Can LLMs read between the lines? Exploring how they compare to human coders in categorising short pieces of ambiguous text

Nine Adler

University College London, London, United Kingdom

Stephen H. Dewitt

University College London, London, United Kingdom

Laura Elaine Strittmatter

ETH Zürich, Zürich, Switzerland

Abstract

While the democratisation of LLMs has proven fruitful in the field of Experimental Psychology, their adoption in some use cases has been slower. Here we explore how LLMs perform on a content analysis task following a predefined coding scheme. Using a qualitative dataset (346 responses, 6 labels with binary options) with an established inter-rater reliability over 95% as our benchmark, we compare models' outputs against human coders. The text responses are often ambiguous and require a certain level of inferential reasoning and subjective interpretation which LLMs still struggle with. We gave GPT 4-o, Qwen 2.5 and Mistral 7B the dataset and label definitions. GPT 4-o matched 59% of human labelling across responses, Mistral 7B matched 46% and Qwen 2.5 matched 42%. We discuss the 'reasoning' element of LLMs and potential ways forward.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*