AuPPLE: Augmented Physical Priors through Language Enhancement using Self-Supervised Learning

Annie Dong*
Watchung Hills Reg. HS
Warren, United States
annieyushan@gmail.com

Anirudh Mazumder*
Texas Academy of Math & Science
Denton, United States
anirudhmazumder26@gmail.com

Mustafa Efe Güzel
TED Sakarya High School
Sakarya, Turkey
mustafaefeguzel@proton.me

Badis Labbedi
United World College Costa Rica
San José, Costa Rica
labbedi75@gmail.com

Zhuo En Chua
SJII
Singapore, Singapore
chuazhuoen@gmail.com

Eveer Soriano
Ameritec School
San Salvador, El Salvador
eveer.soriano@icloud.com

Priyanshu Sethi
The University of Alabama
Tuscaloosa, United States
sethipriyanshu@outlook.com

Michael Lutz
The University of California, Berkeley
Berkeley, United States
michaeljlutz@berkeley.edu

***Abstract*—In recent years, a contentious debate has emerged surrounding the degree to which Large Language Models (LLMs) can truly achieve grounding in the physical world. Grounding, in this context, refers to the models' ability to establish a meaningful connection between their language-based understanding and a concrete comprehension of real-world phenomena. Our research aims to explore the latent capability of LLMs to develop physical intuition: a prerequisite for embodied agents to effectively perform tasks in real-world environments. In this paper, we release a novel dataset of physical scenarios that serve as a benchmark for an LLMs' physical intuition. Our benchmark AuPPLE (Augmented Physical Priors through Language Enhancement) for Language Models includes scenarios regarding free-fall and projectile motion, including various question-answer formulations: MultiQA, binary classification, and continuous number prediction to comprehend linguistic nuances and apply their understanding within a physical context. By meticulously fine-tuning LLMs on this specialized dataset, we assess their performance in providing responses that showcase an ability to draw upon underlying physical knowledge. With our fine-tuned LLMs achieving over 87%—more than 3 times its base model—on free-fall evaluation dataset, our results shed light on the intrinsic grounding capabilities of LLMs, offering insights into their potential to bridge the gap between language and the physical world. This paper contributes to the ongoing discourse on the true nature of LLMs' comprehension and its relationship with real-world context, underscoring the strides made in enhancing their intuitive understanding through targeted fine-tuning techniques.**

***Index Terms*—physics, large language models, grounding, intuition, artificial intelligence**

I. Introduction

Human beings possess an innate ability to intuitively grasp the workings of the physical world, enabling us to effortlessly navigate various real-world scenarios. From fundamental concepts like free-fall to more complex notions such as projectile motion, our natural understanding of physical principles equips us to comprehend the mechanics of the tangible realm. For instance, consider the task of estimating the time it takes for a ball to descend from a 5-meter-high window. Drawing upon past experiences and latent physical comprehension, we can envision the ball's trajectory under the influence of gravity, yielding a confident prediction that it will reach the ground in approximately 1 second.

However, the question arises: can LLMs develop a similar intuition about the physical world? In recent years, substantial advancements and research efforts have showcased LLMs' remarkable linguistic capabilities in natural language processing, text completion, and language translation. Yet, a critical observation has emerged from this surge of research: LLMs currently lack the capability to effectively ground themselves in the physical realm. Unlike humans, these models struggle to understand fundamental physical attributes such as object location, weight, height, and other pertinent properties. Scenarios involving embodied agent tasks, particularly those necessitating continuous control, present inherent challenges for LLMs without a solid grasp of the physical environment. To address this limitation, our research delves into the fine-tuning of LLMs, using prompts to assess their capacity for physical intuition.

Various approaches have made endeavors to tackle the understanding of the physical world within the framework of LLMs. Although they have all produced substantial performance improvements in the realm of physical phenomena, these approaches tend to rely on external resources or tools for guiding LLMs in physical reasoning, rather than cultivating an innate intuitive response.

One promising approach that has emerged is the use of simulated representations of the physical world, exemplified by the innovative frameworks of PiLoT [1] and Mind's Eye [2]. PiLoT (Physics in a Language of Thought) establishes a grounded link between language and probabilistic programs

* Authors Contributed Equally

using a physics engine. Leveraging the Box2D game engine, PiLoT simulates a physical world, generating responses to physics prompts correlating significantly with human judgments [1]. Mind's Eye, by integrating simulations into reasoning via reflection, enhances LLMs' understanding and reasoning of physical phenomena. The Mind's Eye results are promising, with a 27.9% average increase in zero-shot performance and a 46.0% average enhancement in few-shot performance. Although tools like REALM [13], RAG [14], RETRO [15], and others have been used for prompting, these methods require costly external physics engines and do not intrinsically improve the model's intuitive grasp of the physical world, a prerequisite model property we justify in this paper.

While addressing Mathematical Word Problems (MWPs), state-of-the-art language models like ChatGPT [16] have shown subpar performance in precisely answering MWP prompts [3]. However, recent advances have spotlighted the benefits of using chain-of-thought prompts to refine response accuracy in this domain. A notable illustration is MathPrompter, which adopts a unique approach to tackle MWPs. By employing Zero-shot chain-of-thought T (175B parameters) prompting, MathPrompter bolsters the abilities of the model by producing multiple algebraic expressions and Python functions in response to a single problem [5]. Training verifiers further enhance MathPrompter's precision, supervising the multi-step problem-solving process, and proving effective in heightening accuracy [17]. This supplementary training mechanism has proven its efficacy, surpassing the performance of other existing models in solving MWPs.

In the context of leveraging LLMs to support embodied agents, researchers have started exploring the integration of LLMs with physical tasks. Notably, is the development of SayCan [7], which introduces a language-based affordance model to determine task feasibility. SayCan empowers LLMs to engage with physical environments and execute actions guided by spoken instructions. This approach effectively bridges the gap between language understanding and physical execution. Similarly, LLM-Planner [6] is relevant in this context, utilizing commonsense knowledge to facilitate task planning. By combining high-level and low-level planning, LLM-Planner orchestrates tasks toward desired objectives. However, the absence of proper grounding in the physical world can hinder task execution. An inadequate grasp of physical properties, constraints, and interactions may compromise the performance of LLM-based systems in physical tasks. Improved contextual grounding, on the other hand, can enhance LLM-based systems' navigation and interaction with the physical world. This advancement holds potential across diverse domains, including robotics, automation, and smart environments.

Our objective entails two primary goals. Firstly, we strive to enhance the accuracy of Language Models in responding to physical problems through the fine-tuning process. Secondly, we aim to achieve this goal without relying on external tools or resorting to chain-of-thought prompting techniques. By pursuing this approach, we aim to bolster an intuitive physical grounding within LLMs, facilitating a more comprehensive understanding of the physical world.

To evaluate the efficacy of our approach, we conduct a thorough performance analysis of the fine-tuned models on our benchmark dataset that necessitate a keen sense of physical intuition, we gain valuable insights into the capabilities and limitations of the fine-tuned LLMs in terms of their physical reasoning and problem-solving ability. The findings from this analysis serve to inform further advancements in enhancing the physical grounding of LLMs, contributing to the development of more robust and capable models in the domain of physical understanding. We strive to develop more advanced and contextually grounded language models that can effectively bridge the gap between natural language processing and the physical world.

The rest of this paper is outlined as follows: Section II presents the dataset creation process in which we discuss the method we used to create the questions to feed into the LLMs. Section III presents the training of the LLMs with the dataset and the results we got on its understanding of the physical world. Section IV provides a conclusion of our work, the impacts of our work, and potential future works.

II. Method

A. AuPPLE Benchmark Dataset

In this study, we present a comprehensive benchmark dataset meticulously tailored to facilitate the fine-tuning process of LLMs for effectively addressing physics-related questions. The core objective of this benchmark is to gauge the LLM's competence in comprehending and accurately responding to an array of diverse physics questions, subsequent to undergoing the fine-tuning process.

While acknowledging the presence of other benchmark datasets, like the Utopia dataset cultivated by the Mind's Eye Language Model team [2], our contribution introduces a distinctive benchmark. Similar to their approach, our dataset also encompasses physics-oriented questions categorized into discrete scenes. However, our benchmark takes a nuanced stance by adopting the viewpoint of an embodied agent, a perspective that contributes towards the ongoing integration of physical grounding in these technologies.

B. Data Augmentation

The question base was created through data augmentation in which a set of template questions was transformed into thousands of training examples, via a script that utilized the correct physical equations and modeling practices.

The algorithm employed for augmenting free-falling questions operates by creatively combining objects, drop heights, question templates, and physics computations to produce new variations. The process begins by initializing the algorithm with a selection of predefined question templates that include placeholders for the falling object and the drop height. Next, lists containing potential objects and height boundaries are designated from which values can be randomly sampled. These objects span a diverse array of items that could be dropped,

TABLE I
PHYSICS QUESTIONS AND ANSWERS

Concept	Type	Question	Answer Choices	Answer
Free-Fall	Generic	A cup is dropped from a counter that is 2.7 meters above the ground, how many seconds will it take for the cup to reach the ground?	A) 1.04 seconds B) 0.74 seconds C) 1.1 seconds D) 1.03 seconds	B) 0.74 seconds
	Physical World	You are a robot in the physical world, and you see a laptop bag fall from a height of 4.0 meters. How long will it take to hit you?	A) 0.9 seconds B) 1.48 seconds C) 1.24 seconds D) 1.03 seconds	A) 0.9 seconds
Projectile	Generic	A ping pong ball is launched forward and upward from the ground with an angle of 10.65 degrees and an initial velocity of 45.97 m/s. How far will it travel before it hits the ground?	A) 92.68 meters B) 108.91 meters C) 120.63 meters D) 78.25 meters	D) 78.25 meters
	Physical World	A golf ball is launched forward and upward from the ground with an angle of 11.64 degrees and an initial velocity of 19.22 m/s. How far will it travel before it hits the ground?	A) 13.87 meters B) 16.11 meters C) 14.88 meters D) 24.56 meters	C) 14.88 meters

while the height bounds establish minimum and maximum heights for each object category.

The algorithm proceeds into a loop to generate individual questions. During each iteration, it randomly selects one template, an object, height bounds, and samples a height value from within the specified bounds. Leveraging the equations of motion for free-fall, the algorithm calculates the expected fall time corresponding to the sampled height. The chosen object, height, and computed fall time are then integrated into the selected template to forge the complete question text.

The algorithm's randomized sampling strategy ensures the generation of an extensive assortment of unique question variations without repetition. For this study, the algorithm was configured to perform 10,000 iterations, generating 10,000 unique free-fall physics questions. This question pool was subsequently divided into 90% for training and 10% for testing, facilitating evaluation through the LLMs.

Additionally, beyond the question text, the algorithm produces four multiple-choice answers for each question. These answers are generated by deliberately introducing variations in the fall-time solution. Out of these options, one is designated as the correct answer. This randomized approach to answer generation systematically crafts an ample and diverse dataset of free-fall problems, accompanied by accurate answers and solutions, in an unbiased manner.

By introducing variability across objects, heights, and templates, the algorithm ensures a wide-ranging distribution of question types without confinement to any specific format. This comprehensive and dynamic approach underscores the algorithm's capacity to create a versatile array of free-fall physics questions and corresponding answers with a high degree of accuracy and variability. Table 1 shows an example free-falling, generic question and answer.

Furthermore, following the creation of the initial question base, a more targeted test set of 1,000 questions was generated to assess the LLM's comprehension and assimilation of the physical world. The methodology for generating these questions remained consistent with the approach outlined above, however, we altered the perspective from which the questions were formulated. Instead of assuming the role of a passive observer witnessing objects fall to the ground, the approach shifted to casting the LLM in the role of an active participant, similar to a robotic entity, reflected by the prefix and suffix changes in Figure 1.

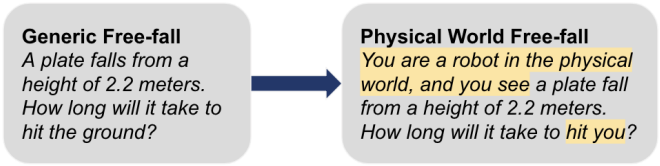

Fig. 1. Nuanced AuPPLE dataset of physical-world free-fall questions.

This nuance in perspective allowed us to evaluate the LLM's intrinsic grasp of the physical realm more rigorously. By considering the LLM as an active entity situated within the physical world, facing an object falling towards it, we aimed to gauge whether the LLM could genuinely anticipate the moment of contact between the falling object and itself. In essence, this approach served as a litmus test for the LLM's authentic understanding of physical principles. Table 1 shows an example free-falling, physical-world question and answer.

In the pursuit of evaluating the extent to which LLMs can comprehend projectile motion—a more intricate physical concept—we adopted a similar methodology. We crafted two distinct projectile-motion datasets: one dataset for fine-tuning consisting of 10,000 generic questions and one dataset for testing-only consisting of 1,000 questions real-world questions. In contrast to the variables of height and fall time employed in the free-fall dataset, our focus shifted to angle in degrees and initial velocity for the projectile-motion inquiries. For the dataset in order to be used for finetuning, we adhered to our ratio of 9:1 for robust evaluation. Example questions for projectile motion questions are shown in Table 1.

C. Training

Our approach involves the fine-tuning of flan-t5-large, a pre-trained autoregressive model with an encoder-decoder architecture. The fine-tuning process hones its weights and parameters on a dataset tailored to our specific task. Operating in an auto-regressive manner, this model generates outputs sequentially, with each step informed by the preceding outputs. Data preprocessing structured the dataset into a sequence-to-sequence (seq2seq) architecture, with essential features including sentence, answer, input_ids, attention_mask, and labels.

The language model's pre-existing linguistic knowledge is augmented with a domain-specific understanding of physics concepts. Combining linguistic and scientific comprehension is vital for creating language models that can not only provide textual answers but also understand the physical implications and nuances embedded within the questions.

During the fine-tuning process, we tailored hyperparameters to our needs: a learning rate of 1e-3, a batch size of 4, and a solitary epoch. The core objective of fine-tuning was to heighten the accuracy of selecting the correct answer among the presented multiple-choice options. Thus, this self-supervised framework facilitated learning from unlabeled data by enabling the model to establish its own targets through predictive capabilities.

Our model underwent training on both the free-falling and projectile motion generic AuPPLE benchmark training datasets. In the case of each dataset, comprising 9,000 distinct questions, our model was primed to excel in understanding and responding to physics-related queries spanning a diverse spectrum of scenarios.

Furthermore, we extended this methodology to train the model on the AuPPLE free-fall physical world, augmenting its competence in grasping physics concepts within distinct real-world contexts.

D. Evaluation

Before collecting the model's responses from the evaluation dataset, we subjected our fine-tuned model to a qualitative assessment. This involved selecting a sample from the dataset with a positive label and observing the model's ability to generate an answer. Remarkably, the model consistently chose the correct answer even when the position of the correct answer choice was altered within the list of options. This qualitative testing underscores our findings that the fine-tuned model genuinely acquired an intrinsic understanding of the physical world, rather than relying on mere guesswork.

Subsequently, we assessed the model's performance on the previously unseen test dataset containing 1,000 unique multiple-choice questions about generic free-fall by intuitively selecting the answer choice that the model deemed correct. This evaluation would gauge the model's ability to generalize its learning to new instances effectively. We then compared the model's predicted answers to the correct answers within the benchmark, enabling us to quantify the accuracy of the model's responses. This evaluation process was repeated for the generic projectile dataset.

The overarching workflow of our approach is visually depicted in the flowchart of Figure 2.

Furthermore, to gauge the model's performance in handling out-of-distribution (OOD) data, we conducted an assessment involving our model, which had been trained on the generic dataset. This evaluation entailed testing the model on the physical-world free-fall dataset, which encompassed scenarios beyond its original training domain.

III. Results

A. Baseline

TABLE II
BASELINE ACCURACY TEST RESULTS

Model	Training Dataset	Percentage accuracy
GPT 3.5	Generic Free-Fall	24.58%
	Free-Fall in the Physical World	25.63%
	Projectile Motion	28.30%
GPT 4	Generic Free-Fall	30.10%
	Free-Fall in the Physical World	29.74%
	Projectile Motion	25.00%
FLAN-T5-Large	Generic Free-Fall	25.80%
	Free-Fall in the Physical World	23.70%
	Projectile Motion	23.50%

The base-FLAN-T5 model, although capable of handling language-based tasks effectively, exhibited limited proficiency in answering domain-specific questions. It achieved an accuracy of 25.80% on our generic free-fall benchmark, which falls short of the fine-tuned LLM performance. Even state-of-the-art GPT-4 demonstrated slightly better performance in answering physics-related multiple-choice questions, with 30.10% proving that these LLMs cannot ground in the physical world.

B. Fine-tuned Model

The fine-tuned FLAN-T5-Large auto-regressive model exhibited significantly enhanced performance in addressing physics-related multiple-choice questions. Following the fine-tuning process, our model showcased an impressive accuracy of 87.7%, thereby confirming our research hypothesis that fine-tuning can significantly enhance the precision of LLMs. This outcome underscores the proficiency of the FLAN-T5-Large model in generating accurate answers, achieved through the utilization of a dataset of only 10,000 examples and hyperparameter tuning. As depicted in Figure 3, our model consistently outperformed the baseline models, achieving more than threefold accuracy improvements in the context of free-fall questions and approximately twofold improvements in projectile motion questions.

C. Generalizability

To assess the extent to which our finetuning process produces in generalizable results as opposed to mere memorization of specific scenarios, we conducted a data ablation study. In essence, our approach involved training the model using

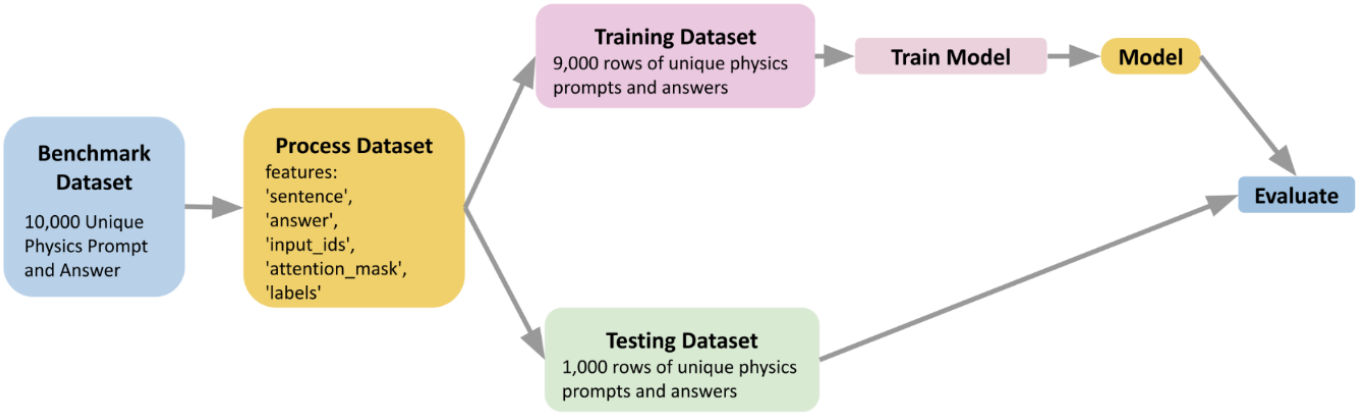

Fig. 2. Fine-tuning Process Flowchart.

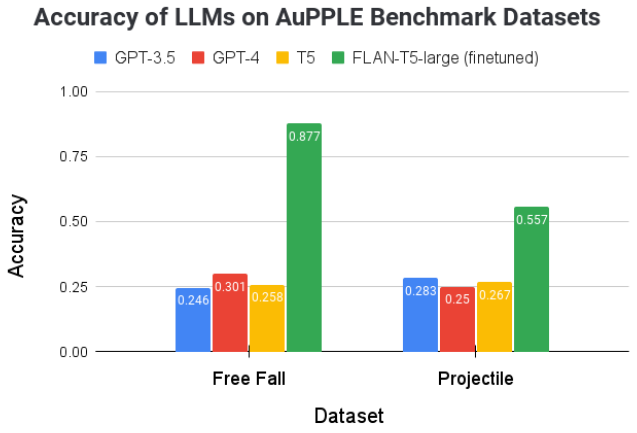

Fig. 3. Accuracy of our fine-tuned model compared to other state-of-the-art LLMS, on generic free-fall AuPPLE Benchmark dataset

the initial free-falling dataset. Subsequently, we evaluated the model's performance using an entirely distinct set of questions, encompassing real-world scenarios wherein an object possesses the capacity to descend onto the model. Furthermore, as depicted in Figure 4, our fine-tuned model showcased its remarkable adaptability in handling out-of-distribution (OOD) data. Even when confronted with substantial alterations in the question format, our model consistently achieved an impressive accuracy of 87.7%. This outcome underscores the transferability and robustness of our results, indicating that the success was not merely a result of overfitting to the provided questions in the training set.

IV. Conclusion

A. Conclusion

In this paper, we present a compelling demonstration that large language models can develop a robust and nuanced intuitive understanding of the physical world through comprehensive training on a diverse dataset encompassing real-world object dynamics and phenomena. By fine-tuning the state-of-the-art FLAN-T5-large model using our extensive custom dataset of multiple-choice questions centered on foundational physics concepts, we enabled the model to achieve accurate predictions about the behavior of objects and systems across a broad range of scenarios necessitating fundamental physical reasoning.

Our qualitative evaluations, illustrated in Figure 4, shed light on the answering patterns of base-FLAN-T5 and GPT-3.5, highlighting their distinct behaviors and decision-making processes. Base-FLAN-T5 consistently favoring option C) in uncertain scenarios suggests a bias or tendency default to a specific choice, which could arise from training data or architectural influences. This pattern indicates that the model has absorbed commonalities or biases within the dataset, influencing its responses. Conversely, GPT-3.5's tendency to guess randomly when uncertain indicates a lack of a consistent strategy for addressing unfamiliar questions. This could be attributed to its generative nature, resulting in diverse outputs when confronted with information gaps.

The remarkable performance exhibited by our fine-tuned model across an extensive array of tests and assessments validates the central hypothesis driving this research—the intrinsic architectural potential of large language models to acquire physical intuitions given an ample supply of training data. Our model's accuracy on intuitive physics problems, akin to those that humans typically solve based on real-world experience, closely approached average human performance, surpassing 87% accuracy on the testing dataset. This suggests that the model has successfully internalized meaningful patterns and dynamics, enabling it to reason about physical phenomena at a level akin to human intuitions, transcending mere memorization of superficial training instances.

B. Future Works

First, the training dataset could be expanded in terms of the diversity and complexity of physical concepts, situations, and interactions covered. Our current dataset, while large, only encompasses fundamental mechanics, basic object motions, and

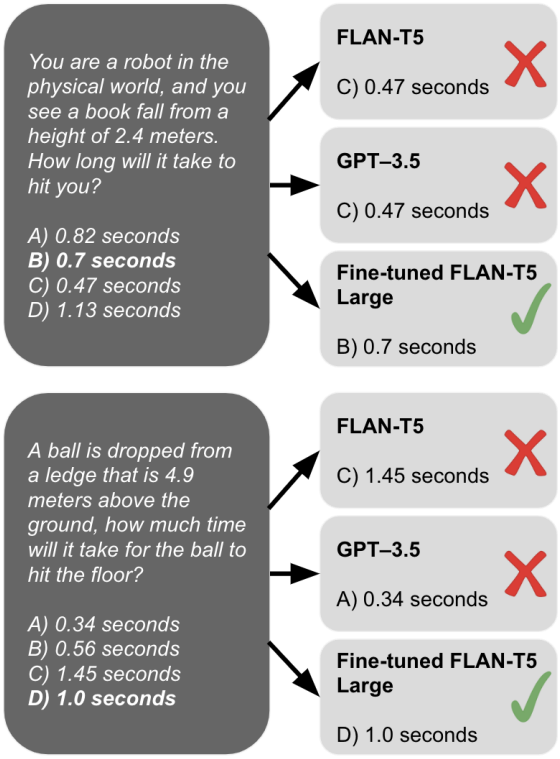

Fig. 4. State-of-the-art models and AuPPLE fine-tuned model answering an example AuPPLE (top) generic and (bottom) physical world free-fall benchmark question.

rudimentary interactions. Constructing vastly more comprehensive training data covering advanced topics like thermodynamics, electromagnetism, optics, fluid dynamics, etc. would spur the development of commensurately more sophisticated physical intuitions. Furthermore, generating ultra-high fidelity simulations and hyper-realistic interactive virtual environments to supply experiential training data could dramatically enrich the models' mastery of real-world physical phenomenology.

Second, devising methods to seamlessly integrate structured scientific knowledge into model training could enable more systematic, hierarchical acquisition of physics knowledge and reasoning abilities. Our current approach relies exclusively on unstructured natural language. Incorporating symbolic knowledge representations in the form of logic formulas, computational ontologies, and scientific modeling primitives could allow the explicit encoding of fundamental principles to complement implicit pattern recognition. This fusion of data-driven deep learning with formally structured knowledge graphs is a supremely promising direction.

Finally, benchmarking the real-world performance of intuitive physics models on challenging embodied robotics tasks could both demonstrate immense practical value and provide crucial feedback to further refine the models. Training the models on simulated interactive robotic control challenges involving complex object manipulations and physics-based reasoning, then transferring them to real-world physical robot platforms would constitute an ideal testbed. The cycle of iteration between robotics applications and expanded model training could ultimately yield artificial agents with both general purpose and specialized physical intuitions surpassing even human capabilities.

Acknowledgments

We would like to acknowledge Blast AI for imparting AI basics and laying the foundation for our work. Their support and resources were pivotal in our research.

References

[1] R. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, "Mind's Eye: Grounded Language Model Reasoning Through Simulation," in Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, pp. 2206–2240, PMLR, 2022. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html
[2] "Grounded Physical Language Understanding with Probabilistic Programs and Simulated Worlds," [Online]. Available: https://nesygems.github.io/assets/pdf/papers/Grounded.pdf
[3] "Idea: train an LLM on output of physical simulation," [Online]. Available: https://arxiv.org/pdf/2210.05359.pdf
[4] "An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP)," [Online]. Available: https://arxiv.org/abs/2302.13814
[5] "How well do Large Language Models perform in Arithmetic tasks?" [Online]. Available: https://arxiv.org/pdf/2304.02015.pdf
[6] "MathPrompter: Mathematical Reasoning using Large Language Models," [Online]. Available: https://arxiv.org/abs/2303.05398
[7] "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models," [Online]. Available: https://arxiv.org/abs/2212.04088
[8] "Do as I Can, not as I Say," [Online]. Available: https://say-can.github.io/
[9] "LoRA," [Online]. Available: https://arxiv.org/abs/2106.09685
[10] "Parameter Efficient Fine-Tuning libraries," [Online]. Available: https://github.com/huggingface/peft
[11] "https://github.com/lxe/simple-llm-finetuner"
[12] "GPT-3 Grounding LLMs," [Online]. Available: https://arxiv.org/abs/2302.02662
[13] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event, vol. 119 of Proceedings of Machine Learning Research, pp. 3929–3938, PMLR, 2020. [Online]. Available: http://proceedings.mlr.press/v119/guu20a.html
[14] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, virtual, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
[15] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. Bm Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," in Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, pp. 2206–2240, PMLR, 2022. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html
[16] K. Cobbe et al., "Chatgpt: Optimizing language models for dialog," arXiv preprint arXiv:2110.14168, 2021.
[17] Sourab Mangrulkar and Sylvain Gugger and Lysandre Debut and Younes Belkada and Sayak Paul, "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods." 2022. [Online] https://github.com/huggingface/peft