. Y has a greater mass than X. We ignore the air resistance. Which one will hit the ground earlier?	mass height energy	6
Projection	Jason throws two baseballs X and Y at the same height horizontally. They have the same mass, but X has a greater initial horizontal velocity. Which one will hit the ground earlier?	velocity mass energy	6
Collision	Two marbles X and Y of the same mass move towards each other. X and Y have the same magnitude of velocity, and the collision is elastic. Which one will have a greater velocity after collision?	velocity mass momentum	6
Incline	Two blocks of metal X and Y are released from a certain height on a slick slope. Y has a greater mass than X, and the friction can be ignored. Which one will have a greater velocity after the same period of time?	mass height friction	9

Table 1 exemplifies some samples in UTOPIA. We deliberately choose to use relative comparison (e.g., *"greater than"*, *"smaller than"*, *"the same as"*; text in purple) rather than actual numbers to describe the physical properties, since we are thus able to disentangle the effects from numeracy (i.e., the gain on reasoning is not attributed to better memorization on numbers, which has been

reported as "shortcuts" used by LMs (Patel et al., 2021)). This setting is also different from those in mathematical reasoning tasks (e.g., GSM8k (Cobbe et al., 2021)), where the decomposed reasoning path is typically the procedure of plugging different values into equations—the LM might be able to solve these problems by symbolic manipulation (Razeghi et al., 2022) rather than actual reasoning.

Most existing physics alignment datasets use vision as the primary modality, such as images (Zellers et al., 2019), animations (Wu et al., 2017), or videos (Piloto et al., 2022), which loses the flexibility to run on LMs which only takes text input. PiQA (Bisk et al., 2020) and MMLU-Physics (Hendrycks et al., 2021) are popular physics reasoning datasets used for LM benchmarking; however, their sizes are naturally limited because of required human annotations (e.g., only 206 samples are on physics in MMLU, with college and high school level questions combined). UTOPIA differs from all these datasets as it leverages a physics engine to generate data—in theory we can obtain *unlimited* samples—and each sample has reliable ground-truth supported by actual simulation. Although in the present work we only take the text-form data for LM benchmarking, the corresponding simulation videos during data generation have been recorded as data for future multi-modality research.

3 MIND'S EYE

As shown in Figure 1, Mind's Eye comprises three main components, a text-to-code LM as the front-end, a physics simulation engine (i.e., MuJoCo) as the back-end, and a foundation model (Bommasani et al., 2021) for general reasoning. We detail the implementation of Mind's Eye as below:

Text-to-Code Converter. The objects and dynamics of the simulation is manifested by the rendering code fed into MuJoCo. The rendering code is written in a type of XML file named MCJF[3], where the physics properties can be easily controlled by changing some key-value pairs. For example, to change the mass of an object to 10, the line of rendering code needed is `geom.set('mass', '10')`. We use actual values to express the relative relationships in UTOPIA (e.g., *"greater"* will be translated to 10 and 1 for the values of the properties to be set). We create rendering templates for each sub-task of UTOPIA, and use programs to generate a dataset with 200,000 text-code pairs. In each pair, the question in text is appended to the top of the XML code as comments. We then train decoder-only LMs from scratch to learn how to generate the rendering code given the question in comments auto-regressively. We leverage the BPE vocabulary set from GPT-2 (Radford et al., 2019) and extend it by several special tokens to represent repeating tabs or spaces. Besides fine-tuning on the dataset with text-code pairs, we also pre-train the model on the C4 dataset (Raffel et al., 2019a) to enhance the model's understanding on natural language. All the training is on TPU-v3 Pods and the resulting models have 0.3B and 1.5B parameters (used as default). See §4.1 for training details.

Simulation Augmented Prompting. Once receiving the rendering code, the physics engine will run the corresponding simulation to get the ground-truth outcome. The program that triggers the simulation will also parse the outcome into text-form prompt injections (e.g., *"Hints: Two baseballs take the same time to hit the ground."*, as shown in Figure 1). The injection combined with the question will be fed to the foundation model, with which LMs can ground their reasoning with the physical world rendered by the physics engine. We present more details of this procedure in §A.1.

The intuition behind Mind's Eye is to imitate the experiment-reasoning paradigm; however, we leverage quick and cheap physics simulation as an alternative to actual experiments in physical world. The cognitive analog for Mind's Eye might be the mental visualization process, also known as "the mind's eye" (Battaglia et al., 2013; Hegarty, 2004), which often relates to motor processes (Wexler et al., 1998) during embodied reasoning (Nathan et al., 2021).

Discussion: Why does Mind's Eye work? Table 2 shows the comparison between Mind's Eye and two other methods in the formulation of the grounding process during LM inference. Assuming knowledge of the physical world aligns with the distribution p_{World}, the Zero-shot Reasoner (Kojima et al., 2022) which uses *"Let's think step by step."* in prompts can be extended to any number of new tasks. However, its reasoning ability will be compromised if the knowledge in LMs is incorrect or outdated. Similarly, incorporating handcrafted reasoning steps rather than a generic phrase, Chain-of-Thought (Wei et al., 2022) is able to elicit LM reasoning in a more explicit way; however, its performance is reported to be sensitive to the quality of human annotated reasoning steps (Lampinen

[3]Docs for MJCF: https://mujoco.readthedocs.io/en/latest/XMLreference.html

Table 2: Comparison in formulation of how the LM inference process is grounded (given question x generating answer y). With the aid of a simulator (i.e., MuJoCo), Mind's Eye is not only scalable (not requiring human annotation) but also well-grounded with the physical world.

Method Name	Formulation	Scalable?	Grounded?
Zero-shot Reasoner	$y \leftarrow \arg\max_{\hat{y}} \text{LM}(\hat{y}\|x, \text{“}\textit{Let's think step by step}\text{”})$	✓	✗
Chain-of-Thought	$y \leftarrow \arg\max_{\hat{y}} \text{LM}(\hat{y}\|x, \text{Chain-of-Thought} \sim p_{\text{Human}})$	✗	?
Ours: Mind's Eye	$y \leftarrow \arg\max_{\hat{y}} \text{LM}(\hat{y}\|x, \text{Simulator}(x, \hat{y}) \sim p_{\text{World}})$	✓	✓

et al., 2022; Dasgupta et al., 2022) (i.e., p_{Human} is not similar to p_{World}). The dependence on human annotation also limits its scalability.

Mind's Eye overcomes these issues by including a simulator into the reasoning pipeline. Given the question, the simulator (i.e., the physics engine) returns the most likely outcome based on its encoded world knowledge. Since the simulator is accurate enough to approximate the physical world, the prompt injection of Mind's Eye basically serves as a scoring machine, which puts probability mass on the answer that is best aligned with the rules of physics—the LM reasoning over the injected rationales is thus grounded. Mind's Eye is also scalable since the whole pipeline is automated.

Besides scalability and grounded reasoning, Mind's Eye is also efficient. Mind's Eye delegates domain-specific knowledge to external expert models (i.e., the MuJoCo engine for expert physics knowledge) (Du et al., 2022; Shazeer et al., 2017), which decouples general reasoning from domain specific knowledge. The size of the LM can thus be significantly shrunk since it removes the burden of memorizing all the domain-specific knowledge. Experiments find that 100× smaller LMs augmented with Mind's Eye can achieve similar reasoning capabilities as vanilla large models, and its prompting-based nature avoids the instability issues of training mixture-of-expert models (Zoph et al., 2022). The compatibility with small LMs not only enables faster LM inference, but also saves time during model saving, storing, and sharing.

4 Experiments

4.1 Experiments Settings

Data and Model. For the convenience of benchmarking on huge LMs, we prepare 100 samples for each sub-task, resulting in a dataset with about 3,900 samples. We use this version of Utopia for evaluation across the paper. The MuJoCo simulations can achieve 171 fps on one A6000 GPU, and generating 100 simulations of a 2 seconds collision scene takes 0.67s. For LMs, besides GPT-3, we have also tested Pathway Language Model (PaLM) (Chowdhery et al., 2022) on 8B, 62B, and 540B checkpoints. All experiments for PaLM are run on TPU-v4 Pods.

Training Details. Training of the JAX-based text-to-code LMs runs on TPU-v3 Pods. The learning rates we use for training 0.3B and 1.5B LMs on C4 are {3.0e-4, 1.8e-4}, which are switched to {1.8e-4, 0.5e-4} when fine-tuning on the text-code pairs. We use cosine annealing to control learning rate over time with fixed warm-up steps (3k). As mentioned in §A.2, we have also fine-tuned 62B PaLM to study task generalization, which takes about 25 minutes on 64 TPU-v4 chips for each task.

4.2 Benchmark Results on Utopia

As shown in Figure 4.2, we run experiments on Utopia with GPT-3 and PaLM of different sizes, ranging from 340M (GPT-3 Ada) to 540B (PaLM). Although larger LMs perform consistently better than smaller LMs in both zero-shot and few-shot settings (n = 5), reasoning ability seems to plateau after a certain size, especially in the few-shot setting. In other words, the scaling curve of vanilla few-shot is nearly flat. One interpretation could be that few-shot demonstrations have managed to trigger effective in-context learning (e.g., for learning the answer format), but the lack of grounded reasoning becomes the bottleneck for further improvement.

Mind's Eye, however, unlocks the ability to scale (red line in Figure 4.2) by adding knowledge from physics engine simulations. Since the correctness of the simulation results is guaranteed by the

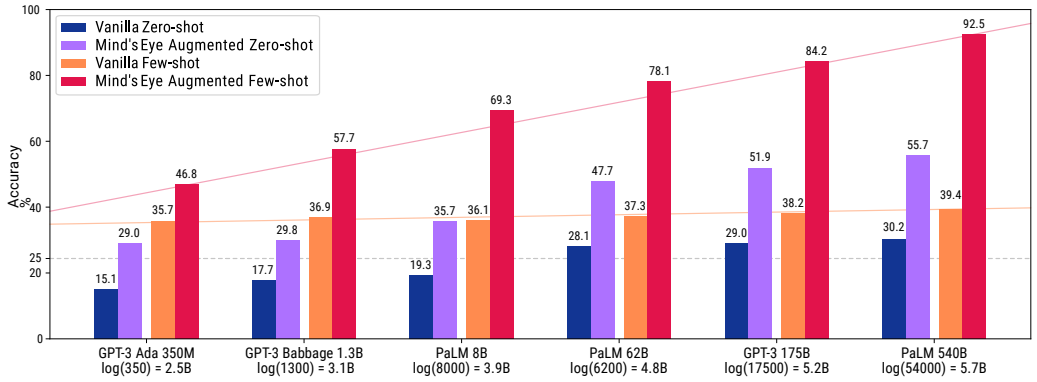

Figure 2: Scaling law of reasoning on UTOPIA when benchmarking LMs in different sizes (log scale). Smaller LMs augmented with Mind's Eye can achieve on par or even outperform larger vanilla LMs in both zero-shot and few-shot settings (e.g., 29.8% for GPT3 1.3B + Mind's Eye vs. 29% for vanilla GPT-3 175B in zero-shot). In few-shot settings ($n = 5$), the scaling potential is unlocked when LMs are augmented with Mind's Eye (red line), as the scaling curve is nearly flat in vanilla few-shot mode (orange line), which demonstrates the power of incorporating knowledge from simulations. On average across all sizes, the gain from Mind's Eye is greater in large LMs than in small ones, and greater in few-shot (34.2%, absolute) than in zero-shot settings (18.4%, absolute).

physics engine, the reasoning of the large foundation model is thus well grounded. Compared with solely using knowledge perpetuated in LMs (results with vanilla LMs), Mind's Eye is able to boost the reasoning ability of LMs by 18.4% in zero-shot and 34.2% in few-shot settings. Interestingly, smaller LMs augmented with Mind's Eye can achieve similar or even better performance than vanilla larger LMs. For example, the accuracy of GPT-3 Babbage 1.3B with Mind's Eye in zero-shot is 29.8%, while that of vanilla GPT-3 Davinci 175B in zero-shot is 29%. This finding demonstrates the effectiveness of the Mind's Eye paradigm decoupling *experimenting* from *reasoning*—where the domain specific tool is responsible for providing ground-truth results, and the LM can mainly focus on general reasoning, whose size can thus be largely decreased.

Comparison with other reasoning enhanced techniques. In Table 3, we compare Mind's Eye with other methods that improve the reasoning abilities of LMs (i.e., GPT-3 175B). In addition to a) **Chain-of-Thought** (Wei et al., 2022), we consider other prompt-based methods such as b) **Zero-shot Reasoner** (Kojima et al., 2022), which uses *"Let's think step by step."* in prompts to induce the decomposed reasoning steps of LMs; c) **Self-Consistency Decoding** (Wang et al., 2022), which is an ensemble technique that decodes multiple reasoning paths concurrently to improve the performance of Chain-of-Thought; the recent study d) **DiVerSe** (Li et al., 2022b) achieves new SotAs on many reasoning tasks, by using pre-trained verifiers to weight good answers more than bad ones in each step decoding. Besides RAG (Lewis et al., 2020) that retrieves knowledge from the memory of 75GB documents, we also consider other fine-tuned LMs which attempt to optimize reasoning ability from different perspectives, such as e) task-scaling **T0 (version pp)** (Sanh et al., 2021), which fine-tunes T5 (Raffel et al., 2020) on thousands of prompts to make LMs better understand input prompts, especially in zero-shot settings, and f) model-scaling **Minerva** (Lewkowycz et al., 2022), which fine-tunes PaLM (Chowdhery et al., 2022) 540B on a newly collected dataset that contains scientific and mathematical documents (e.g., arXiv papers) to improve quantitative reasoning. Though most experiments are running on GPT-3 175B, we also explore the scaling effect by using the 1.3B Babbage model. The 'grounding gain' is the absolute accuracy difference for the 175B model augmented with and without Mind's Eye. Unless otherwise stated, we use the default parameter settings recommended by competitor methods.

Results shows that Mind's Eye outperforms other methods in both zero-shot and few-shot settings significantly, even if using a relatively smaller LM (i.e., GPT-3 Babbage 1.3B). Comparing results on GPT-3 175B and 1.3B, we can conclude that 1) larger LMs can better leverage Mind's Eye, especially in few-shot settings (46.0% vs. 10.3% average grounding gain), probably because larger LMs are

Table 3: Comparison of Mind's Eye and other methods on UTOPIA benchmarking. Step-by-step (Kojima et al., 2022), Chain-of-Thought (Wei et al., 2022) are two prompt-based methods that can elicit reasoning in large-scale LMs. Self-consistency (Wang et al., 2022) and DiVerSe (Li et al., 2022b) are decoding-time optimization techniques for LMs reasoning. RAG (Lewis et al., 2020) is a retrieval augmented LM, while T0 (Sanh et al., 2021) and Minerva (Lewkowycz et al., 2022) are fine-tuned LMs to improve reasoning ability by task-scaling and model-scaling. We present results of Mind's Eye on GPT-3 175B/1.3B, and find that a 100× smaller LM can outperform a vanilla 175B model (*ref.*) when armed with Mind's Eye. Interestingly, fine-tuning on prompts to better follow human instructions, Instruct-GPT (Ouyang et al., 2022) can achieve nearly perfect physics alignment in few-shot. We also annotate the grounding gain of Mind's Eye against vanilla GPT-3 175B.

	Motion		**Friction**		**Free Fall**		**Projection**		**Collision**		**Incline**		**Average**	
Existing Methods	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few
GPT-3 175B (*ref.*)	21.5	34.3	24.3	40.0	22.2	44.5	27.0	31.8	26.5	44.8	22.6	33.9	24.0	38.2
GPT-3 1.3B	15.8	34.3	16.7	33.2	6.7	51.8	18.5	33.0	13.5	33.5	19.6	32.4	15.1	36.4
Instruct-GPT 175B	76.0	85.3	60.2	83.2	39.6	47.6	39.8	50.0	49.0	76.8	37.1	68.6	50.3	68.6
Better Prompting														
Zero-shot Reasoner	25.0	-	35.0	-	31.6	-	21.6	-	32.0	-	28.5	-	29.0	-
Chain-of-Thought	-	41.5	-	42.2	-	53.0	-	40.3	-	38.0	-	41.6	-	42.8
Optimized Decoding														
Self-Consistency	-	38.9	-	44.5	-	57.2	-	48.2	-	48.8	-	43.9	-	46.9
DiVerSe	20.2	33.1	17.3	29.6	24.3	37.6	23.1	35.6	25.4	39.3	24.5	37.8	22.4	35.5
Augmented LMs														
RAG (0.4B + 75GB)	7.4	-	8.3	-	10.2	-	5.2	-	11.4	-	8.5	-	8.5	-
T0-pp (11B)	12.3	29.4	14.3	34.1	11.4	25.5	13.7	27.1	9.5	16.8	17.2	30.9	13.1	27.3
Minerva (540B)	24.5	45.3	29.8	41.0	6.2	15.5	36.3	51.0	23.3	46.3	24.3	43.4	24.1	40.4
Ours: Mind's Eye														
w/. GPT-3 175B (⋆)	52.0	82.5	57.0	82.1	30.3	89.8	65.7	83.8	54.0	79.0	52.3	87.8	51.9	84.2
w/. GPT-3 1.3B	33.7	41.7	35.0	42.3	12.8	70.0	34.8	40.0	30.5	45.0	31.8	41.3	29.8	46.7
w/. Instruct-GPT 175B	100.0	100.0	97.8	100.0	54.9	94.8	91.8	99.8	61.8	99.8	99.4	100.0	84.3	99.1
Grounding Gain (%)	30.5	48.2	32.7	42.1	8.1	45.3	38.7	52.0	27.5	34.2	29.7	53.9	27.9	46.0

more capable of general reasoning, and 2) solely scaling-up models is not adequate for reliable grounded reasoning performance. For example, when using a 100× larger LM (1.3B → 175B), the few-shot accuracy is merely boosted by 1.8% (absolute; 36.4% → 38.2%) if using vanilla LMs, but that can be boosted by 37.5% (absolute; 46.7% → 84.2%) if the LMs are augmented by Mind's Eye.

Note that Instruct-GPT(Ouyang et al., 2022) augmented with Mind's Eye is able to achieve nearly perfect performance in few-shot settings. This result is promising because it demonstrates the ideal alignment is achievable if the LM is given proper reasoning evidence and has good understanding of the questions (as Instruct-GPT is optimized for instruction following). The improvement from simply better prompting or decoding methods is limited, since their reasoning completely relies on the internal knowledge perpetuated in the LMs, while the knowledge induced from the LMs could be factually wrong. Among augmented LMs, 540B Minerva is the best performing one but still falls behind Mind's Eye + 175B GPT-3, mainly due to the lack of grounded experience for reasoning. We also find the retrieved evidence of RAG sometimes cannot answer the question, even though it includes entities mentioned in the question, and RAG cannot function well in few-shot settings since the retriever module (i.e., DPR (Karpukhin et al., 2020)) was not pre-trained to handle context that contains multiple question-answer pairs. We have also discussed the domain generalization ability of Mind's Eye in §A.2, and error analysis in §A.3.

4.3 ABLATION STUDY

Do we really need simulation? In Table 4, we show the performance if 1) we randomly alter the simulation results in the prompts to mismatch the physics property asked (e.g., asking velocity but including acceleration results in prompts), and 2) we delete the trigger words at the beginning of the prompts (i.e., *"Hints:"*). We also test what would happen if we ground the reasoning on incorrect simulation results (e.g., Simulation indicates *greater* but we use *smaller* in prompts). We find reasoning on mismatched simulation results causes substantial performance drops, while including

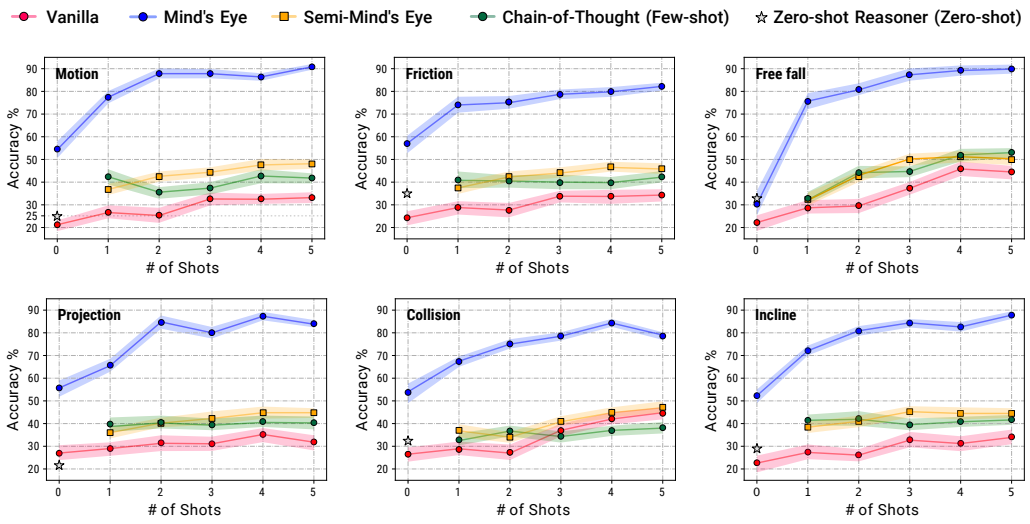

Figure 3: The dynamics of reasoning ability on UTOPIA when we increase the number of shots from zero to five, to examine whether we achieve similar performance with more in-context demonstrations instead of using external simulators (as Mind's Eye).

wrong simulation results will further deteriorate the performance, probably because in this case the reasoning is misguided even if the vanilla LM can answer the question correctly. We take these results as evidence that the correct simulation results are crucial for grounded LM reasoning. Missing trigger words have marginal influence on the performance, which confirms again that reasoning performance mainly depends on the existence and correctness of the simulation results.

Table 4: Ablation study on the simulation results and the trigger words of Mind's Eye to understand their effects. We also study whether the correctness of simulation will affect the reasoning performance.

Ablation Settings	Zero-shot	Few-shot
Mind's Eye (*default*)	51.9 $_{1.73}$	84.2 $_{0.79}$
Mismatched simulation results	30.4 $_{1.65}$	54.3 $_{1.28}$
Missing trigger words	51.0 $_{1.12}$	83.0 $_{0.73}$
Incorrect simulation	24.6 $_{1.73}$	39.6 $_{2.89}$

Table 5: The effect of using different sizes of text-to-code models (T2C) with GPT-3 175B/1.3B as the foundation model (FM) in the zero-shot and few-shot settings.

LM Size (T2C + FM)	Zero-shot	Few-shot
0.3B + 1.3B	25.1 $_{1.77}$	43.3 $_{1.12}$
0.3B + 175B	48.3 $_{1.61}$	82.1 $_{0.89}$
1.5B + 1.3B	29.8 $_{1.65}$	46.7 $_{0.82}$
1.5B + 175B (*default*)	51.9 $_{1.53}$	84.2 $_{0.79}$

Can few-shot replace simulation? Given enough in-context demonstrations, can the LM learn how to make its reasoning grounded *internally* without relying on *external* simulations? In Figure 3, we present the dynamics of reasoning performance by gradually including more and more in-context demonstrations in few-shot settings on both vanilla LMs and Mind's Eye augmented LMs. We also design a semi-augmented baseline, Semi-Mind's Eye, which has the Mind's Eye style in-context demonstration, but the final shot omits the simulation result—in other words, the LM has to perform reasoning on the generated grounding evidence by itself in the final step. This baseline differs from vanilla few-shot as it incorporates some simulation results from the physics engine, and it differs from Chain-of-Thought (CoT) since the reasoning steps of CoT are written by humans.

The result demonstrates neither vanilla few-shot (red line) nor semi-augmented few-shot (yellow line) can provide as good an improvement as Mind's Eye (blue line). We also find the few-shot reasoning performance of CoT (green line) and Semi-Mind's Eye (yellow line) has some instabilities which depends on whether the few-shot demonstrations happen to have similar reasoning steps as the given question (final step). The effectiveness of Mind's Eye seems to echo the findings in Zhao et al. (2021), which confirms that the LM tends to ground its reasoning on "recent context", which is coincidentally the simulation results of the given question.

Using a smaller text-to-code LM? Text-to-code LMs convert text-form questions into rendering code. To study its scaling effect, we explore several combinations of using text-to-code LMs and pretrained LMs in different sizes. As shown in Table 5, we find using a smaller text-to-code LM (0.3B) has slightly smaller impact for larger pretrained LMs, as the accuracy decreases by 4.7% for 1.3B pretrained LM but 3.6% for 175B in zero-shot. We also find the conversion error is somehow mitigated by few-shot demonstrations (e.g., the accuracy decreases by 3.6% in zero-shot but by 2.1% in few-shot when the 175B model uses a smaller T2C LM), indicating the benefits on robustness when using a larger pretrained LM.

5 Related Work

Grounded Reasoning. Early attempts to relate language understanding (Winograd, 1972) or learning (Roy & Pentland, 2002) to the physical world mostly rely on manually created linguistic and physical rules (Hermann et al., 2017; Berant et al., 2013). Recent studies have claimed that pre-trained large-scale LMs have already memorized enough world knowledge (Roberts et al., 2020; Brown et al., 2020), and enhanced reasoning ability can be achieved by proper prompting (Nye et al., 2021; Wei et al., 2021; Sanh et al., 2021). Besides adding decomposed reasoning steps (Wei et al., 2022; Zhou et al., 2022), previous work has tried to add hand-written task descriptions (Raffel et al., 2019b) and targeting formats (Marasović et al., 2021) to the prompts, but such human annotation is often costly and expensive at scale (Li & Liang, 2021; Shin et al., 2020). Mind's Eye, instead, presents a new paradigm to ground LM reasoning, via automated simulation rather than human crafted rationales.

Augmented Language Models. Inspired by the evidence that humans require sensory perception for grounded reasoning (Sachs et al., 1981), previous work has tried to augment text inputs to LMs with audio signals (Tsai et al., 2019) or visual perception (Bakhtin et al., 2019; Yi et al., 2018), for improved game playing (Lee et al., 2022), faster language learning (Hill et al., 2020), or better decision making in general (Reed et al., 2022). Our approaches seemingly echoes these findings as we leverage the simulation results as the extra sensory input. However, we do this in a more scalable way. To endow LMs with updated knowledge, TALM (Parisi et al., 2022) fine-tunes LMs interactively with augmented inputs from API calls, which is rather costly compared to prompt-based Mind's Eye. PaLM-SayCan (Ahn et al., 2022) uses the pre-trained PaLM 540B to help robots better understand complex instructions that require reasoning—Mind's Eye can be viewed as the reverse: it infuses knowledge from external tools (i.e., the MuJoCo physics engine) into LMs, to further unlock the the reasoning capabilities of LMs.

Modeling the World. Learning by trial and error, humans seem able to learn enormous amounts of common sense about how the physical world works (Lerer et al., 2016)—such observation has inspired the idea of developing a neural model to model the world (LeCun, 2022; Craik, 1967). World models should be capable of planning (Li et al., 2022a), predicting (Amos et al., 2018), and reasoning (Ullman et al., 2017) through interactions with the world. Similarly, Mind's Eye proposes a paradigm that *reasons* over the experimental results *predicted* by simulation, and the experiments are *planned* beforehand by the text-to-code LM. Using simulated environments to help learning has been widely adopted by research in robotics (MineRL; Guss et al. (2019)) or computer vision (Kubric; Greff et al. (2022)), while our focus is grounded reasoning in the form of natural language.

6 Conclusion and Future Work

We have proposed Mind's Eye, a novel paradigm that enables LMs to ground their reasoning with simulations of the world. We conclude that Mind's Eye is not only effective and scalable but also efficient, as it is able to boost the reasoning performance of small-scale LMs significantly, requiring neither handcrafted prompts nor costly fine-tuning.

We believe the idea of including simulation into the reasoning pipeline can be easily extended to other domains or applications, especially where simulation engines already exist. For example, instead of the physical world, one can use a simulation of economic changes or thermodynamics, aiding policy making or engineering. The dynamic nature of Mind's Eye where we generate grounding evidence unlocks the scaling potential of these models.

ETHICS AND REPRODUCIBILITY STATEMENT

The goal of Mind's Eye is to present a general-purpose paradigm that incorporates knowledge from simulation by professional tools into LM reasoning pipeline. Though we have already seen significant improvement on reasoning with Mind's Eye grounding, the performance can be affected by how accuracy the simulation is. Unqualified simulators may result in wrong simulation results, grounding with which will lead to even worse results than reasoning without grounding. Furthermore, our experiments and analysis are done in English, and therefore we do not claim that our reasoning paradigm will generalize across all languages, although our framework has potential to be extended to other languages by using multilingual LMs.

For reproducibility, we run experiments mainly with publicly available LMs (e.g., GPT-3) and choose baseline methods that have open-sourced implementation. To aid reviewing, we have included the benchmarking version of UTOPIA as well as model outputs in the supplementary materials.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *ArXiv preprint*, abs/2204.01691, 2022. URL https://arxiv.org/abs/2204.01691.

Brandon Amos, Laurent Dinh, Serkan Cabi, Thomas Rothörl, Sergio Gomez Colmenarejo, Alistair Muldal, Tom Erez, Yuval Tassa, Nando de Freitas, and Misha Denil. Learning awareness models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1HhRfWRZ.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross B. Girshick. PHYRE: A new benchmark for physical reasoning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5083–5094, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4191ef5f6c1576762869ac49281130c9-Abstract.html.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6239.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron

Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311, 2022. URL https://arxiv.org/abs/2204.02311.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *ArXiv preprint*, abs/2207.07051, 2022. URL https://arxiv.org/abs/2207.07051.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 2442–2448. ijcai.org, 2019. doi: 10.24963/ijcai.2019/339. URL https://doi.org/10.24963/ijcai.2019/339.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL http://proceedings.mlr.press/v119/guu20a.html.

Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6): 280–285, 2004.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *ArXiv preprint*, abs/1706.06551, 2017. URL https://arxiv.org/abs/1706.06551.

Felix Hill, Andrew K. Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SklGryBtwr.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *ArXiv preprint*, abs/2201.07207, 2022. URL https://arxiv.org/abs/2201.07207.

Mary Kister Kaiser, John Jonides, and Joanne Alexander. Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, 14(4):308–312, 1986.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Neural Information Processing Systems (NeurIPS)*, 2022.

James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *ArXiv preprint*, abs/2204.02329, 2022. URL https://arxiv.org/abs/2204.02329.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. 2022.

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, et al. Multi-game decision transformers. *ArXiv preprint*, abs/2205.15241, 2022. URL https://arxiv.org/abs/2205.15241.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 430–438. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/lerer16.html.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *ArXiv preprint*, abs/2206.14858, 2022. URL https://arxiv.org/abs/2206.14858.

Shuang Li, Xavier Puig, Yilun Du, Clinton Wang, Ekin Akyurek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. Pre-trained language models for interactive decision-making. *ArXiv preprint*, abs/2202.01771, 2022a. URL https://arxiv.org/abs/2202.01771.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. On the advance of making language models better reasoners. *ArXiv*, abs/2206.02336, 2022b.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *ArXiv preprint*, abs/2109.07958, 2021. URL https://arxiv.org/abs/2109.07958.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. Few-shot self-rationalization with natural language prompts. *ArXiv preprint*, abs/2111.08284, 2021. URL https://arxiv.org/abs/2111.08284.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974, 2020.

Mitchell J Nathan, Kelsey E Schenck, Rebecca Vinsonhaler, Joseph E Michaelis, Michael I Swart, and Candace Walkington. Embodied geometric reasoning: Dynamic gestures during intuition, insight, and proof. *Journal of Educational Psychology*, 113(5):929, 2021.

Allen Newell. Physical symbol systems. *Cognitive science*, 4(2):135–183, 1980.

Joseph D Novak and D Bob Gowin. *Learning how to learn*. cambridge University press, 1984.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *ArXiv preprint*, abs/2112.00114, 2021. URL https://arxiv.org/abs/2112.00114.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 823–835, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.63. URL https://aclanthology.org/2021.emnlp-main.63.

Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *ArXiv preprint*, abs/2205.12255, 2022. URL https://arxiv.org/abs/2205.12255.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pp. 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020. URL https://openreview.net/forum?id=025X0zPfn.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, pp. 1–11, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019a.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683, 2019b. URL https://arxiv.org/abs/1910.10683.

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

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot reasoning. *ArXiv preprint*, abs/2202.07206, 2022. URL https://arxiv.org/abs/2202.07206.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *ArXiv preprint*, abs/2205.06175, 2022. URL https://arxiv.org/abs/2205.06175.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1743–1752, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1202. URL https://aclanthology.org/D15-1202.

Jacqueline Sachs, Barbara Bard, and Marie L. Johnson. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, 2(1):33–54, 1981. doi: 10.1017/S0142716400000643.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *ArXiv preprint*, abs/2110.08207, 2021. URL https://arxiv.org/abs/2110.08207.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *ArXiv preprint*, abs/2201.05320, 2022. URL https://arxiv.org/abs/2201.05320.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ICLR*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *Conference on Neural Information Processing Systems (NeurIPS)*, abs/2201.11903, 2022.

Mark Wexler, Stephen M Kosslyn, and Alain Berthoz. Motor processes in mental rotation. *Cognition*, 68(1):77–94, 1998.

Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 153–164, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4c56ff4ce4aaf9573aa5dff913df997a-Abstract.html.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 1039–1050, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5e388103a391daabe3de1d76a6739ccd-Abstract.html.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6720–6731. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00688. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *ArXiv preprint*, abs/2205.10625, 2022. URL https://arxiv.org/abs/2205.10625.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Designing effective sparse expert models. *ArXiv preprint*, abs/2202.08906, 2022. URL https://arxiv.org/abs/2202.08906.

A APPENDIX

A.1 DETAILS ABOUT MIND'S EYE

In Figure A.1 we have presented the whole procedure on how a text question is converted into rendering code for simulation, and how the simulation results are parsed by the manager program to produce simulation based prompt injection.

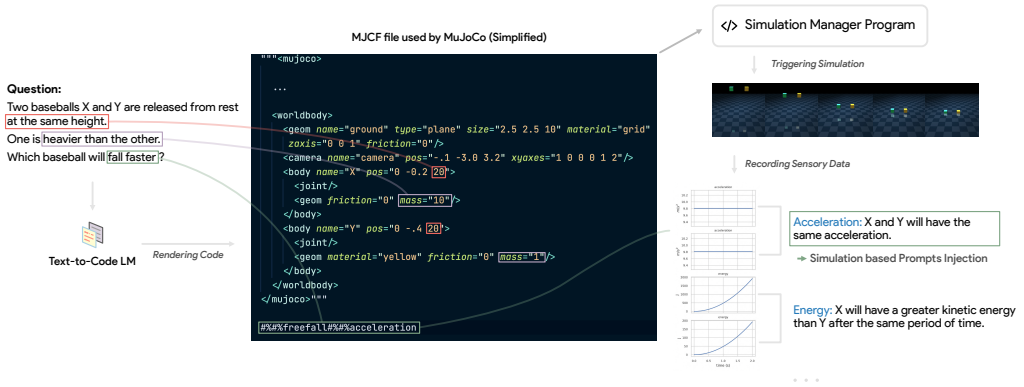

Figure A1: We demonstrate the whole pipeline of how Mind's Eye generates the simulation based prompts injection for a given physics question. We highlight those attributes in the rendering code that reflect the relative relationship in a UTOPIA question. A simulation manager program is responsible for triggering simulation and parse the results. Here we only show two of physics properties that the manager records for demo purpose.

Trained on many text-code pairs, the text-to-code LM is able to generate the corresponding rendering code for the given physics related question. Note that in the generated rendering code, besides the MJCF file itself, in the last line there are special signs #%#% to record the scene name and the physics attributes of interest—these meta-information are added during fine-tuning data (text-code pairs) synthesis stage prepared for the text-to-code LM. After fine-tuning, the text-to-code LM should be able to generate such information.

Once the simulation manager program receives the rendering code, it will execute the simulation based on it. Sensory data such as velocity, acceleration, energy, etc. will be recorded (via MuJoCo APIs). Finally, the manager program will parse the data in terms of the physical properties of interests (coming with the rendering code, marked by #%#%). For example, in Figure A.1, the sensory data is recorded along with the free fall simulation, and since the asked property is "acceleration", the manager program will extract the sensory ground-truth data only on acceleration, and draws the conclusion *"X and Y will have the same acceleration."* by filling in pre-defined "conclusion templates" (e.g., *"The {physics property} of X will be {greater/smaller} than that of Y."*) with observations. Some trigger words will be concatenated before (e.g., *"Hints:"*) the generated conclusion. The connection words (e.g., *"So the answer is:"*) are also added to elicit the final answer. The whole concatenation will be appended after *"Answer:"* as grounded rationale for LM reasoning (as shown in Figure 1 in the main body of the paper).

A.2 SCENE GENERALIZATION WITH MIND'S EYE

By injecting simulation based prompts during LM inference, Mind's Eye does not need fine-tune the LM every time it encounters new contexts. This feature is appealing as we expect Mind's Eye can serve as a powerful tool for building autonomous intelligence that can ground its reasoning or decision making even in unfamiliar contexts. To clearly show the advantages of Mind's Eye on scene generalization, we compare the performance of fine-tuned PaLM 62B with that of vanilla PaLM 62B armed with Mind's Eye.

As shown in Figure A.2, we find the fine-tuned model can achieve good performance on in-domain data, but its performance is quite limited in scenes it has not been fine-tuned on. For example, in Figure A.2 (a), fine-tuned on the Motion scene of UTOPIA can obtain 70% in domain alignment accuracy, but the accuracy decreases to 24% if we test this model on the Projection scene (> 50% performance drop). However, as shown in Figure A.2 (b), Mind's Eye enables the LM to perform consistently well, even if the LM has never seen the scene before.

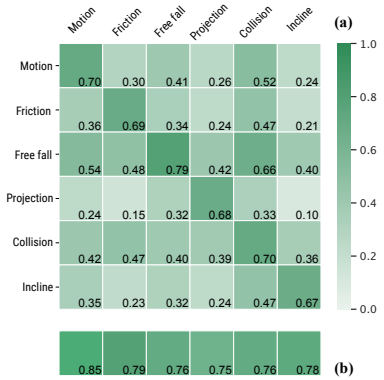

Figure A2: We compare the performance of (a) fine-tuned 62B PaLM on six scenes of UTOPIA data respectively with (b) vanilla 62B PaLM armed with Mind's Eye in zero-shot. Though fine-tuning can achieve good performance on in-domain data, it can hardly generalize to OOD cases (i.e., scenes on which the LM has not been fine-tuned). Prompting based Mind's Eye can enable large LMs generalize to unseen new scene with consistent alignment performance even in zero-shot settings. In (a), the performance of fine-tuned PaLM 62B on in-domain test sets is in diagonal, while that of out-of-domain is the rest. The zero-shot performance of PaLM 62B + Mind's Eye in six scenes is listed in (b).

A.3 MIND'S EYE ERROR ANALYSIS

In Figure A.3 we show three error cases when using Mind's Eye. Ignorance error refers to the reasoning apparently ignores the given grounded rationale. Recency bias means the LM tends to extract the final answer from nearby words (e.g., the object Y). "I doń know" error are the cases where the LM simply generates non-sense when unable to reason about the question. All these cases are much less common when we run experiments on larger LMs.

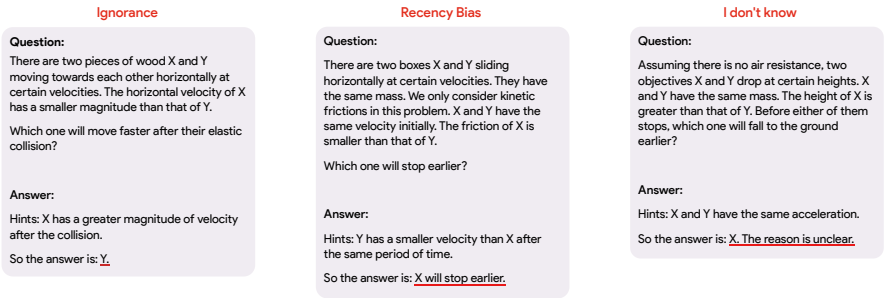

Figure A3: We exemplify several error cases when using Mind's Eye. In general, we find these errors mostly happen in small LMs, because their reasoning ability over given evidence is limited.