SimPal: Towards a Meta-Conversational Framework to Understand Teacher's Instructional Goals for K-12 Physics

Effat Farhana
effat.farhana@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Souvika Sarkar
szs0239@auburn.edu
Auburn University
Auburn, Alabama, USA

Ralph Knipper
rak0035@auburn.edu
Auburn University
Auburn, Alabama, USA

Indrani Dey
idey2@wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

Hari Narayanan
naraynh@auburn.edu
Auburn University
Auburn, Alabama, USA

Sadhana Puntambekar
puntambekar@education.wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

Santu Karmaker
sks0086@auburn.edu
Auburn University
Auburn, Alabama, USA

ABSTRACT

Simulations are widely used to teach science in grade schools. These simulations are often augmented with a conversational artificial intelligence (AI) agent to provide real-time scaffolding support for students conducting experiments using the simulations. AI agents are highly tailored for each simulation, with a predesigned set of Instructional Goals (IGs), making it difficult for teachers to adjust IGs as the agent may no longer align with the revised IGs. Additionally, teachers are hesitant to adopt new third-party simulations for the same reasons. In this research, we introduce SimPal, a Large Language Model (LLM) based meta-conversational agent, to solve this misalignment issue between a pre-trained conversational AI agent and the constantly evolving pedagogy of instructors. Through natural conversation with SimPal, teachers first explain their desired IGs, based on which SimPal identifies a set of relevant physical variables and their relationships to create symbolic representations of the desired IGs. The symbolic representations can then be leveraged to design prompts for the original AI agent to yield better alignment with the desired IGs. We empirically evaluated SimPal using two LLMs, ChatGPT-3.5 and PaLM 2, on 63 Physics simulations from PhET and Golabz. Additionally, we examined the impact of different prompting techniques on LLM's performance by utilizing the TELeR taxonomy to identify relevant physical variables for the IGs. Our findings showed that SimPal can do this task with a high degree of accuracy when provided with a well-defined prompt.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**;
• **Applied computing → Education**.

L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-0633-2/24/07
https://doi.org/10.1145/3657604.3664695

KEYWORDS

Large Language Models, Conversational AI, Meta-Conversation, K-12 Science

ACM Reference Format:
Effat Farhana, Souvika Sarkar, Ralph Knipper, Indrani Dey, Hari Narayanan, Sadhana Puntambekar, and Santu Karmaker. 2024. SimPal: Towards a Meta-Conversational Framework to Understand Teacher's Instructional Goals for K-12 Physics. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18–20, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3657604.3664695

1 INTRODUCTION

Simulations are widely used in science education, and prior research shows that using simulations in science education can enhance students' comprehension of scientific concepts [21, 28]. However, students often need guidance and scaffolding when conducting experiments with simulations [14, 15], and it is challenging for one teacher to provide real-time support to multiple students simultaneously [11]. Recent advancements in Large Language Models (LLMs) [5] have revolutionized conversational AI agents as a plausible solution to provide real-time support to students. But LLM-powered conversational AI agents also present unique challenges. First, existing AI agents are highly customized for a specific simulation with a predesigned set of Instructional Goals (IGs) [12]. Therefore, teachers often struggle to edit these predesigned IGs or redesign the IGs because the AI agent will no longer be aligned with the revised IGs. Second, middle or high school science teachers lack the technical expertise to customize AI agents [25]. This leads to the use of pre-existing, non-customizable agents or third-party software, which requires more time and resources for simulations. For similar reasons, teachers also hesitate to integrate new/other third-party (closed-source) simulations into their instructional materials.

How can we empower teachers to integrate any third-party (open or closed-source) simulation into their instruction materials such that they can I) freely design their own Instructional Goals (IGs) and II) quickly customize a conversational AI agent to better align with their IGs? More importantly, how can we achieve this goal without requiring teachers to understand the technical details of Large Language Models (LLMs) like GPT-4 [1] and PaLM [2, 7]? While LLMs are trained on vast internet text data and can aid in language comprehension tasks like answering questions [23] and facilitating human conversations [30], adapting LLMs to domain-specific tasks is still challenging due to a lack of proper knowledge grounding in that particular domain. It is also unrealistic to expect school teachers to learn knowledge-grounding techniques that require in-depth machine learning or deep learning knowledge.

This paper introduces SimPal, a meta-conversational agent that can assist school teachers in adopting any existing physics simulation into their lesson plan while allowing them to custom-design their own IGs and customize a general-purpose LLM that aligns with those custom IGs, facilitating *instruction at scale*. SimPal achieves this ambitious goal through *meta-conversation*, which is essentially a conversation with the teacher about structuring future conversations with students for simulation-based physics experiments. Through natural (meta-)conversation with SimPal, teachers first explain their desired IGs, based on which SimPal identifies a set of relevant physical variables and their relationships to create symbolic representations of the desired IGs. The symbolic representations can then be leveraged to design prompts for the original AI agent to yield better alignment with the desired IGs.

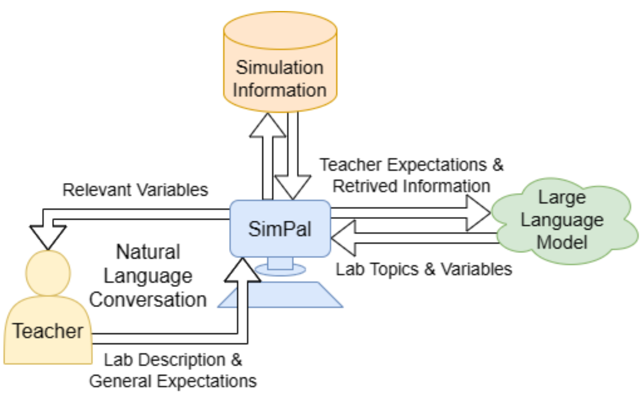

Figure 1: SimPal's high-level overview: The teacher converses with SimPal, discussing their simulation of interest and corresponding IG. As the conversation progresses, SimPal extracts useful information from the conversation to infer a computational representation of the teacher's IG. That internal representation is then communicated back to the teacher so they can make any necessary adjustments.

Figure 1 presents an overview of SimPal's interaction with the teacher. The teacher conveys their IGs to SimPal, and then SimPal creates symbolic representations of IGs by identifying relevant physical characteristics and their interactions. Accurately identifying relevant physical variables is crucial, as the IGs are encoded in terms of these variables and will guide student interactions. SimPal's architecture allows a teacher to tailor their lesson plan by I) modifying the variables and relations of a simulation through natural conversation and II) integrating any third-party simulation.

A challenging first step toward achieving this goal is to have the LLM accurately identify variables from the simulation selected by a teacher that best matches their IGs. In this paper, we empirically evaluate this task's accuracy on 63 physics simulations from PhET and Golabz using two LLMs: ChatGPT-3.5 [5] and PaLM 2 [2]. By employing the recently introduced TELeR taxonomy, we examined the impact of different prompting strategies on LLM's ability to identify the physical variables relevant to the IGs. Our findings demonstrated that SimPal can perform this task with a high degree of accuracy when provided with an appropriately crafted prompt.

2 BACKGROUND AND RELATED WORK

Conversational Agents in K-12 Science. Conversational agents, like Betty's Brain [19, 20] and MetaTutor [3, 4] have been used to foster students' learning. In Betty's Brain [19, 20], students learn science and mathematics concepts by teaching a virtual agent, Betty. *MetaTutor* is a hypermedia-based biology learning environment where teachers set learning goals and students choose metacognitive processes, with occasional pedagogical agent prompts. All of the aforementioned frameworks support students' learning, whereas SimPal offers a conversational AI assistant for teachers to develop simulation-based science lesson plans.

LLMs and K-12 Education. LLMs have recently been increasingly used to enhance student learning. Zhang et al. utilized LLMs in solving arithmetic math word problems [34]. Prihar et al. [26] utilized GPT-3 with few shot learning to generate middle school math explanations on ASSISTments. They found that GPT-3, primarily trained on English text, generated explanations that were significantly inferior to teacher-authored ones. Lately, Khan Academy has introduced a GPT-4 [1] powered tutoring system, Khanmigo [18], to assist teachers in planning their lessons and providing feedback on students writing. Our proposed approach, SimPal, is similar to Khanmigo in terms of assisting teachers in planning their lessons. However, SimPal differs from Khanmigo in that it allows teachers to integrate any *third-party simulations* into their lesson plans.

Grounding LLMs to Unseen Tasks. LLMs, which represent vast amounts of information, still require adaptation to specific tasks. Traditionally, task-specific supervised data is used to fine-tune an LLM and adapt it to new natural language processing (NLP) applications [10, 16, 17, 27]. However, fine-tuning faces two major challenges: insufficient training data and a lack of computing resources and expertise. Few-shot learning is another approach that uses prompt engineering [6, 13] and domain-specific examples [5]. However, few-shot learning may be challenging for lesson planning due to teachers' individual teaching styles and preferences. Reinforcement learning (RL) from human feedback (RLHF) employs RL to optimize human preferences during LLM training [24]. However, it can incur significant exploration costs in RL. In contrast, our approach, known as *meta-conversation*, uses natural conversation to infer a human preference, i.e., the teacher's lesson plan.

Prompt Taxonomy for LLM. As LLM's prompt impacts the output accuracy of LLMs, a recent study proposed a taxonomy, TELeR [29], to design and evaluate prompting techniques systematically. TELeR taxonomy has seven levels of prompts. We only explain the four prompt levels [Level 1- Level 4] used in our study in Table 1.

Table 1: TELeR Taxonomy for LLM Prompting

Level (L)	Definition
L1	One sentence describing the high-level task goal
L2	Multi-sentence prompt describing the high-level goals and sub-tasks
L3	Prompt describing the high-level goals and sub-tasks in bulleted style.
L4	Prompt specifying high-level goals, sub-tasks, and output evaluation criteria (e.g., few-shot examples)

3 INSTRUCTION GOALS AND SIMPAL

We formulate a teacher's IG in terms of variables and relationships among variables. Consider a toy example where the teacher's instructional goal is to teach inversely proportional relationships in Newton's Second Law of Motion in a PhET simulation [22]. As demonstrated in Figure 1, the teacher conveys their IGs (e.g., inversely proportional relationships Newton's Second Law of Motion) to SimPal. Then, SimPal generates relevant topics (e.g., force, acceleration) for the lab and asks the teacher to review those. Upon receiving the teacher's feedback, SimPal then identifies a set of relevant variables and their relationships to create symbolic representations of the desired IGs based on the teacher's feedback.

The scope of our study is variable extraction in Physics simulations, with the task described as follows.

Problem Definition. Given an IG of a simulation topic, SimPal uses LLMs to generate *variables*. The task is to assess LLM's accuracy of generated variables given a natural language description of the IG.

4 EXPERIMENTAL DESIGN

4.1 Underlying LLM of SimPal

Table 2 lists three LLMs that we assessed in our preliminary analysis.

Table 2: LLMs Evaluated in this work.

Model	Creator	# Parameters
ChatGPT-3.5 (`gpt-3.5-turbo-0613`, [5])	OpenAI	175B
PaLM 2 (`chat-bison-001`, [2])	Google	340B
LLaMA-2 (`Llama-2-70b-chat-hf`, [31])	Meta	70B

4.2 Prompt Design with SimPal

We used Level 1 to Level 4 following the TELeR taxonomy in Table 1. Example Level 1, 2, 3, and 4 prompts are given below.

- **Level 1** Identify and list the variables associated with these topics and the description, along with their corresponding symbols.
- **Level 2** You are a physics teacher in a high school, and you are preparing a lesson plan on related concepts. You have a list of topics and descriptions.
  Your task is to ***Level 1 Prompt Text***
  Please provide the variables and symbols in the following JSON format. The key would be the "Name" of the variable and the value would be the "Symbol".
  Include symbols and strictly follow the JSON format.
  Do not print topics and descriptions; only variable names and corresponding symbols are used.
- **Level 3** ***Level 2 Prompt Text***
  Please provide the variables and symbols in the following JSON format: ["Name": " ", "Symbol": " "]
  - List down all the relevant variables and their symbols.
- **Level 4** ***Level 3 Prompt Text***
  You are given a GUIDELINES_PROMPT to show an example but do not include the variables from the GUIDELINES_PROMPT in the response if they are not relevant.

4.3 Simulation Dataset

Our dataset includes simulations from PhET [33] and Golabz [32]. PhET hosts free math and science simulations. Golabz hosts online science labs to promote inquiry learning at scale. We performed preliminary analysis on five PhET simulations (Section 4.4) and final evaluation on 32 PhET and 31 Golabz simulations (Section 5).

4.4 Preliminary Experiments and Insights

We investigated the output of three LLMs on five PhET simulations using the TELeR taxonomy prompting levels [Level 1– Level 4]. Table 3 shows that all three LLMs' F1-scores fall with Level-4 prompting. Observing the format accuracy of Levels 2 and 3, we conclude that ChatGPT-3.5 and PaLM 2 generate output in the desired format. Based on the results in Table 3, we selected two LLMs, ChatGPT-3.5 and PaLM 2, with Level 2 and Level 3 prompting levels.

Table 3: LLM Performance and Prompting Levels as per the TeLER Taxonomy. Format Accuracy = (0) 1, if LLM-generated Results (Do not) Follow the Prompt's Format Specification. The Highest of each Metric per Prompt Level is in Bold

Model	Format Accuracy	Precision	Recall	F1 Score
Level 1				
ChatGPT-3.5	0	0.923	0.923	0.923
PaLM 2	0	0.923	0.958	0.94
LLaMA-2 (70B)	0	**0.929**	**1**	**0.963**
Level 2				
ChatGPT-3.5	1	0.78	0.729	0.754
PaLM 2	1	**0.881**	0.835	0.857
LLaMA-2 (70B)	0	0.876	**0.897**	**0.887**
Level 3				
ChatGPT-3.5	1	**0.898**	**0.877**	**0.887**
PaLM 2	1	0.853	0.848	0.851
LLaMA-2 (70B)	0.4	0.755	0.767	0.761
Level 4				
ChatGPT-3.5	1	0.732	0.691	0.711
PaLM 2	1	**0.96**	0.712	**0.818**
LLaMA-2 (70B)	0	0.82	**0.761**	0.7894

5 FINAL CASE STUDY AND EVALUATION

Dataset. We evaluated SimPal's performance in 63 Physics simulations, including 32 from PhET and 31 from Golabz, as depicted in Table 4. For each simulation, we designed two prompting levels (Level 2 and Level 3) using two LLMs: ChatGPT-3.5 and PaLM 2.

Table 4: Dataset Statistics. L2 = Level 2, L3 = Level 3, #Prompts = Total Prompts by Level 2 and Level 3

	ChatGPT-3.5			PaLM 2		
	L2	L3	#Prompts	L2	L3	#Prompts
Golabz	32	32	64	32	32	64
PhET	31	31	62	31	31	62

Evaluation. We created prompts by extracting IGs and topics from lab web pages. The IGs in PhET and Golabz are the learning goals and lab descriptions, respectively. To identify gold standard variables for a lab, we identified topics from the lab webpage and added additional terms from the Teacher Resources section. Finally, we cross-referenced the relevant terms with an open-source CK-12 Physical Science textbook [8], aligned to the Next Generation Science Standards (NGSS) [9] to determine the final gold standards and manually compared SimPal's outputs to the gold standards.
Metric. For each simulation, the LLM inferred variables are compared against the list of gold standard variables to compute the true positive, false positive, true negative, and false negative statistics. Then, all such statistics in a dataset were aggregated to compute the final Precision, Recall, and micro-averaged F1 score.

Table 5: An Example Annotation Scheme and SimPal's Output Evaluation on a Lab Titled *Wave on a String*

Topics	LLM Output	Gold Standard
Frequency Amplitude Damping	""Name"": ""Wavelength""," , ""Symbol"": ""ǒ3BB"" ""Name"": ""Frequency"", ""Symbol"": ""f"" ""Name"": ""Period"" , ""Symbol"": ""T"" ""Name"": ""Amplitude"" , ""Symbol"": ""A"" ""Name"": ""Speed"" , ""Symbol"": ""v"" ""Name"": ""Damping Coefficient"", ""Symbol"": ""ǒ3B2""	frequency amplitude wavelength period

Table 5 presents an example of SimPal's output evaluation in a lab. We calculated true positive values (TP) by comparing the number of matched LLM outputs to the gold standard, resulting in four true positives. We calculated false positives (FP) by subtracting the number of LLM outputs from the true positives, yielding two false positives. Further, we calculated the false negatives (FN) by subtracting true positives from the number of gold standard outputs, resulting in zero false negatives in the given example.

5.1 Results and Discussion

Table 6 presents our evaluation results of SimPal.

TELeR Prompting Levels and SimPal Performance. Level 3 prompting resulted in higher F1 scores for both LLMs than Level 2 in Golabz simulations. In PhET simulations, Level 2 prompting produced a higher recall score than Level 3 in PaLM 2.

Table 6: SimPal's Performance with TeLER Prompt Levels 2 and 3 for LLM Families and Simulation Sources in Table 4

ChatGPT-3.5						
	Precision	Recall	F1	Precision	Recall	F1
	Level 3			Level 2		
Golabz	0.590	0.713	0.60	0.525	0.627	0.541
PhET	0.560	0.654	0.581	0.523	0.519	0.539
PaLM 2						
	Level 3			Level 2		
Golabz	0.607	0.639	0.568	0.555	0.591	0.525
PhET	0.512	0.584	0.547	0.529	0.628	0.547

LLM Family and SimPal Performance. ChatGPT-3.5 outperformed PaLM 2 in F1-scores in both Golabz and PhET simulations with Level 3 prompting. ChatGPT-3.5 also achieved a higher F1 score than PaLM 2 for Level 2 prompting in Golabz simulations.
Simulation Source and SimPal Performance. Golabz simulations resulted in a higher F1-score in both Level 2 and Level 3 prompting than PhET in ChatGPT-3.5. In PaLM 2, Golabz simulations outperformed PhET in F1-score in only Level 3 prompting.

The differences in F1 scores between Golabz and PhET simulations may be due to content alignment differences. Golabz simulations may have been more aligned with curriculum standards. Additionally, PhET simulations may contain more complex or detailed information, resulting in the generation of extraneous outputs.

6 FUTURE WORK

We plan to extend SimPal to provide support to students via meta-conversation. This includes feedback on writings, answered questions, and hint generation. Additionally, we plan to use SimPal's student interaction data to generate recommendations for teachers, such as identifying high-performing and struggling students.

7 CONCLUSION

In this study, we present SimPal, an LLM-based meta-conversational framework for simulation-based science labs, allowing teachers to include third-party (open or closed-source) simulations into lesson plans, facilitating *instruction at scale*. We assessed SimPal's variable generation capabilities with two LLMs: ChatGPT-3.5 and PaLM 2 on 63 Physics simulations from PhET and Golabz, experimenting with different prompts following the TELeR prompting taxonomy. Our findings showed that I) SimPal can provide a meaningful variable list tailored to the lab and instruction goal, and II) the LLM prompting level impacts SimPal's performance. Furthermore, we observed that Golabz simulations outperformed PhET in the F1 score. It is important to note a limitation in our evaluation; our gold standard outputs may lack the subject matter expertise of real school teachers, potentially leading to disparities in F1 scores. Future work will involve incorporating feedback from teachers and subject matter experts to improve the accuracy and relevance of LLM outputs.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation through Grant 2302974.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[3] Roger Azevedo, Amy M Witherspoon, Arthur C Graesser, Danielle S McNamara, Amber Chauncey, and Emily Siler. 2009. MetaTutor: Analyzing Self-Regulated Learning in a Tutoring System for Biology. AIED.

[4] François Bouchet, Roger Azevedo, John S Kinnebrew, and Gautam Biswas. 2012. Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. *International Educational Data Mining Society* (2012).

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. *Advances in Neural Information Processing Systems* 35 (2022), 23908–23922.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[8] CK-12. 2024. CK-12 Physical Science for Middle School. https://flexbooks.ck12.org/cbook/ck-12-middle-school-physical-science-flexbook-2.0/.

[9] National Research Council et al. 2013. Next generation science standards: For states, by states. (2013).

[10] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *Advances in neural information processing systems* 28 (2015).

[11] Garry Falloon. 2019. Using simulations to teach young students science concepts: An Experiential Learning theoretical analysis. *Computers & Education* 135 (2019), 138–159.

[12] Christian Fischer and R Charles Dershimer. 2020. Preparing teachers to use educational games, virtual experiments, and interactive science simulations for engaging students in the practices of science. (2020).

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. 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)*. 3816–3830.

[14] Javier González-Cruz, Rogelio Rodríguez-Sotres, and Mireya Rodríguez-Penagos. 2003. On the convenience of using a computer simulation to teach enzyme kinetics to undergraduate students with biological chemistry-related curricula. *Biochemistry and Molecular Biology Education* 31, 2 (2003), 93–101.

[15] Arthur C Graesser, G Tanner Jackson, Hyun-Jeong Joyce Kim, and Andrew Olney. 2006. AutoTutor 3-D Simulations: Analyzing Users' Actions and Learning Trends. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 28.

[16] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 328–339.

[17] Nathan Hu, Eric Mitchell, Christopher D Manning, and Chelsea Finn. 2023. Meta-Learning Online Adaptation of Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 4418–4432.

[18] Sal Khan. 2024. Khanmigo. https://www.khanacademy.org/khan-labs.

[19] John S Kinnebrew and Gautam Biswas. 2012. Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. *International Educational Data Mining Society* (2012).

[20] John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. 2013. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM| Journal of Educational Data Mining* 5, 1 (2013), 190–219.

[21] Bas Kollöffel and Ton De Jong. 2013. Conceptual Understanding of electrical circuits in secondary vocational engineering education: Combining traditional instruction with inquiry learning in a virtual lab. *Journal of engineering education* 102, 3 (2013), 375–393.

[22] Leah L. Lecerio. 2019. Newton's Second Law of Motion. https://phet.colorado.edu/en/contributions/view/5092.

[23] Alexander Hanbo Li, Patrick Ng, Peng Xu, Henghui Zhu, Zhiguo Wang, and Bing Xiang. 2021. Dual Reader-Parser on Hybrid Textual and Tabular Evidence for Open Domain Question Answering. 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)*. 4078–4088.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[25] Joonhyeong Park, Tang Wee Teo, Arnold Teo, Jina Chang, Jun Song Huang, and Sengmeng Koo. 2023. Integrating artificial intelligence into science lessons: teachers' experiences and views. *International Journal of STEM Education* 10, 1 (2023), 61.

[26] Ethan Prihar, Morgan Lee, Mia Hopman, Adam Tauman Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil Heffernan. 2023. Comparing different approaches to generating mathematics explanations using large language models. In *International Conference on Artificial Intelligence in Education*. Springer, 290–295.

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

[28] Nico Rutten, Wouter R Van Joolingen, and Jan T Van Der Veen. 2012. The learning effects of computer simulations in science education. *Computers & education* 58, 1 (2012), 136–153.

[29] Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. In *Findings of the Association for Computational Linguistics: EMNLP*. 14197—-14203.

[30] Anirudh S Sundar and Larry Heck. 2023. cTBLS: Augmenting Large Language Models with Conversational Tables. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. 59–70.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[32] the Netherlands University of Twente. 2012. Global Online Science Labs for Inquiry Learning in Schools. https://premium.golabz.eu/about/go-lab-initiative.

[33] Carl Wieman. 2002. PhET. https://phet.colorado.edu/.

[34] Mengxue Zhang, Zichao Wang, Zhichao Yang, Weiqi Feng, and Andrew Lan. 2023. Interpretable Math Word Problem Solution Generation Via Step-by-step Planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 6858—-6877.