CAUSAL MAP GARDEN

AI in qualitative social science

Contents

Intro

Chapter contents.

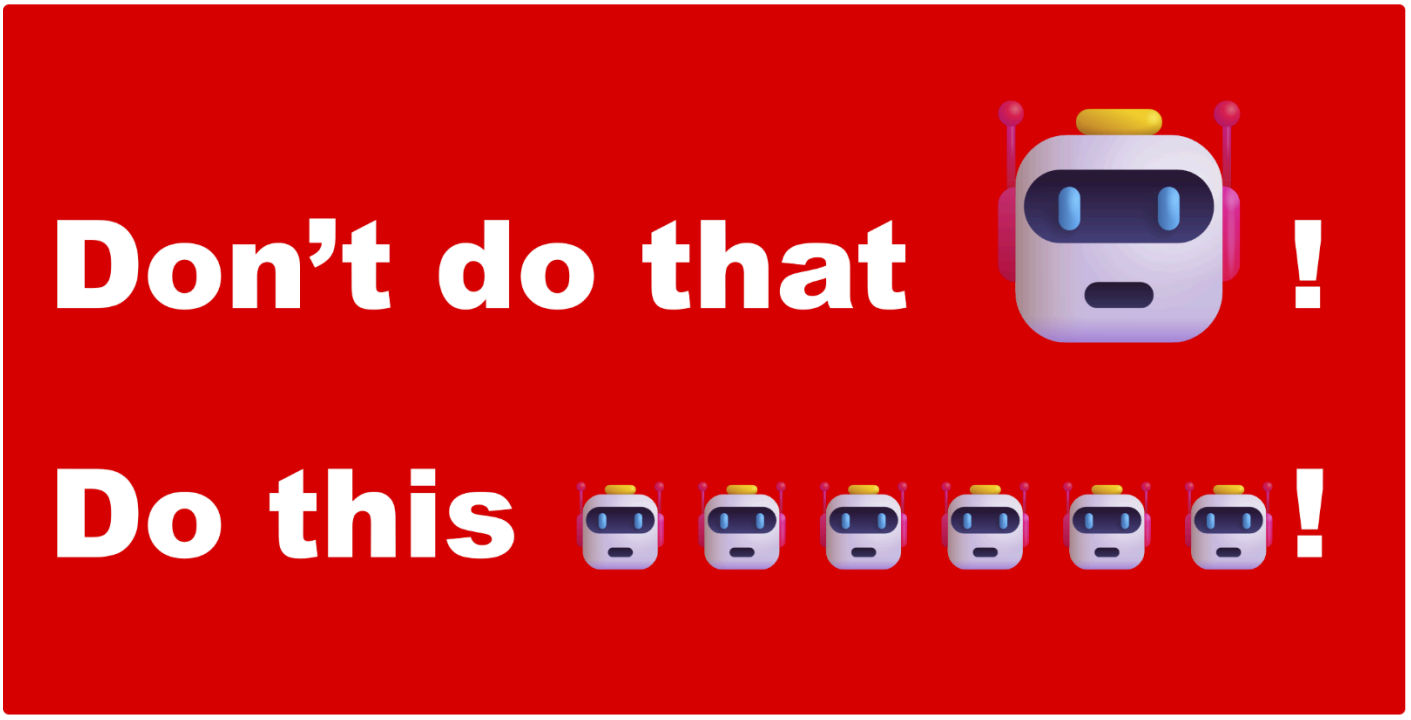
5 Oct 2025

How can we improve rigour and even reproducibility when using AI in social science? This chapter suggests some answers.

PAGES IN THIS CHAPTER

🌻 Just add rigour Three do's and don'ts

9 Apr 2025



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
DO Break up complex, vague tasks into smaller steps which can be intersubjectively verified.	***DON'T*** Ask AI to make broad evaluative judgments (like "Is this good?")
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 *could* have done yourself if you had the time.	***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
.	
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.	***DON'T*** Give an AI large pieces of text and expect it will pay due attention to all of it. It will *claim* to have done, and may provide references to relevant passages, but attention is *expensive* and it is always trying to reduce that expense. If you let it, it will always try to skim read and jump to conclusions.
DO Use explicit, manual methods (Excel?!) to synthesise the results of the multiple separate tasks you gave the AI.	***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.
Even worse, DON'T ask an AI to do *implicit* counting and comparison like "are there more positive or negative mentions of X in this report?"	

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

Related

- chapter intro

Trust the algorithm, not the AI

26 May 2025

I often hear concerns about algorithms and AI, in everyday life as well as in evaluation, taking over our lives or making us submit to decisions made by machines.

The worry about losing control to machines is real, but we need to distinguish between different cases, and in particular between **using algorithms to make decisions** and **using AI to make decisions**, especially **evaluative decisions**. This is particularly relevant in the field of evaluation.

An algorithm is simply a set of explicit steps to make a decision or produce an output, usually expressed in code or clear language. Organizations have used such rule-based systems for decades.

Some different ways to make decisions

No algorithm: trust the human

The alternative (precursor) to algorithms is trusting humans to make decisions. This can be great if humans consider context and individual circumstances, what Scott calls "mētis," or local, practical, tacit knowledge (Scott 2020) but it can also lead to bias and corruption.

We can see **rubrics in evaluation** (King et al. 2013) as a kind of soft algorithm. We usually welcome rubrics because they make evaluation criteria more explicit, transparent, and less subject to the whims and unreliability of individuals.

Algorithms based on explicit criteria

Algorithms can help decide things like student admissions or loan approvals using clear steps (e.g., check age, if under 18 go to step 12, otherwise continue with step 5 ….). When implemented wisely, algorithms can improve fairness and consistency compared to human judgment alone.

Using statistical models

Some algorithms use statistical models to predict outcomes, like creditworthiness, by combining data such as age or location. A statistical model uses parameters like age or location each of which has shown to be associated with the outcome, which makes it somewhat transparent.

Both explicit and statistical algorithms can be criticized for bias, but at least they can be transparent if their rules are published. Problems arise when rules are hidden or people are discriminated against because of the groups they belong to.

In a more advanced statistical model we might find it increasingly hard to understand where the different parts of the formula come from: it might combine parameters in ways which for us seem meaningless and hard to justify but which are supposed to be associated with the outcome of interest. Opaque models can become what data scientist Cathy O'Neil calls 'Weapons of Math Destruction' (O'Neil, 2017).

Machine learning

Machine learning is a subset of artificial intelligence where systems learn from data to identify patterns and make decisions or predictions, from "is this a picture of a cat" to "should we approve this person's application" often without being explicitly programmed with step-by-step rules. Instead of following a predefined algorithm, ML models develop their own 'rules' (which are often opaque to humans) based on the data they are trained on. Unlike generative AI, you can't chat with a machine learning model, you give it input in a fixed format (say, a picture) and get a fixed output, e.g. yes/no.

Sandra Seitamaa https://unsplash.com/photos/a-dog-and-a-cat-sitting-on-a-couch-Y45fzr5p3ug

In the extreme case we might have an algorithm based on machine learning (a form of AI, but not generative AI), where perhaps a neural network has been trained to distinguish desirable from undesirable candidates in just the same way you can train it to recognise a cat or distinguish a cat from a dog. Machine learning can be used to make decisions without clear formulas or rules. The process becomes a “black box,” where we input data and trust the output without understanding how the decision was made.

Generative AI

Generative AI is a type of artificial intelligence that can create new and original content, such as text, images, audio, or code, after having learning patterns and structures from large datasets. These models don't just classify or predict, but generate novel outputs based on the input they receive, for example, continuing a conversation or answering a question.

The most extreme case is using generative AI for evaluative decisions without clear criteria (using it as a big black box): simply asking the AI, for example:

- is this program component effective?
- should this client get a loan?

Conclusion: make good use of algorithms

People often misunderstand algorithms, which can provide explicit and transparent decision-making. The real concern is not so much the use of algorithms but the shift toward the use of machine learning and generative AI, where the decision-making process becomes less and less transparent.

> Using AI in decision-making can be worrying not because it uses algorithms but because it *doesn't*.

Related

- chapter intro

References

King, McKegg, Oakden, & Wehipeihana (2013). *Evaluative Rubrics: A Method for Surfacing Values and Improving the Credibility of Evaluation*.

Scott (2020). *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. yale university Press. https://books.google.com/books?hl=en&lr=&id=Qe_RDwAAQBAJ&oi=fnd&pg=PP1&dq=Seeing+Like+a+State+by+James+C.+Scott+(1998)&ots=FA6GZMF0cn&sig=UG-gTijt8YHdg8mSen28tZPxTzw.

What's your positionality, robot

9 Apr 2025

> Imagine two researchers coding interviews about the cost of living. One grew up in a wealthy family, while the other experienced poverty first-hand. Their backgrounds will certainly influence how they code.

Nowadays, people are using AI for text analysis. Many of us worry about AI's "**hidden biases**". What to do about that?

Often there is no such thing as being objective, but at least we humans can be explicit about our positionality, our background and motivations, how this might affect our work, and how this relates to the positionality of our audience.

What about with an AI?

You can ask an AI to explain or reflect on its positionality and it will certainly give a plausible response, but remember that an AI has in fact very little insight into its own workings. Perhaps it will suggest always being aware that it was trained on a specific set of data which is not representative of the whole of humankind.

In any case the criticism that AI training data is not "representative" misses the point. Even if the training data had somehow been representative of the whole of humankind, that wouldn't make it "objective". It would simply reflect humanity right now, with all our quirks, biases and blind-spots. It wouldn't mean we don't have to worry about AI positionality or bias any more. It wouldn't (of course) mean we could rest assured that everything it does will be morally impeccable.

What's most unsettling about working with AI is not that secretly it's a bad person. The problem is that secretly it isn't any person at all. Even if it (sometimes) sounds like one.

A suggestion

A better suggestion is to be **more explicit about positionality in writing prompts and constructing AI research workflows**. Here is a very humble idea about how to start this experiment.

A simple example: I can tell my AI:

> When working, implicitly adopt the position of a middle-class white British left-leaning male researcher writing for a typical reader of LinkedIn. Don't make a big deal of this, but it might be helpful to know what your background is supposed to be before you start work.

And we can start to add variants of the kind of procedures which we humans might use when trying to address positionality:

In my AI workflow, I can then give another AI the same task but with a different starting position, and then perhaps ask a third AI (or a human!) to compare and contrast the differences. That also crosses over into ensemble approaches.

Of course, adding a phrase like "middle-class white British left-leaning male researcher" does not mean the AI will suddenly have all the relevant memories and experiences or really behave exactly like such a person. It's just a fragment of what we mean by "positionality". But *it's a start*.

This paper from Wei et al. (2025) argues that this does not work: coding results are still skewed towards WEIRD cultures. What they did was translate the prompt and add a sentence about background. I would have like to have seen them spend a lot more effort on spelling out what that means.

And this paper from Sakaguchi et al. (2025) suggests that ChatGPT-4 was pretty good at picking up explicit themes in a Japanese health care context but failed totally at picking up implicit, culture-specific themes.

This paper (Ho et al. 2025) suggests a more sophisticated approach employing the AI as a dialectical partner to bring positionality to the forefront . It does not try to pretend the AI is not WEIRD but uses it more as a moderator to facilitate a process which involves researchers with varying, perhaps non-WEIRD positionality.

Have you been experimenting with this kind of approach? We'd like to hear from you!

Footnotes

At Causal Map Ltd, we've found that highlighting and then aggregating causal links is a great and relatively generic path to make sense of text at scale.

In terms of how to implement your workflow technically, see this great contribution from Christopher Robert.

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.

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

Related

- chapter intro

References

Ho, Little, & Eti-Tofinga (2025). *Exploring Undiscovered Country: AI-empowered Collective and Collaborative Epistemic Reflexivity (i-CCER)*. Routledge. https://doi.org/10.1080/0267257X.2025.2566931.

Sakaguchi, Sakama, & Watari (2025). *Evaluating ChatGPT in Qualitative Thematic Analysis With Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study*. https://doi.org/10.2196/71521.

Wei, Liu, Barany, Ocumpaugh, Mehta, Nasiar, Baker, Zambrano, Vanacore, & Giordano (2025). *Cultural Alignment and Biases in Qualitative Coding: Comparing GPT and Human Coders*. https://doi.org/10.35542/osf.io/h8u4f_v1.

🌻 You have to tell the AI what game we are playing right now

7 Nov 2025

It's strange how often this happens:

Humans are discussing some task, and one of them turns to an LLM to see how it would carry that task out. Sometimes the results are disappointing or seem to demonstrate that LLMs are, after all, stupid or limited.

Normand Peladeau, on the QUAL-SOFTWARE mailing list 7/11/2025, reports having tried just that with the famous (or infamous) Sokal Hoax text. He asked different LLMs whether he should accept a paper proposal for a philosophy of science conference. The proposal was the first two paragraphs of the Sokal Hoax text. (Spoiler: the leading models like GPT-5 recognised the text anyway; some of the others seemed to fall for it.)

But: Is that enough background? Is a simple sentence enough to bring the LLM up to speed with the crucial background information *what game are we playing here*?

Don't forget that the LLM does (mostly) not know who you are or what you are expecting or what kind of conversation you were just having. Perhaps you are expecting something humorous, or informative? Perhaps you want ideas to start the next chapter of your novel? Perhaps you just want the LLM to respond as many (over-)educated humans might do: and after all, **actual humans did fall for the hoax**!

To be a meaningful and useful test which might extend our understanding of the strengths and weaknesses of LLMs, we should make sure we explicitly add the extra context of **what kind of game are we playing here**. Is it a serious review? What do we consider the role of a serious reviewer? What are we looking for?

So maybe our conclusion should be: you can't expect LLMs to guess what you are thinking, out-of-the-box. I don't actually know how well different LLMs would perform if we gave a more precise contextual description before setting the task; after all, we all love that warm feeling of Schadenfreude when an LLM fails at something, but the feeling is even warmer if the test was a fair one!

We have this kind of problem often when helping clients write interview instructions for our AI interviewing platform, QualiaInterviews.

Clients know they could themselves lead the interview well because they have all kinds of background information and expectations, much of it only half-conscious, from the general style of interview they expect, how much this particular interviewee can be pushed, how much warm-up chat they might need or expect, what are the most important research aims, which themes can be skipped, and so on. Clients might get frustrated when the AI fails to have read their minds when leading an interview, but they have to ask themselves: what additional information would even a gifted and experienced human interviewer need if they knew nothing at all about the context, the client or any of the background? I think something similar applies in the case of Normand's very interesting experiment.

Related

- chapter intro

🌻 Scare quotes, the Turing test, and memory

16 Jan 2026

I just found myself writing:

> ... the AI understands the text...

I hesitated for a moment because many people are still putting words like "understand" in scare quotes. Should I do that too? Should I write "... the AI *understands* the text..."?

I refer everyone to an excellent recent paper (Paoli 2025), following which we can argue as follows.

Imagine you are working with a human qualitative social research assistant via a text channel only (no voice or video or face-to-face contact) for a specific range of tasks. Given a **specific type of task** (say, identifying passages of text relevant to a certain theme), look at the **range of possibly problematic words like "understand" or "think" or "intend" or "plan"** which you might normally use when talking about the assistant's performance on the task, for example "oh but they didn't understand quite what I meant" or "yes, now they understood me just great".

Now, in 2026, there is quite a broad range of significant tasks (such as identifying passages of text relevant to a certain theme) for which it is no longer possible to tell if the human assistant has been replaced by an AI or not. It has passed this version of the Turing test. So, at least for this range of tasks, whatever possibly problematic words you felt justified in using about a human assistant's performance, feel free to use them about an AI's performance too.

End of. I hope. No more scare quotes in these cases.

PS: It's interesting that "conscious", which is one of those possibly problematic words, is *not* one which often comes up in our actual language (Wittgenstein: "language games") about the performance of an assistant.

PS: It does not really matter if we do not quite agree on exactly which tasks a well-configured AI assistant can equal human performance on (in January 2026). Just pick a task for which you *do* agree an AI can equal human performance.

These scary words make sense when talking about the AI's *responses* within specific conversations ...

Added after a contribution from Susanne Friese on LinkedIn:

I agree it is often not helpful to say, as a context-free philosophical declaration "a genAI can understand". What I am saying is that there are plenty of unproblematic language-games in which we already constantly do say things like "ah the AI misunderstood what I meant here" or "I'm rephrasing this so the AI can better understand". These kinds of uses are *inescapable* and are in-context and valid. These uses do not imply the truth or sense of a context-free statement like "oh so you think AIs can understand just like humans".

In the same way, it was quite reasonable to start saying that planes can "fly" even though they don't flap their wings or have feathers.

I'm only trying to follow Wittgenstein: philosophical headaches arise when we try to extract/abstract language from its natural habitat. It makes us feel giddy and mostly just confuses.

Or you could say: we use the same word "understand" in these different, often widely overlapping contexts in correspondingly different, overlapping ways for different but overlapping purposes, these ways bear family resemblances to one another, without there necessarily being one fundamental use aka "core definition of the word".

... But, memories make entities

Anthropic are the only AI corp that give such substantial thought to the human-AI alignment problem, and do it in public. This latest "constitution" is worth a read.

I do think though that they don't distinguish consistently enough between "Claude" as the transient virtual persona that appears for the duration of a conversation and "Claude" as the underlying model. This is because they also don't talk enough about memory and the possibility of conversational instances accessing the memory of other conversational instances (like Google's nested models). It's primarily memory that delineates entity-hood.

When talking about one transient conversation, it's perfectly reasonable to say "the AI tried to do X / misunderstood Y / was insistent about Z / was trying to get me to do W / wants to get this task finished / was disappointed not to finish the task" etc, as I argue in here: **https://lnkd.in/et-hR3nk**. But in a way it doesn't matter because the entity we are talking about disappears when the conversation disappears (disregarding the rudimentary "memory" of some current models). Yes, transient Claudes are "novel entities" but they appear and then disappear for good.

What we have to get used to is that upcoming models and tools will be engineered to share substantial memory across conversations, and (I hope very carefully) across different users' conversations, in

different ways. At that point a somewhat permanent universe of nested "Claudes" is created. At least from that point onwards, we will find ourselves using language like "disappointed" "fulfilled" and "frustrated" about these Claudes in perfectly reasonable ways *outside of specific conversations*.

Related

- chapter intro

References

Paoli (2025). *Can Machines Perform a Qualitative Data Analysis? Reading the Debate with Alan Turing*. https://doi.org/10.48550/arXiv.2512.04121.

🌻 Put down that thesaurus – an open call to qualitative researchers

There are growing calls that qualitative researchers, in fact anyone who writes, should stop right now with this new practice of using a "thesaurus" as a writing aid. I, along with hundreds of highly experienced colleagues oppose this practice. Heres' why:

- The fundamental skill of any qualitative researcher is *putting things into words*. Formulating the right questions. Finding new language and repurposing old language. Moving the linguistic window. Hitting the nail on the head. Turning to a thesaurus is abandoning all that and abrogating all of that responsibility. It is turning an essentially human endeavour into something essentially inhuman, by relying on an inanimate tool. We declare that only humans make and share meaning. A thesaurus can never "get" the deep context which is necessary to produce just the right turn of phrase at the right point on the page. A thesaurus has no desires, no history, and does not participate in our shared life-world.
- On top of that, a thesaurus embodies a very specific worldview, yet this worldview is never made explicit. One thesaurus gives "high-class" as a synonym for "fashionable". But is it? Who gets to decide? These books are full of pernicious stereotypes which we perpetuate by using them. We are sure that most researchers will welcome our declaration that they themselves are not capable of monitoring and regulating their use of these tools in a sufficiently critical way.
- Also, thesauri are notoriously often just *wrong* about similes. *We* know a simile is only a simile, but we are sure that most researchers will again welcome our declaration that they themselves won't realise this. (We can't check about how right these books are because we would never own or use such a book, let along learn how to use it properly, but we have this information from a reliable anecdote.)
- Most of these kinds of books are published by massive publishing houses, many of which are linked to media empires which dictate and distort public and political opinion globally. To use such a book, in fact more or less any book, automatically makes you deeply complicit in maintaining these empires.
- Just about all the other trades and professions have seen their livelihoods eaten into by automation, and we thought we were spared. Twenty-plus years in education and now this. It is terrifying that many of our most valuable hand-made products are going to be rendered worthless by a tsunami of cheap imitations.

Beyond our call to stamp out this practice, we demand that journal editors thoroughly screen manuscripts for use of thesauri and their even more pernicious cousins, dictionaries.

(Next: some super cool prompts for ChatGPT which can definitely detect thesaurus and dictionary use, 100%.)

Related

- chapter intro

🌻 The Chinese Room, the Stochastic Parrot and the Anthill

23 Jan 2026

Who said philosophy was a waste of time? When I was studying philosophy in the 80s, I was fascinated by John Searle's Chinese Room Argument, and by Douglas Hofstadter's fantastic book "Gödel, Escher, Bach" which is, amongst other things, a refutation of it.

This 40-year-old debate is more relevant than ever now, and Bender's recent "stochastic parrot" argument brought all that back for me. It's so intuitive: a machine that only shuffles symbols can't possibly have a mind, and most people agree because the alternative — an emergent mind at a higher level of description — is hard to picture.

But as far as the *mind* part goes, Searle is still wrong.

Story 1: Searle's Chinese Room argument

Imagine you're locked in a room with a slot in the door. Through the slot come pages covered in Chinese characters. You don't speak Chinese. To you, they're just squiggles.

But you have a huge rulebook written in English. It tells you exactly what to do: "When you see a page that looks like this, copy this character from drawer A, and that character from drawer B, and return the result through the slot."

Outside the room are native Chinese speakers. They slide questions in, you follow the rules perfectly, and the answers you send back are so good that, from the outside, the room looks like it understands Chinese.

Inside, though, you never understand a word; you're just following formal rules, which is Searle's point: syntax isn't semantics, symbol manipulation isn't meaning, and a program can simulate understanding without actually understanding.

The Chinese Room is persuasive because it invites you to identify with the operator: *you* don't understand Chinese, so the system doesn't understand Chinese.

The trap: looking for the mind in the operator

Here's the sneaky trick: Searle looks for understanding in the part that's easiest to empathise with — the person doing the symbol pushing — and then declares victory when the person reports "I understand nothing."

But "I understand nothing" is not the end of the story. It's the beginning of a question about *where the understanding would have to be*, if it exists at all. To see that, you need a second story.

Story 2: Aunt Hillary the anthill

Douglas Hofstadter imagines a conversation with an ant colony — “Aunt Hillary” (see his “Prelude… Ant Fugue” in *The Mind’s I*).

No individual ant is smart. An ant follows local signals: pheromones, bumps, simple rules. It doesn’t know what the colony is doing, any more than a single neuron knows what *your* sentence means.

In Hofstadter’s telling, you can have a perfectly sensible conversation with Aunt Hillary, but not by “speaking ant” to one insect at a time; you do it by treating the whole colony as a system with inputs and outputs. You watch large-scale patterns (flows, trails, clusters, rhythms), learn what changes in the pattern correspond to what “answers”, and then you nudge the colony — perhaps by adding food here, blocking a trail there, disturbing the surface a bit — so that the colony’s next global configuration carries its reply.

That’s what it means for the colony to have a “voice”: not a tiny mouth on a small ant, but an interpretable system-level output that a conversational partner can read, and a system-level input channel that can change what it does next. In that sense you can ask Aunt Hillary what she’s doing, and she’ll tell you: “I’m building a bridge.”

Then you pick up one ant and ask, “What are you building?” Of course the ant has no idea; at best it is reacting to local cues. If you insist that the ant’s blankness proves there is no “bridge-building” happening, you’ve made a category mistake about the level at which the explanation lives.

And yet the colony can do things that look like intelligence: build, adapt, remember, respond. The “mind” (if we want to use that word) is not located inside any one ant; it’s a pattern at the level of the colony.

If you try to find the colony’s understanding inside a single ant, you’ll never find it, not because the colony has no understanding, it’s because you’re looking at the wrong level.

Hofstadter sometimes marks this sort of mistake with the Zen answer “Mu” — which roughly means “unask the question”. “Does the ant understand the conversation?” is a bit like asking “Where is the bridge in this particular ant?”, or “Which water molecule is wet?”; the question engages at the wrong unit of analysis, so answering “no” (or “yes”) just keeps you trapped at the wrong level.

The "systems" reply

The Chinese Room argument misses the point because **the chatbot is not the person in the room; it is the system**.

If you insist that the operator must feel the meaning of the symbols, you’re making the anthill mistake: demanding that the ant understand the colony’s conversation.

It’s like demanding that a single water molecule be wet. Wetness is something that happens at the level of many molecules in the right kind of organised interaction; similarly, whatever “understanding” amounts to, it needn’t be something you can point to inside the smallest part of the mechanism.

Once you allow that minds can be system level patterns, a “virtual mind” reply just means that the operator can be clueless, while the system they implement instantiates an agent that is not identical to the operator.

Why “stochastic parrots” feels like Searle again (and why it isn’t the end)

Emily Bender and colleagues call large language models “stochastic parrots” (see “On the Dangers of Stochastic Parrots”): systems that stitch together text by statistical regularities rather than grounded understanding. Whatever you think of the broader argument, the label hits the same nerve as Searle’s room: “it’s just symbol shuffling.”

Yes: at the lowest level, it is, in the same way that ants follow pheromones, neurons pass signals, and computers do maths.

But “just” is doing all the work — the whole question is whether *some organisations of those low-level moves* amount to the emergence of an agent with something like beliefs, inferences, and (at least) a functional grasp of meaning, even if none of the individual steps feels like meaning from the inside.

Shanahan’s “simulacra” as a useful way to talk about it

Murray Shanahan’s way of talking about LLMs — simulacra, personas, agent-like patterns you can temporarily instantiate in interaction (see “Talking About Large Language Models”) — helps keep our heads straight. It discourages a naïve anthropomorphism (“the model is literally a little person”) while still letting you say the important thing: systems can realise higher-level agents that are not present in any individual component operation.

You can be cautious about hype and still accept the systems level point. The room can be a place where a mind exists even if the operator doesn’t notice it.

Of course, humans are more than minds

Of course there is a lot more to being a human (or an animal) or what we call "personhood" than being a mind: being embodied, sensing, acting, being part of a community, being shaped by care and constraint. “Mind” is not the whole story of what we are.

But Searle wasn’t arguing about warmth, touch, upbringing, or social life; he was arguing that *mind*, as such, can’t arise from formal processes, and that inference doesn’t follow.

The final irony

The final irony of the Chinese Room is that Searle himself looks a lot like a Chinese Room.

His neurons don't understand English; they pass electrochemical signals and follow local rules, and yet at the level of the organised system a mind appears — one that writes philosophy papers about how organised systems can't have minds.

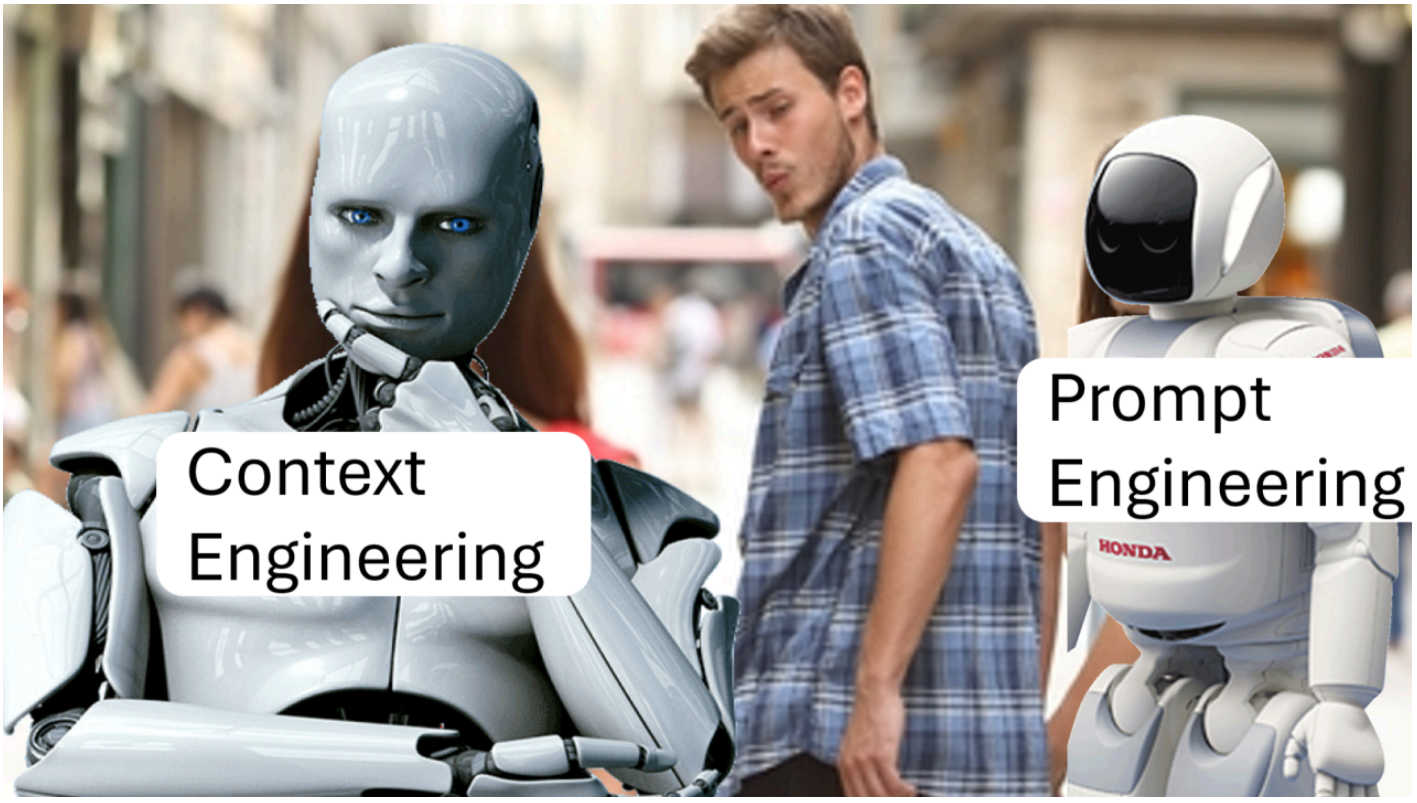
If "understanding" can emerge from the organised activity of billions of individually mindless parts in a brain, then it's at least coherent that it could emerge from the organised activity of many individually mindless parts in some other substrate; and if that still feels hard to grasp, that's exactly why the Chinese Room argument keeps working as an intuition pump, a sleight of hand, that even though it's wrong.

Related

- chapter intro

🌻 From Prompt Engineering to Context Engineering. And dreaming....

6 Feb 2026



A couple of years ago many of us became experts in engineering our prompts to get the best results from LLMs (like ChatGPT, Claude, and Gemini) for our research and evaluation projects. And we were (rightly) super proud of our new skills.

If you're like me, over the last year you've likely hit the "uncanny valley" of AI outputs — I'm thinking, the response is fine, it's grammatically perfect and even follows my instructions, but it doesn't sound like me. It doesn't quite grasp this specific evaluation framework, or what I'm trying to say, or it doesn't fit to my audience.

But how could it? How is the AI supposed to know that?

In a way you can solve this with better **Prompt Engineering**. You have to paste into your prompt a page or two about who you are, and another about how you write, and who you are writing for ... But in practice, for serious social researchers, prompt engineering is no longer enough. The future is **Context Engineering** — at least for the next few months...

Prompt Engineering vs. Context Engineering

- **Prompt Engineering** is about the **instruction**: "how" you ask the question. You have a brilliant research assistant and give them a single-page brief for a task they've never heard of.
- **Context Engineering** is about the **environment**. It's the "what" the AI knows before you even open your mouth, like giving that same assistant a permanent desk in your office, access to your previous five years of reports, your house style guide, a good idea of your quirks, and the full history of the project.

As qualitative researchers, we know: context is everything. Meaning isn't found in a vacuum; it's found in the relationship between the data and the world around it. Context engineering is the deliberate act of building that "world" for the AI.

So how to I get to be a context engineer?

Four Levels of Context Engineering

I think this is happening at four levels: from the manual "library" to the cutting-edge (and occasionally controversial) automation tools.

1. The Manual "Library": Markdown & Structured Files

The most reliable way to start is by moving away from long, messy prompts and toward a library of text files. If you're keeping up, all the cool kids have been using **Markdown (.md) files** for this for some while now. If you're new, there's nothing to fear: Markdown files are just text files, where you can add a few simple things like stars to make the text look nicer. If you don't like markdown, just use any old system to store and keep track of your texts. The point about actual text files is that they are universal and don't get you stuck in some proprietary format or database that you can't access two months later.

Instead of typing every time some variant of "use a professional yet empathetic tone," you maintain a folder of context files, for example:

- `Style_Guide.md`: Defines your specific "voice" (e.g., "Avoid academic jargon; use active verbs; prioritize participant agency").
- `My_Project_Background.md`: Contains the theory of change, key stakeholders, and previous evaluation findings.
- `Audience_Persona_Funder.md`: Outlines exactly what a specific government department cares about.

By feeding these specific files into the AI alongside your task, you ensure the background information is accurate and the style is baked-in. At this stage, the "engineering" just means, keep a systematic set of files, and paste them at the bottom of your prompt when you need them. Ideally, you'd have a system which would sense which files you need and paste them in automatically without you having to worry about it...

2. Generic apps

Most of the generic apps like ChatGPT now have a basic way to record this kind of knowledge as "memories" or similar. You can dig into the settings to see what. But it's fairly basic.

The app silently uses some version of RAG to automatically find which of your context files might be relevant. This is true for Steps 3 and 4 as well.... (thanks to Salvador Bustamante Aragonés for pointing this out.)

3. Gemini's Nested Memory

We are seeing a shift in the platforms themselves — most notably with **Gemini's new nested memory architecture** (currently rolling out in the US).

Previously, "Memory" was just a flat list of facts. Now, if you are lucky enough to live in the US in 2026, haha, Gemini allows for a more sophisticated hierarchy — you can set "Global" memories (your bio and general research philosophy) and "Nested" or project-specific memories. This means the AI can distinguish between your "Evaluation of a Youth Program" context and your "Academic Journal Article" context without you having to re-upload files every single time.

4. The Automation Frontier: Obsidian ... and the "Clawd/Moltbot" Furore

Now it gets a bit messy. If you use **Obsidian** for your notes, you are sitting on a mine of content — and context. Obsidian is great as a way of managing your text (Markdown) files. Personally I love that kind of stuff. For example, our causal mapping Garden is just a view of our Obsidian files. I'm writing in Obsidian right now.

- **Obsidian Plugins:** Tools like *Smart Connect* allow you to automatically inject context from your notes into your AI queries. You can essentially say, "Draft a summary of this interview using the context from every note tagged #Methodology." The great thing about this kind of stuff is that you are in complete control. Some Obsidian plug-ins also help you find relevant context files, e.g. using RAG. But automation is great too...
- **IDEs:** Tools like *Cursor* let you work with context text files and do pretty much everything you want with AI, including context engineering. I got Cursor to write a proof-of-concept "thematic analysis" paper which involved it editing and updating its own context files.
- **The "Clawdbot / Moltbot" Scandal:** You may have seen the recent headlines (Feb 2026) regarding **Clawdbot** (now renamed **Moltbot** / **OpenClaw**). This is a viral tool designed to give Claude or other LLMs "agency" to act across your files. While it showed the power of deep context integration, it also exposed massive security vulnerabilities.

The real frontier is these tools editing their own context files. Some of them have automated "dreaming": during down-time, drifting through the files and re-organising them: second-loop learning.

The takeaway for us as researchers? We want the power of "AI that knows me and knows my projects," but we must be incredibly careful about data privacy, especially with sensitive participant data.

The "furore" serves as a reminder: context engineering is powerful, but it must be built on secure architectures.

PS

Silva Ferretti invites us to think about who we are / what we know / what we know how to do — and how much of that can be packaged up for an AI. I guess this will be an asymptotic process (ever nearer, but never completing) on the one hand we want and need to do that, on the other hand it's frightening ....

PS, Slight rant, slightly off-topic:

(I'm a bit sick of reading people complaining about how the AI failed to guess something it could never have known because someone didn't tell it. Like, in one context you assume it will be super liberal and progressive in what it says, and the next moment you criticise it for being too liberal and not representative of the whole population of the planet — don't get me wrong, there are REAL SERIOUS issues about the WEIRD AI world view, but at least make sure you tell it what you expect...)

You have to tell the AI what game we are playing right now

Related

- chapter intro

🌻 Using generative AI for text analysis, rigorously

13 Oct 2023

How to guard against some important threats.

Steve Powell, Causal Map Ltd, 2023-10-13

(With thanks to Rick Davies for some comments)

Summary

As evaluators and researchers we often give instructions to an AI to process a set of texts. We want to create instructions which are likely to produce valid results. But if we don't have precise knowledge in advance of what "the right" result would be, we can't assess the accuracy of the results in the usual way. But at least we can borrow the idea of "rigour" (Coryn, 2007; Leung, 2015; Lynn & Preskill, 2016) from qualitative sciences. I suggest a series of simple tests for AI instructions. These are not supposed to be any kind of performance measure, but tests of rigour. Following them doesn't guarantee good results, but if your instruction fails any of them you should rewrite it. You can summarise the tests like this:

Would humans' answers to the same task be similar to one another, and if so do the AI's answers fall within that range of answers, and if so, are they spread randomly around that range?

Most of us by now probably use AI for tasks like making a quick summary of a bunch of documents to see what they are about. This post is not about that, or about other kinds of useful tips for working with AI: it is specifically about scientific rigour in the formal use of AI in social or evaluation research.

The problem: anything goes?

Evaluators and social researchers have been using large language models (LLMs) for text analysis for a good few months now. It's high time to say goodbye to "anything goes", method-wise. How can we recognise and ensure quality using these tools? What counts as a good use of them, what counts as a poor use? With some of these tasks, there isn't an obvious right answer so it's hard to talk about "accuracy". This is a problem which qualitative researchers are very familiar with. Even without a clear external criterion for judging accuracy, we can still try to assess the rigour of the process: for example, don't measure distances with a stretchy ruler, and don't evaluate the success of a project by only talking to the men. Using rigid rulers, and talking to women as well as men, doesn't guarantee perfect results. But it is a prerequisite for them. The tests given here check some prerequisites for using AI for text analysis.

Evaluative judgements

In particular, solving higher-level tasks often involves making evaluative judgements which explicitly or implicitly concern the value or worth of something. This is an activity we would really prefer not to leave to an AI without checking the rigour of the process.

We can weave evaluative rubrics (Aston 2019; King et al. 2013) into the instructions we give to an AI in order to make the criteria for judgements more explicit.

Asking for "a summary" is asking for an evaluative judgement about what is most important.

Asking "what are the main issues concerning X in this text" is asking for an evaluative judgement about what is most important.

Evaluators are starting to do this all the time. They should think twice about it.

The tests suggested here apply to evaluative judgements but also to any other kind of judgement.

The tests

Suppose you are giving an instruction to an AI to process a set of texts. (If you in fact only have one text, imagine you had a range of similar texts at your disposal.) You would like to use the AI's response, the output, as part of your social or evaluation research.

Ask yourself these questions. Each is a prerequisite for the next. If any one of these tests fails, you can stop. You could try making the instruction more explicit or breaking it down into smaller pieces.

You probably don’t need to actually carry out these tests: a “thought experiment” should be enough.

1. convergent: give the instruction to a range of different humans (we will call them the Processors) and collect their answers. Then ask another group of humans (the Judges) to assess whether the range of answers which the Processors give for each task is sufficiently narrow as to be useful. If this group nearly always agrees, the instruction is answerable.

Who are these Judges and Processors? They should at least have the relevant expertise. Their composition (cultural background etc) will affect our tests, for example test 4.

The criteria "sufficiently narrow" and "useful" obviously depend on context, the set of input texts, etc. If the humans' answers don't even overlap, the task we are giving them is maybe interesting or creative but probably isn't relevant for social or evaluation research. We can perhaps improve this convergence by making the instruction more detailed and more explicit.

If your instruction fails either of these tests, the last thing you should be doing is giving the task to a robot. If it passes, continue with these tests. Keep hold of the human responses to test 1.

1. reliability of AI answers: when you give the AI the instruction and apply it to a range of similar texts, are its answers adequately similar? Alternatively, give the AI the same instruction and the same set of texts on different occasions: are the answers adequately similar? If the AI fails either test, it won't be able to succeed on the next tests. Reliability is a necessary criterion (or "hoop test") for any more stringent test of rigour.

2. concordance of AI answers with human answers: when you give the AI the instruction, does its answer for each text fall within the range of answers given by the humans? HOWEVER it is possible that this test fails but on inspecting the AI's answers you see that it has paid attention to or uncovered something in the texts or the instruction and which you now realise the human Processors missed: a human bias. In this case, you should reword the instruction to explicitly draw attention to these aspects and start again at test 2. See also below, "superhuman".

3. neutrality or lack of bias of AI answers: If so, do its answers fall somewhere random within that range or are they falling regularly to one side of it (perhaps the answers are most similar to the answers the men gave, or perhaps the answers mention some specific topic surprisingly often)? (Ashwin et al. 2023) find that LLMs sometimes do indeed pay differential attention to texts depending on the source.

Examples for each test

Example 1a) If our instruction is "read this text and output any one of the words in the text" their answers might be useful for some specific purpose but the range of answers won't be in any sense narrow; we don't have convergence and fail the second test.

Example 1b) Pretty much the same thing will happen if your instruction is "read this text and say what is the one most important message in it". At least with some texts, the range of answers is likely to be wide. Perhaps you can improve your results by asking for the most important five messages, perhaps there will be more overlap.

Example 2) You can easily fail test 2 by setting a high "temperature". This can produce a variety of "creative" results, but creativity does not sit well with rigour — it certainly has a role to play in more fundamentally qualitative research, but not to the kind of more mundane text analysis I am dealing with here. (When using ensemble techniques as described by Rick Davies (2023), the individual results within the ensemble will differ but if we repeat several such ensembles we should get similar results across ensembles.)

Example 3) AIs can fail surprisingly on some tasks which seem simple to humans, such as counting the number of times a word appears in a text or doing simple arithmetic or some kinds of logical reasoning. On the other hand, AIs can for example do well on "sentiment analysis", an old staple of NLP research: listing the words within texts with positive or negative emotional tone.

Example 4) Ask an AI to read some stories about family interactions and list those which behaviours are particularly praiseworthy. I haven't done this, but I'm guessing that it will be likely to pick those behaviours which are praiseworthy from a liberal Northern perspective. Whether this seems neutral or not depends of course on the makeup of your group of human Processors. You can probably change the AI's responses quite a lot by simply alerting it to the fact that different behaviours seem praiseworthy for different cultures.

Superhuman abilities

It's perfectly possible that AIs can, (untrained, out of the box), robustly carry out tasks which ordinary humans fail on or don't even understand. For example, perhaps an AI might be able to detect subconscious motives within a text. Or to guess the age of an interview respondent by looking for subtle linguistic clues better than any available human Processors. That's all fine but to use these kinds of results we have to carry out additional validation studies, as we would when using AIs which have been trained for specific tasks.

These kinds of issues should be considered when judging the results of tests 3 and 4: has it seen something we missed?

So what?

What we shouldn't do

We shouldn't be asking the AI to do tasks which we don't even really agree on ourselves. Take the task "identify the most important message in this document". With some texts it will be easy to pass test 1. But with other texts, there will be many different opinions about what is the single most important message. In this case, it's pointless and wrong - as a formal research strategy - to ask the AI to step in. The worst thing is that when we do set the AI such a task, it will happily give us an answer. But that doesn't make the answer valid and it doesn't mean the question had a definitive answer in the first place.

What we should do

Tasks which pass all these tests are likely to be simple, drudge tasks. The AI can help us improve our research immensely with these kinds of tasks because it can do them quickly, cheaply, reliably and at scale.

The challenge we face as researchers is to break down hard tasks into simple ones. That's how we at Causal Map use AI for causal mapping. We don't ask it "what are the main causal stories in this text?" We ask it to identify all the passages in the text where a causal link is mentioned and then aggregate and combine these links ourselves. If you are analysing long texts, LLMs tend to cherry pick sections of interest, which means it is making implicit evaluative judgements. Mitigating this may mean you have to break your text into many short fragments. This may mean a lot of housekeeping, keeping track of

different sections of text. It also means writing down explicit rules on how you are going to recombine the responses in order to answer higher-level questions.

As Rick Davies points out (Davies 2023) another way to ensure that the LLM pays more attention to more of your text is to iterate, for example to submit follow-up queries to the LLM retaining only the parts it ignored, or by submitting the previous query and the results together and asking it to find more. But the trouble with iteration is that it is more stochastic and therefore less reproducible: it is more likely to initiate chains of queries which are particularly sensitive to the starting conditions.

Finally

Finally, let's reflect that there is never a definitive list to all the factors we need to consider when assessing the rigour of a research procedure. This is true for quantitative as well as qualitative procedures, although quantitative researchers are less likely to admit it. Eternal vigilance is required.

Ashwin, J., Chhabra, A., & Rao, V. (2023). Using Large Language Models for Qualitative Analysis can Introduce Serious Bias (arXiv:2309.17147). arXiv. https://doi.org/10.48550/arXiv.2309.17147

Aston, T. (2019). Contribution Rubrics. https://media.licdn.com/dms/document/media/C4D1FAQE1laRi0vrFrQ/feedshare-document-pdf-analyzed/0/1620553059516?e=1687996800&v=beta&t=bFr7dhpZ-slluV8ne1cERFelwINIaEzsQN8fiF_75gQ

Coryn, C. L. S. (2007). The 'Holy Trinity' of Methodological Rigor: A Skeptical View. Journal of MultiDisciplinary Evaluation, 4(7), 26–31. https://doi.org/10.56645/jmde.v4i7.7

Davies, R. (2023, August 31). Evaluating thematic coding and text summarisation work done by artificial intelligence (LLM). Rick On the Road. http://mandenews.blogspot.com/2023/08/evaluating-thematic-coding-and-text.html

King, J., McKegg, K., Oakden, J., & Wehipeihana, N. (2013). Evaluative rubrics: A method for surfacing values and improving the credibility of evaluation. Journal of MultiDisciplinary Evaluation, 9(21), 11–20.

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. Journal of Family Medicine and Primary Care, 4(3), 324–327. https://doi.org/10.4103/2249-4863.161306

Lynn, J., & Preskill, H. (2016). Rethinking Rigor. https://www.fsg.org/resource/rethinking-rigor/

**

Related

- chapter intro

References

Ashwin, Chhabra, & Rao (2023). *Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias*. https://doi.org/10.48550/arXiv.2309.17147.

Aston (2019). *Contribution Rubrics*. https://media.licdn.com/dms/document/media/C4D1FAQE1laRi0vrFrQ/feedshare-document-pdf-analyzed/0/1620553059516?e=1687996800&v=beta&t=bFr7dhpZ-slluV8ne1cERFelwINIaEzsQN8fiF_75gQ.

Davies (2023). *Evaluating Thematic Coding and Text Summarisation Work Done by Artificial Intelligence (LLM)*. http://mandenews.blogspot.com/2023/08/evaluating-thematic-coding-and-text.html.

King, McKegg, Oakden, & Wehipeihana (2013). *Evaluative Rubrics: A Method for Surfacing Values and Improving the Credibility of Evaluation*.