AI in evaluation actually show your working

10 Apr 2023

First published on the Causal Map blog on 10 April 2023, a few months after ChatGPT was released, and republished here as it was written. Read it as an early statement of the position; what it says about the Causal Map app describes the app as it was then.



There's been a lot of talk about using AI and in particular large language models in evaluation and specifically in coding and processing texts.

Here at Causal Map we've been working very hard on just that (and on automating interviewing too, but that's another story). And we see fantastic potential. Our Causal Map app now has a beta version of that big "auto code" button we'd always dreamed of (and feared).

However, I wanted to draw attention to a really big distinction which I think is important. There's a continuous spectrum between on at the one end **transparent**, reproducible approaches founded in social science and the other end of the spectrum **black box** approaches where responsibility is shifted from the evaluator to the AI. There may be use cases for the latter, "black box" kind of approach. Maybe one day doctors will abrogate all responsibility to medical AI. Maybe one day evaluators will abrogate all responsibility to evaluation AI. But here I'd like to set out reasons why right now we should prefer transparency.

Black box coding is possible today in its rudiments and it's going to get a lot more accessible and powerful quite quickly.

At its most extreme, you simply say to the AI 'Here's a load of documentation from a project. You tell me if the project is efficient, effective, sustainable, draw some conclusions and make recommendations according to criteria C, D and E. This is an extreme case, but the basic idea is submitting a long text and asking for a black box judgement about what themes are present and even what conclusions can be drawn.

To be sure, it's possible to say to a model 'Yes and also show your working or print out some quotes or examples to backup your findings.' But it's very important to realise that this "show your working" question is **spurious** because AI at the current state of development has no more insight into its inner workings than does a human being has into his or hers, and probably less so. So while it can (and will) competently bullshit about what steps somebody might have taken to reach that conclusion it doesn't mean it's actually the steps that it did take. So basically, you have no way of knowing how the AI came up with a particular finding or conclusion using this approach and it's a massive abrogation of responsibility for an evaluator to sign off this kind of output without further analysis.

Now at the other, "**transparent**" end of the spectrum, what we recommend is **using AI merely to follow established procedures of manual coding and do it faster, more reliably and more reproducibly**. That's a big win. The old school way: First of all, highlighting individual sections of text according to explicit rules set by the evaluator and then aggregating and combining those codings, again according to explicit rules.



As an aside, we believe that even before we get into the AI possibilities, **causal mapping in particular** is a really good way to summarise documents and in particular sets of documents. Obviously, there is more to documents than simply the causal claims made within them, but if you had to pick a type of content an evaluator might want to extract from a document, causal claims are pretty central and the procedure for identifying, extracting and aggregating those claims are an order of magnitude more straightforward than any other kind of useful text analysis (unless you count word clouds...). In particular, causal mapping is particularly good at making summaries from *sets* of documents, such as semi structured interviews with comparable respondents, rather than only the special case of making one summary of just one document.

It is already possible to say to an AI, 'please read this long document and draw a causal map saying what do you think are the main causal drivers and outcomes and intermediate links and just print out the specification of a diagram'. And the job's done. That's exactly the sort of approach we are **warning against** because you have no way of knowing how the model has reached that conclusion.

When we use AI to help code a set of documents we tell it to explicitly identify causal claims *and provide the relevant quote for each individual claim*, following rules we give it and in each case, it's possible to

look at the actual quote it identifies and check if it really is appropriate evidence for the causal claim. Just as with human coding, in the sort of way causal mapping has been carried out for 50 years or more



It's been a lot of work to develop the right set of prompts (and they are still a work in progress) embedded in our app, but the prompts we use in any given case are pretty simple and transparent: around half a page of standard prompts which are pretty much the same across use cases and another half a page or so of prompts which are specific to the use case; these themselves are 90% derived in an automated way.

Nevertheless, the evaluator bears 100% responsibility for overseeing these prompts, which are plain English. They can be followed by a team of postgrads or by the AI: there is no difference in principle. There is no black box and no magic, and any human can follow every step of the argumentation.

At present, the AI is much faster and more reliable and transparent than a human coder; and a human coder is much better at seeing larger connections, reading between the lines and linking up the parts of a

larger story. The most interesting part of causal coding with AI is to add this human inspiration back into the AI prompt in a transparent way.

In order to then aggregate, synthesise and simplify the causal maps which result, we can use the many, more or less standard, causal mapping procedures which have been developed over the years and in particular our open source set of causal mapping algorithms. So an interested outsider can follow the chain of argument right away from the original text to the final conclusion.

Responsibility is the issue here. If you feed data or documents into an AI and let it come up with its own conclusions, they aren't your conclusions and as an evaluator you can't sign off on them. Maybe this will change in the future as we learn to find our way around in this new world. But right now, you need to show your working.

Of course the big worry in all of this is that higher-level, black-box approaches are **much** quicker and easier to apply, putting together black-box approaches to get from documents to findings to (*evaluative*) judgements in just a few clicks, given some generic definitions of evaluation criteria. Black-box approaches could be the beginning of the end of evaluation as we know it, but they'd be really tempting for a commissioner: for a purely document-based review, who'd bother with the time and expense to commission an evaluator if you can get your report written in a few minutes? With black-box approaches, people's fears about bias are really justified.