🌻 Can AI accurately map causal claims - A validation study

3 Mar 2026

_Summarised from "AI-assisted causal mapping: a validation study" by Steve Powell and Gabriele Caldas Cabral.

Extracting causal claims from qualitative interviews is a notoriously slow process. Evaluators spend hours combing through transcripts to work out what respondents think influences what. Generative AI offers a solution, but a critical question remains. Is an untrained AI assistant actually any good at identifying and labelling these causal links compared to human experts?

We conducted a validation study to find out. The results show that treating AI as a tireless, low-level coding assistant works surprisingly well.

The approach: naive causal coding

We used a pragmatic, minimalist approach to causation. We did not ask the AI to model complex systems, determine the strength of causal effects, or make predictions. We simply asked it to identify evidence of causal influence within the text.

The task was reduced to a clear set of instructions: read the text, find the causal claims, and tell us what influences what. By asking the AI only to extract claims rather than interpret their ultimate validity, we keep the human evaluator firmly in charge of the actual analysis.

The experiment

We used a dataset from a Qualitative Impact Protocol (QuIP) study evaluating an agriculture and nutrition programme. Expert human analysts had already coded these transcripts by hand. This provided our benchmark.

We took a subset of the data (163 statements from three sources) and instructed GPT-4.0 to code it. The AI processed the text statement by statement, extracting the cause, the effect, and the verbatim quote proving the claim. We ran the test in two ways:

1. Radical zero-shot: The AI was given research context but no codebook or suggested labels. It had to invent its own factor labels.
2. Closed coding: The AI was provided with a basic codebook of 29 top-level labels derived from the prior human coding.**

The Results: high precision and high recall

The AI performed at a level comparable to a human assistant. We evaluated the results on both precision (were the links correct?) and recall (did it find all the links?).

- Precision: Without a codebook, 84% of the causal links identified by the AI were rated as perfectly correct across four strict criteria. When given a basic codebook, this precision rate rose to 87%. The errors the AI did make were usually minor labelling discrepancies or 'noise' rather than complete misunderstandings of the text.
- Recall: The AI successfully identified approximately as many valid causal links as the human experts. In some transcripts, it actually found valid links that the human coders had missed.

The bigger picture: comparing the maps

In qualitative data analysis, two human coders rarely produce the exact same set of detailed labels. The AI was no different. The raw, detailed maps produced by the AI and the humans differed in specific terminology.

However, evaluation rarely relies on looking at every single microscopic link. When we zoomed out to create high-level overview maps showing the most frequently mentioned factors and links, the structures were very similar to each other.

[Insert Figure 1: AI-coded map] Caption: Overview causal map generated from AI coding.

[Insert Figure 2: Human-coded map] Caption: Overview causal map generated from human expert coding.

The AI map and the human map told the same overarching story. The AI tended to prefer labels like "received training", whereas the human coders preferred "increased knowledge", but both were perfectly reasonable interpretations of the underlying text.

The verdict

Generative AI is highly capable of extracting causal claims from qualitative data. Using a naive approach to causal coding, AI achieved a precision rate well over 80% GPT-4.0, which is now an

outdated model.

This does not turn the AI into a black box that replaces the evaluator. Instead, the AI functions as a transparent assistant. It does the heavy lifting of initial extraction, reading through pages of text to pull out claims and attach the exact quotes that prove them. This allows evaluators to process qualitative data rapidly and at scale, freeing up time for the human judgment and synthesis that actually matters.