



JUST ADD RIGOUR THREE DO'S AND DON'TS

9 Apr 2025

Don't do that !



Do this !



Three do's and don'ts when using AI for text analysis.

A lot of evaluation work is a kind of text analysis: processing reports, interview transcripts, etc. A bit like qualitative social science research. So this little piece is for evaluators in particular and (qualitative) social scientists in general.

How do we get from texts to evaluative judgements?

Recently many evaluators and researchers have been turning to AI to help.

BUT if you didn't have a clear workflow from data to judgements *before* AI, don't lean on the black box of the AI to cover that up. Here is my first set of Do's and Don'ts. More soon.

1) DO Break up big, vague tasks into multiple smaller, clearer steps

Do	Don't
<p>DO Break up complex, vague tasks into smaller steps which can be intersubjectively verified.</p>	<p>DON'T Ask AI to make broad evaluative judgments (like "Is this good?")</p>
<p>DO Document your methodology so that you can explain step by step how you reached your conclusions in a way which anyone can check. No black boxes. Use the AI to speed up many simple tasks which you <i>could</i> have done yourself if you had the time.</p>	<p>DON'T Trust the AI's explanations of how it reached its conclusions. AIs often create plausible-sounding but unreliable explanations after the fact. Normal AIs have very limited information about their inner processes</p>
<p>.</p>	
<p>DO Break up the data into pieces for AI analysis. Ideally run each piece as a separate prompt. Failing that, number each section and ask for a numbered, section-by-section answer, for example in a table.</p>	<p>DON'T Give an AI large pieces of text and expect it will pay due attention to all of it. It will <i>claim</i> to have done, and may provide references to relevant passages, but attention is <i>expensive</i> and it is always trying to reduce that expense. If you let it, it will always try to skim read and jump to conclusions.</p>
<p>DO Use explicit, manual methods (Excel?!) to synthesise the results of the multiple separate tasks you gave the AI.</p>	<p>DON'T Ask an AI to do maths for you, like adding up the number of positive or negative findings on a rubric. AIs are still terrible at maths.</p>
<p>Even worse, DON'T ask an AI to do <i>implicit</i> counting and comparison like "are there more positive or negative mentions of X in this report?"</p>	

AIs excel at specific, well-defined tasks that can be verified intersubjectively, like rubrics. Most importantly they can answer lots of them, quickly.

"Intersubjectively verifiable" just means that most people will more or less agree on the answer most of the time.

- It creates transparency and allows others to verify your work.
- Clear instructions lead to more reliable results.
- If you can't check it, you can't trust it.

Example of an intersubjectively verifiable task:

- Does this paragraph mention water and sanitation?
- If so, are any recent changes mentioned?
- If so, do these sound like positive changes according to the interviewee?

Notice that here we've broken down a larger task into three smaller and simpler steps.

Examples of tasks which are *not* intersubjectively verifiable:

- Is the intervention described in this report efficient and effective?

Text needs breaking up into sections, judgements on efficiency and effectiveness need breaking down into pieces, e.g. using rubrics.

- What are the main themes in this document?

This is a very common question in qualitative research, but it's a terrible task to give to an AI without further details. What do we mean by a theme? Are we interested in economic aspects? Interpersonal aspects? How are the themes to be identified and refined? Here, a whole world of qualitative social science experience, skills and workflows ([grounded theory](#), [thematic analysis](#)) have been bypassed in a single sentence.

- Summarise this document!

Yes, everyone does it. Evaluators do it. Schoolchildren do it. Pets will be doing it soon. As a quick time-saver for low-stakes tasks, it's very useful. But it's the vaguest, highest-level instruction, not a systematic analysis.

How do you break down a high-level judgement into a workflow of smaller tasks? Well isn't that what evaluation methods and qualitative research methods are for? Go read a book!

*We're not saying you have to specify *in advance* exactly what methods you will use. That's a bit too positivistic. But you should at least document them as you go along and be prepared to defend them when your analysis is done. That's the untranslatable [Nachvollziehbarkeit](#).*

At Causal Map Ltd, we've found that [highlighting and then aggregating causal links](#) is a great and relatively generic path from text data to the brink of evaluative judgement.

In terms of how to implement your workflow technically, see this [great contribution from Christopher Robert](#). At Causal Map, we're also working on ways to make workflows accessible. See how we currently use AI in Causal Map [here](#).

This post is based on my recent contribution to the [NLP-CoP Ethics & Governance Working Group](#), along with colleagues [Niamh Barry](#), [Elizabeth Long](#) and [Grace Lyn Higdon](#). In the next couple of weeks we'll look at two more do's and don'ts.

This post was originally published by Steve Powell on LinkedIn and has been republished here. See the original article [here](#)