For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

SEMINAR ON DATA SCIENCE

Can large language models provide useful feedback on research papers? A large-scale empirical analysis

By

Mr. Weixin LIANG

Stanford University

Abstract

Expert feedback lays the foundation of rigorous research. However, the rapid growth of scholarly production and intricate knowledge specialization challenge the conventional scientific feedback mechanisms. High-quality peer reviews are increasingly difficult to obtain. Researchers who are more junior or from under-resourced settings have especially hard times getting timely feedback. With the breakthrough of large language models (LLM) such as GPT-4, there is growing interest in using LLMs to generate scientific feedback on research manuscripts. However, the utility of LLM-generated feedback has not been systematically studied. To address this gap, we created an automated pipeline using GPT-4 to provide comments on the full PDFs of scientific papers. We evaluated the quality of GPT-4's feedback through two large-scale studies. We first quantitatively compared GPT-4's generated feedback with human peer reviewer feedback in 15 Nature family journals (3,096 papers in total) and the ICLR machine learning conference (1,709 papers). The overlap in the points raised by GPT-4 and by human reviewers (average overlap 30.85% for Nature journals, 39.23% for ICLR) is comparable to the overlap between two human reviewers (average overlap 28.58% for Nature journals, 35.25% for ICLR). The overlap between GPT-4 and human reviewers is larger for the weaker papers (i.e., rejected ICLR papers; average overlap 43.80%). We then conducted a prospective user study with 308 researchers from 110 US institutions in the field of AI and computational biology to understand how researchers perceive feedback generated by our GPT-4 system on their own papers. Overall, more than half (57.4%) of the users found GPT-4 generated feedback helpful/very helpful and 82.4% found it more beneficial than feedback from at least some human reviewers. While our findings show that LLM-generated feedback can help researchers, we also identify several limitations. For example, GPT-4 tends to focus on certain aspects of scientific feedback (e.g., `add experiments on more datasets'), and often struggles to provide in-depth critique of method design. Together our results suggest that LLM and human feedback can complement each other. While human expert review is and should continue to be the foundation of rigorous scientific process, LLM feedback could benefit researchers, especially when timely expert feedback is not available and in earlier stages of manuscript preparation before peer-review.

Biography

Weixin Liang is in the third year of his Doctorate studies in Computer Science at Stanford University, working under the supervision of Professor James Zou. Previously, he obtained a Master's degree in Electrical Engineering from Stanford University and a Bachelor's degree in Computer Science from Zhejiang University. His research is primarily focused on the areas of trustworthy AI, data-centric AI, and natural language processing.

Date : 9 November 2023 (Thursday)

Time : 11:00am

Zoom Meeting : https://hkust.zoom.us/j/5616960008 (Passcode: hkust)

All are Welcome!