Towards Grounding Large Language Models with the Physical World

Bohan Lyu[*1] **Yadi Cao**[2]
Duncan Watson-Parris[2] **Leon Bergen**[2] **Taylor Berg-Kirkpatrick**[2] **Rose Yu**[2]
[1] Tsinghua University [2] University of California, San Diego

Abstract

Due to the lack of a physics-grounded dataset, and the limitation of the seemingly correct human preferences, existing supervised fine-tuning and preference optimization methods for large language models (LLMs) struggle with complex physical systems. To address these challenges, we propose a new paradigm, **Physics-Informed Fine-Tuning (PIFT)**. This approach consists of three main steps: 1. creating a dataset using world simulators, 2. performing World Knowledge Distillation, and 3. conducting World Preference Learning. Experiments on climate-related problems demonstrate PIFT's effectiveness, with our models outperforming state-of-the-art models like GPT-4o and Claude-3.5.

Large language models (LLMs) demonstrate robust analysis and reasoning capabilities in daily scenarios. These models have shown proficiency in addressing simple school-level scientific problems (Cobbe et al., 2021; Hendrycks et al., 2021). However, our preliminary experiments reveal that LLMs cannot solve problems derived from complex physical simulations, such as soft and rigid bodies, fluid dynamics, climate science, and epidemiology.

Supervised fine-tuning methods for LLMs predominantly rely on 1. leveraging existing corpora for additional dataset construction (Brown et al., 2020), 2. knowledge distillation from more advanced models (Sanh et al., 2019), and 3. rule-based sampling of model outputs (Bai et al., 2022). Scholars further enhanced fine-tuned LLMs using preference learning, improving their instruction-following ability (Ouyang et al., 2022). These approaches involve labeling the preferences of different responses, either from human feedback or from another LLM (Lee et al., 2023).

However, our preliminary results revealed that both above techniques fail to answer questions that involve complex physics. We attribute the failures to: 1. Existing training datasets (Sun et al., 2023; Wang et al., 2024; Zhong et al., 2023; Arora, Singh, and Mausam, 2023) primarily cover school-level scientific problems. These scenarios, while offering clear-cut solutions, often fail to capture real-world complexities. 2. Existing methods align LLMs with human preferences, which, while often intuitively correct, may not always adhere to true physical laws (Ouyang et al., 2022).

These limitations motivate us to enable physics-grounded LLMs. To achieve this goal, we propose Physics-Informed Fine-Tuning (PIFT) technology, which includes 3 steps, as shown in Figure 1. We first construct a comprehensive dataset involving complex and typical physical systems in the real world with corresponding simulators. The physical systems include rigid- and soft-body dynamics (Todorov, Erez, and Tassa, 2012), fluid dynamics (Kochkov et al., 2021; Dresdner et al., 2022), climate science (Niu et al., 2024), and epidemiology (Wu et al., 2023). We create two types of questions: multiple-choice questions and open-ended questions. The correct choice for Multiple-choice questions is generated during the question synthesis. For open-ended questions, such as those involving planning and causal inference, we prompt a critic LLM to use domain-specific simulators to evaluate the quality of responses generated by the LLM being fine-tuned. Additionally, we sample a portion of these evaluation scores for human expert validation, ensuring the validity of our pipeline.

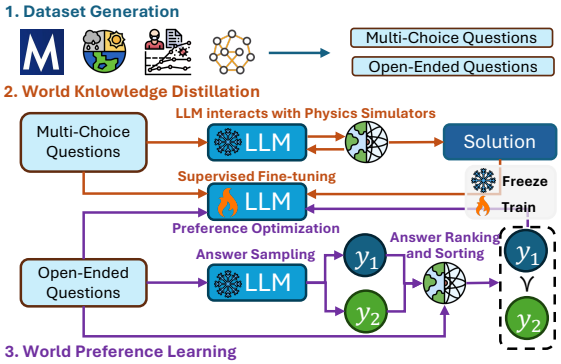

Figure 1: The 3 steps of PIFT pipeline.

Secondly, we design **World Knowledge Distillation (WKD)**, where we prompt LLMs to analyze problems using results from world simulators and derive detailed solution steps correspondingly. Then, we filter the generated solutions to retain only those that align with correct answers and use these filtered solutions to fine-tune the target LLM.

Lastly, we propose **World Preference Learning (WPO)**, where the critic LLM utilizes world simulators to rank different responses for open-ended problems. This ranking procedure involves designing proper domain-specific criteria based on the simulated results, such as ranking the cost or reward in the planning settings. We then employ Direct Preference Optimization (DPO) (Rafailov et al., 2024) to train the target LLM with these ranked responses.

Preliminary experiments on climate problems demonstrate our methods' effectiveness, while untrained `Llama-3.1-8B-Instruct` has a 32.2% accuracy, `GPT-4o` shows 51.1% and `Claude-3.5` shows 37.8%. `Llama-3.1-8B-Instruct` trained with WKD reaches 47.8% compared with 43.3% if just trained with correct choices. Additional WPO increases this score to 55.6%.

[*]Work done while visiting UCSD

References

Arora, D.; Singh, H.; and Mausam. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7527–7543. Singapore: Association for Computational Linguistics.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Dresdner, G.; Kochkov, D.; Norgaard, P.; Zepeda-Núñez, L.; Smith, J. A.; Brenner, M. P.; and Hoyer, S. 2022. Learning to correct spectral methods for simulating turbulent flows.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874.

Kochkov, D.; Smith, J. A.; Alieva, A.; Wang, Q.; Brenner, M. P.; and Hoyer, S. 2021. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21).

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267.

Niu, R.; Wu, D.; Kim, K.; Ma, Y.-A.; Watson-Parris, D.; and Yu, R. 2024. Multi-Fidelity Residual Neural Processes for Scalable Surrogate Modeling. In *International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sun, L.; Han, Y.; Zhao, Z.; Ma, D.; Shen, Z.; Chen, B.; Chen, L.; and Yu, K. 2023. SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research. *arXiv preprint arXiv:2308.13149*.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of the Forty-First International Conference on Machine Learning*.

Wu, D.; Niu, R.; Chinazzi, M.; Vespignani, A.; Ma, Y.-A.; and Yu, R. 2023. Deep Bayesian Active Learning for Accelerating Stochastic Simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364.