Contribution ID: **13** Type: **not specified**

Exploring LLM performance on Physics 101 coursework in different languages

Large language models see rapid adoption in various domains, prompting us to rethink established teaching paradigms. We examine their utility in university-level physics education, focusing on two main aspects: Firstly, how reliable are publicly accessible models in answering exam-style multiple-choice questions? Secondly, how does the question's language affect the models' performance? We benchmark a number of LLMs on the mlphys101 dataset, a new set of 929 university-level MC5 questions and answers released alongside this work. Using a GPT-4 powered response parser, we compare the other models' responses against sample solutions. While the original questions are in English, we employ GPT-4 to translate them into various other languages, followed by revision and refinement by native speakers. Consistent with related works, GPT-4 outperforms the other models across all languages and tests, including simple multi-step reasoning problems that involve calculus. Publicly available models such as GPT-3.5 and Mistral-7B produce more incorrect answers, sometimes struggle to maintain the desired output format, and show a preference for English inputs, necessitating more precise prompt engineering. In conclusion, the most advanced LLMs already perform well on basic physics courses and LLM powered translations are a viable method to increase the accessibility of materials. Further improvements may lead to PhysGPT, a teaching assistant for instructors and personalized tutor for students, redefining how we learn and teach in the age of AI-assisted education.

Primary author: VÖLSCHOW, Marcel (Hamburg University of Applied Sciences)

Presenter: VÖLSCHOW, Marcel (Hamburg University of Applied Sciences)

Track Classification: Alignment, Ethics, and Reliability