Contribution ID: **107** Type: **not specified**

Assessing inter-rater reliability of LLMs with the probability of agreement

Inter-rater reliability, the quantification of agreement between individuals who assign scores to the same phenomenon, is an important consideration in all fields for which data drives decision-making (e.g., business and industry, healthcare, social and behavioural sciences, education, etc.). Traditionally, the raters scoring the phenomenon have been human beings. With the proliferation of AI, a natural question arises: can a large language model (LLM) perform this task as well as humans? Central to this question is the assessment of inter-rater reliability of LLMs relative to humans. In this talk, we describe one such problem in the ed-tech space, where the goal is to establish the reliability of LLM-evaluation of educational material. In particular, we describe an end-to-end framework implemented at a prominent ed-tech company in which agreement studies are designed and analyzed to compare LLM raters with human raters via the probability of agreement.

Special/ Invited session

ISEA session

Classification

Both methodology and application

Keywords

agreement, reliability, concordance

Primary author: STEVENS, Nathaniel (University of Waterloo)

Presenter: STEVENS, Nathaniel (University of Waterloo)

Track Classification: Other/special session/invited session