Hi All,

Further to Robert’s email about ‘physics answering efficacy’ for various AI models I’ve created a textbook question and answer bank with a simple API script that bulk-sends all the questions and registers the response (see attached). Below are two plots showing a) the correct answer rate overall and by subject and b) how the AI scores itself when you give it a correct question/answer pair.

This question and answer bank has been created from introductory physics textbooks (see excel file)

but the scores are XXXX

The API script uses both the free legacy model (davinci-002) and the February version of ChatGPT (gpt-3.5-turbo) as ChatGPT4 isn’t available on the API yet, but I imagine it will be soon. However, even with gpt-3.5-turbo it still scores XXXX

The view here is to build out this bank of questions from different difficulty levels to get a more standardized idea of the overall efficacy. Broadly questions are either going to be numerical, derivation based or word/written based. To get an idea of how it performs with different subjects I stuck to numerically answerable questions.

There is also the idea of getting it to answer and coursework/exam in one go to see how vulnerable it currently is to AI. On this, I tried a few different optical character recognition modules to see if you could just ‘scan in’ an exam paper but it kept tripping up on mathematical symbols so for typing out the questions is required.

Cheers,

Will