Results

22 Aug 2025

The sub-headings within each question form our criteria for answering that question.

Question 1: can an AI interviewer gather causal information at scale?

Efficiency

As we were still experimenting with the process, it took us around 8 hours to write, test, deploy and monitor the interviews.

We spent around $40 on API fees, including both tests and real interviews. The time and cost involved were significantly less than what it would have taken for humans to create an interview guideline and interview the same number of participants.

Validity

This is a difficult question to answer fully. However, in the interview prompt, we instruct the AI to summarise the conversation at the end of the interview and ask the respondent to verify its accuracy. We can use these answers to make a rough assessment of how valid the original summaries were: if the interviewee expresses no dissatisfaction, we can assume that the interviewer successfully elicited valid information.

The final section of all 163 interviews was analysed. We classified each interview into 3 groups:

1. No summary provided;
2. The respondent explicitly expressed dissatisfaction and/or asked for changes in the summary;
3. The respondent finished the interview and did not explicitly express dissatisfaction nor ask for changes in the summary.

78.5% of the respondents (**group 3**) didn't ask for changes in the summary, implying at least no dissatisfaction with what the AI produced (128 out of 163 interviews). Only in 7 interviews (4.29%) did the interviewee ask the AI to change or correct something in the summary, and/or the respondent explicitly expressed dissatisfaction (**group 2**). Of these, three then explicitly expressed satisfaction with the revised summary offered by the AI. The other 25 interviews (15.3%) were not summarised (**group 1**), mainly due to the participants breaking off before the end of the interview.

We used a much simpler architecture to manage the interview process than Chopra and Haaland (2023), however, our interviews were much shorter than theirs (their average interview length was about 30 minutes), raising the question of whether longer interviews might need more elaborate management architecture.

Question 2: Can automated causal mapping code causal information at scale?

Efficiency

It took around 5 hours to write and test the coding instructions and validate the results.

The cost of using the API was around $20.

Recall

Recall can be defined as the extent to which the AI finds "all" the causal links (Resnik & Lin 2010).

We made a separate assessment of the number of links "really" present within each interview, a "ground truth" of 1154 links. In comparison, the automated coding identified 1024 links, or 89%. However this is before assessing which of those codings were correct: the precision of the links, as follows.

Precision

Precision can be defined as the proportion of the identified links which were accurate/correct (Resnik & Lin 2010). To define "correct" we used the following informal criteria, which were assessed for each link by the second author:

1. The cause and effect in each link correctly name phenomena which are named in the text;
2. The coding represents an actual causal claim within the text (rather than, for example, merely events listed in sequence);
3. The coding represents a factual claim rather than a wish or hypothetical statement.
4. The coding is in the correct direction (cause to effect).

We gave each causal link a 0-2 score on the four criteria of precision as detailed in the Supplementary Material. 65% of the links had a perfect score, and 72% dropped only one point (a "not sure" on only one criterion). The errors we identified seem to take place approximately at random, except that there were more errors with causal claims which human analysts themselves judged to be difficult to code.

A more systematic assessment of the coding process on a real-life dataset had similar results and is currently in press (redacted).

Question 3: can automated causal mapping help answer evaluation questions?

Can an overall causal map be generated which includes much of the information?

The map in Figure 1 is filtered to show only the top 11 factors (in terms of the number of respondents mentioning them); links mentioned by only one source are also removed, meaning many less frequently mentioned factors and links are not shown.

We introduce a measure which we call **coding coverage**: given any map based on any recoding or filtering of the original data, what percentage of the original codings are included? There are balances to be struck: a map with more factors will usually have higher coverage but will be harder to understand and less useful. More homogeneity in sample and theme usually mean higher coverage. Very granular clustering will mean lower coverage or a larger map.

The first result can be seen in Figure 1. This map contains only 11 factors but covers 42% of the raw causal claims.

Insert Figure 1 around here.

Most (113 out of 136) sources have contributed at least some citations to this summary map. The numbers on the factors and links (and the sizes of the factors and the widths of the links) represent the number of sources mentioning each. Factors with darker backgrounds have proportionately more incoming than outgoing links: they have greater “outcome-ness”.

At this coarse level of “granularity”, many of the factors are bundles of cause-effect stories, as shown by the “self-loops” such as the 10 sources that mentioned links between different environment/climate change issues.

In this map, it is mostly not possible to distinguish between constituent factors with different valence or sentiment. For example, “military strengthening” and “military weakening” are two codes which have been included under “International conflict”. Indeed they are not so far from one another in the overall space of embeddings, something which is quite hard to understand from a positivistic, Cartesian point of view but which is perhaps more familiar to those more used to thinking in terms of "themes" than in terms of “variables”.

Face validity

Does the overall causal map present a plausible picture of the most important factors and how they influence one another (in the opinion of respondents)? Yes, even in the absence of a particular research

focus, this causal map has a lot to tell us about the causal worlds of the respondents.

“It’s the economy, stupid”: economic stress is mentioned by the largest number of sources and is central to most of the narratives. Covid-19 appears as a pure driver of economic stress.

Ability to answer other evaluation questions

Regarding the differences between timepoints, there were significant differences for several of the links. For example, of the five sources that mentioned the link from Political conflict to International conflict overall, all of them were from the third time-point, which is unsurprising considering the situation in Israel/Palestine at that timepoint.

In this analogue study, we did not have any additional information e.g. about the sociodemographic characteristics of the respondents which would have enabled us to look at differences between subgroups.

In a more realistic evaluation context, it would be possible to further investigate narratives about the causes and effects of specific factors of interest.

References

Resnik, & Lin (2010). *Evaluation of NLP Systems*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444324044#page=291.