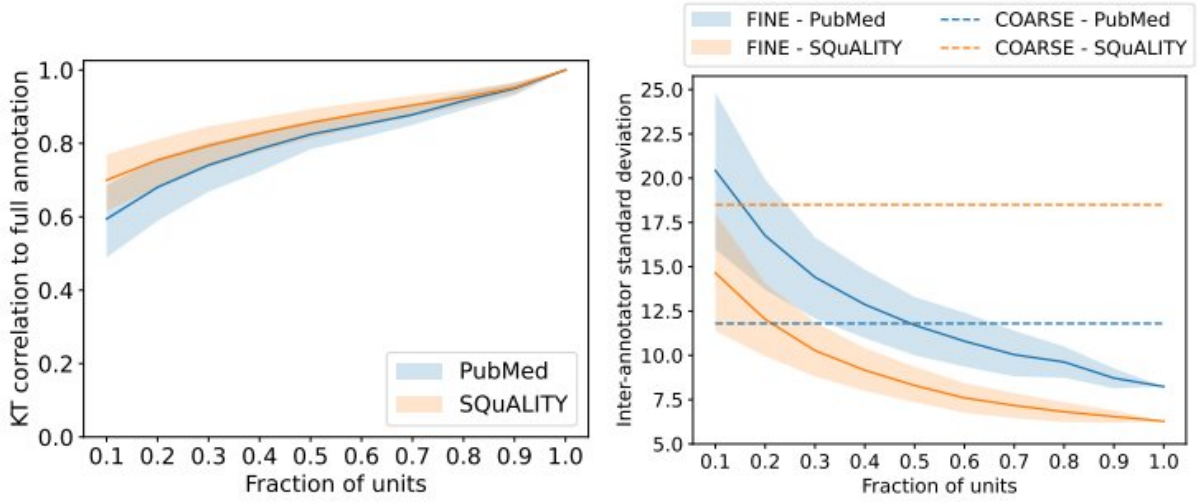
Computer scientists release guidelines for evaluating AI-generated text

July 7 2023



Accuracy and variance after annotating a fraction of units per summary (X-axis) with FINE. Despite annotating just a fraction of the summary, we observe a high segment-level Kendall tau correlation with a full annotation (left). However we observe higher inter-annotator variance as the fraction reduces (right). Confidence intervals shown are 95% and computed across 1000 random subsets (see Appendix F for left plot with Pearson). Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2301.13298

The public release of AI text generators, such as ChatGPT, has caused an enormous stir among both those who herald the technology as a great leap forward in communication as well as those who prophesy the

technology's dire effects. However, AI-generated text is notoriously buggy, and human evaluation remains the gold-standard in ensuring accuracy, especially when it comes to applications such as generating long-form summaries of complex texts. And yet, there are no accepted standards for human evaluation of long-form summaries, which means that even the gold-standard is suspect.

To rectify this shortcoming, a team of computer scientists, led by Kalpesh Krishna, a graduate student in the Manning College of Information and Computer Sciences at UMass Amherst, has just released a set of guidelines called LongEval. The guidelines were presented at the European Chapter of the Association for Computational Linguistics, for which it was awarded the Outstanding Paper prize.

"There is currently no reliable way to evaluate long-form generated text without humans, and even current human evaluation protocols are expensive, time-consuming and highly variant," says Krishna, who began this research during an internship at the Allen Institute for AI. "A suitable human evaluation framework is critical to build more accurate long-form text-generation algorithms."

Krishna and his team, including Mohit Iyyer, assistant professor of computer science at UMass Amherst, combed through 162 papers on long-form summarization to understand how human evaluation works—and in doing so, they discovered that 73% of the papers did not perform human evaluation on long-form summaries at all. The remaining papers used widely divergent evaluation practices.

"This lack of standards is problematic because it hampers reproducibility and does not allow for meaningful comparison between different systems," Iyyer says.

To further the goal of efficient, reproducible and standardized protocols

for human evaluation of AI-generated summaries, Krishna and his co-authors developed a list of three comprehensive recommendations that cover how and what an evaluator should read in order to judge the reliability of the summary.

"With LongEval, I am very excited about the prospect of being able to accurately and quickly evaluate long-form text generation algorithms with humans," says Krishna. "We have made LongEval very easy to use and released it as a Python library. I am excited to see how the research community builds upon it and uses LongEval in their research."

The research is published on the *arXiv* preprint server.

More information: Kalpesh Krishna et al, LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization, *arXiv* (2023). DOI: 10.48550/arxiv.2301.13298

Provided by University of Massachusetts Amherst