1 Introduction

Large language models (LLMs) have demonstrated the ability to pass exams from individual college-level courses [17]. However, a systematic evaluation of this ability across entire curricula that students would be expected to go through to obtain their college degrees has yet to be explored due to the lack of a central repository of questions from actual exams and assignments at the curriculum level. In this paper, we introduce MITCOURSES, a curated dataset of 4,550 questions and their solutions spanning exams and assignments from all courses that form the curriculum for MIT's Mathematics and Electrical Engineering and Computer Science (EECS) majors. The dataset is seven times larger than our previously released dataset [7, 8] and covers all the course requirements for an undergraduate degree.

Using this dataset, we benchmark four state-of-the-art language models, GPT-4 [17], GPT-3.5, StableVicuna [22], and LLaMA [23], which vary in their sizes, their capabilities, and whether their weights are publicly available or not. We evaluate different LLM prompting techniques (few-shot [5], chain of thought [12], self-critique [11, 14, 19, 21, 9]) and document their effect on the model success rates. We propose a new prompting technique which we call *expert prompting*, where we ask the model to suggest named experts on a given question, then ask for the answer the named experts would have given and subsequently make a collective decision. Our experiments show expert prompting improves performance relative to prior prompt engineering techniques.

For models whose weights are publicly available, we also demonstrate that fine-tuning improves performance on our test set and other reasoning task benchmarks.

In addition to providing performance measures across various models, we demonstrate the application of our models for curriculum design for college majors. An essential aspect of curriculum design is determining the appropriate sequence of courses to ensure prerequisites are established effectively. Traditionally, this process is manual, relying on human input to identify key concepts and learning outcomes. However, this method is subjective, and coordinating input from various faculty members teaching different courses can be challenging. Instead, we use the embeddings for the questions of different courses to discover dependencies between courses. Given the challenges of accurate student evaluation in a world where large language models are readily available, we also propose the development of new meta-questions that focus on assessing the correctness and completeness of students' understanding rather than their ability to generate correct answers.

To prevent our dataset being incorporated as part of LLM training corpora, the dataset will not be made publicly available but will be made available to researchers upon request through a data use agreement (DUA). This is both to preserve the value of this dataset as a resource for benchmarking LLMs—which is lost if the dataset becomes part of the training set for future public models—as well as to respect the wishes of instructors who contributed to it and who want to preserve their ability to use problems from this dataset in the future. We release our open-source LLaMA models fine-tuned on our dataset, called MIT-LLM, in the supplemental material.

Related Work. Large language models such as GPT-4 demonstrate excellent results on standardized AP tests [17, 6]. Their performance across many reasoning tasks has been enhanced by few-shot learning and by providing intermediate reasoning steps and chain-of-thought (CoT) prompts [16, 25, 24, 26]. Iterative prompting techniques such as self-critique, self-refinement, self-feedback, self-reflection, and self-improvement [11, 14, 19, 21, 9] have further been shown to improve performance. While much of the early work on eliciting better reasoning capabilities of LLMs has focused on prompt engineering, their capabilities have recently expanded by combining LLMs with reinforcement learning. In this work, we write an optimization objective that accurately describes the problem, formulates these improvements, and provides an ablation study. LLMs are now routinely used for mathematical reasoning, solving university-level mathematics and computer science problem sets and final exams [7, 13, 20, 8, 2] at the human level. Our work extends this to entire undergraduate degrees.