Automatic Generation and Evaluation of Reading Comprehension. Test Items with Large Language Models

Andreas Säuberli, Simon Clematide

Reading comprehension tests are used in a variety of applications, reaching from education to assessing the comprehensibility of simplified texts. However, creating such tests manually and ensuring their quality is difficult and time-consuming. In this paper, we explore how large language models (LLMs) can be used to generate and evaluate multiple-choice reading comprehension items. To this end, we compiled a dataset of German reading comprehension items and developed a new protocol for human and automatic evaluation, including a metric we call text informativity, which is based on guessability and answerability. We then used this protocol and the dataset to evaluate the quality of items generated by Llama 2 and GPT-4. Our results suggest that both models are capable of generating items of acceptable quality in a zero-shot setting, butGPT-4 clearly outperforms Llama 2. We also show that LLMs can be used for automatic evaluation by eliciting item responses from them. In this scenario, evaluation results with GPT-4 were the most similar to human annotators. Overall, zero-shot generation with LLMs is a promising approach for generating and evaluating reading comprehension test items, in particular for languages without large amounts of available data.