Generating and Grading Exam Questions with the Help of Large Language Models

Clemens H. Cap

September 2024

1 Introduction

Students often ask for sample exam questions to evaluate their state of knowledge on certain instructional material. With the advent of large language models (LLMs) such as ChatGPT there is a wide spread expectation that the task of developing exam questions as well as providing feedback might partially be accomplished by artificial intelligence systems.

While the systems themselves seem to be pretty knowledgeable in numerous areas themselves, they should, however, not produce contents according to their own understanding but follow the curriculum as provided by material from the instructor.

There are already several examples for prompts which guide text-generating AI systems in developing multiple choice questions and thus there are big expectations that these systems can be used in educational settings. There is, however, insufficient neutral, systematic or critical evaluation for the true abilities of such approaches.

The task of this Thesis is an empirical study and evaluation based on texts in the domain of undergraduate computer science study material.

2 Research Questions

1. Collect a test corpus of instructional texts on which LLMs shall generate exam questions of different types (yes/no, single choice, multiple choice, free text and more).

2. Conduct tests with different text-generating AI systems.

3. Let these AI systems grade correct and incorrect answers of students and evaluate systematically the correctness of the grading as well as the feedback given to incorrect answers.

4. Develop an evaluation scheme whether the respective LLMs follow the contents provided in the instructional texts.

5. Two particularly interesting border cases which should be studied are the following:

   What happens when the contents of the instructional texts and the knowledge available to the LLM are in contradiction? Does the LLM rather follow the instructional text or its own belief system?

   To which extent is the LLM capable to correct the user and provide meaningful feedback to wrong answers? This question reflects on the character trait of some LLMs to rather not frustrate the user but follow his suggestions even in cases where the user objectively is wrong.