Controlling for Bias in Assessment Assistant

Dr. Paul Embleton
February 2025

Problems

One major concern regarding the use of AI in healthcare settings is the threats of bias and hallucination. Both of these concepts exist in the context of computer information systems.

Bias refers to the idea that some aspect of the information system was designed in such a way that introduced some type of undue influence. For example, large language models are complex systems that are 'trained' on a corpus of material. Many questions come to mind. What is the source of that corpus? What implicit bias already existed in that corpus? Who authored, vetted, and approved that corpus? Is it right to use that source material in my professional work? And so on.

Hallucination refers to the idea that generative AI (usually language models) are prone to 'making up' information that is false or misleading. This is common in commercial language products; the product is designed to always give an answer without significant regard to the level of confidence of the answer. In the professional psychological sciences, the threshold for confidence values is very strict. Most providers use a confidence threshold of at least 80% when interpreting standardized test data or concluding research results. Some providers hold a 95% confidence threshold standard.

Solutions

So, how have we controlled for these problems in the engineering of our product?

For bias, the solution is two parts: 1) grounded language models with expert vetted source material and 2) explicit declaration of potential bias to end users through materials such as this article. To start, we carefully selected a number of extremely high quality source documents to be made available to the language model for generative processing. We then restricted the language model to only use information provided by these source materials (also known as 'grounding') in its analysis and response.

For hallucinations, we combine the bias controls with both a confidence value threshold and a source citation system. For all generative AI text provided in our application, we cite the source from which it came. We also let the provider choose what confidence value they would like the language model to respect. So if a model's overall response does not return a high enough confidence value, it won't be returned at all.