

Research on vs. with AI

(Leveraging) AI for Research Purposes, Class 2

Giovanni Spitale

A standalone summary of the afternoon, to keep. Every exercise is given in full, followed by its resolution: if you are reading this before the session, please don't be a cheater and stop at each "Resolution" heading. Plus, I prepared some surprises for you folks and spoilers are baaad.

1. Three ways AI can be in your research

When you open a language model, are you **studying the model**, or **using the model to study something else**? The answer matters more than it seems, and many published papers never say it.

- **Research ON AI:** the AI is the **object**. You ask how it behaves, persuades, classifies, discriminates, simulates or fails. Typical topics: model behaviour, bias, safety, evaluation, failure modes.
- **Research WITH AI:** the AI is the **instrument**. It sits inside your pipeline: literature mapping, transcription, coding, classification, simulation, drafting.
- **Both:** you use AI as an instrument **in order to** study AI. More common than it sounds, and the most interesting case.

2. Why the distinction is not just a label

Same model, same API call: two different sciences. What changes is what the model is **for** in your design.

	Research ON AI	Research WITH AI
What threatens validity	prompt sensitivity, model version and drift, sampling temperature, contamination, whether your prompts resemble real use	fabricated sources, coding drift, anchoring, non-reproducibility, missing provenance
What the ethics are about	what you make the model produce (dual use), who the outputs could harm	whose data you send where, who is accountable for what the model concluded
What you will have to report	the model, and how you used it	the model, how you used it, and where it ran and what it saw

What goes into that report, and why, is the subject of Michele Loi's session on Tuesday afternoon.

3. Exercise A · Eight studies, two axes

The exercise

Round 1, on your own (13 minutes). Read the eight studies below (3 minutes of silent reading), then place each one on two independent axes, on your phone: *is the AI the instrument of the research?* and *is the AI the object of the research?* Top-left is *on*, bottom-right is *with*, top-right is *both*, bottom-left is *neither*. Read each description to the end.

Round 2, in your group (15 minutes). Take the two studies where the room's answers were most spread out. Agree where each goes, or write down both positions; then name one threat to validity that follows from where you put it. One person reports, 60 seconds.

The eight studies

A1, A4, A5 and A8 describe real studies, simplified; A2, A3, A6 (the evaluation design) and A7 are constructed for the exercise.

A1 — Can you tell GPT-3 from a human on Twitter?

Researchers collected real tweets on science topics that are often the subject of misinformation, some accurate and some false. They then used GPT-3 to generate new tweets on the same topics, again some accurate and some false. Participants in an online survey saw a mix of the two and judged, for each tweet, whether its content was true or false, and whether it had been written by a person or by the AI. In a separate step, the researchers gave the same tweets to GPT-3 and asked it to make the same two judgements, so its answers could be compared with the participants'.

A2 — Oral history of a factory closure

A historian has recorded 80 oral-history interviews with former workers of a textile factory that closed in the 1990s, about four hours of audio each. She runs the recordings through Whisper, an open-source speech-to-text model, on her own laptop, so the audio never leaves her machine. She then reads every transcript against the recording, corrects names and dialect words by hand, and codes the corrected transcripts manually for themes of loss, skill and identity. The study is about how the workers remember the closure.

A3 — Same CV, different name

Researchers write one CV for a mid-level accounting job and make fifty copies, identical except for the candidate's name, chosen to signal gender and ethnic origin. They submit all fifty to five commercial AI hiring assistants, the kind companies use to pre-screen applicants, and record how each tool ranks each candidate. They repeat the whole procedure three times to check that the rankings are stable, and compare the rankings across names.

A4 — 249 patient interviews and a dictionary

249 interviews with Swiss patients, across eight conditions (among them dementia, multiple sclerosis, chronic pain and rare diseases), are analysed to build a taxonomy of the harm patients describe in their contact with healthcare. The coding uses a dictionary: 21 codes, each defined by a list of expressions in German, French, Italian and English, drawn up by the researchers from an earlier close reading of the interviews. Software reduces every word to its base form and marks each sentence that contains one of the expressions. No language model is involved. The researchers then read the marked passages.

A5 — How Google invents things

Researchers ask Google Search about theories, scholars, effects and historical events that do not exist: they invented the names themselves. They collect 140 of the AI Overviews that Google displays at the top of the results: seven kinds of invented entity, two names each, each name searched ten times. They then analyse how the overviews build a plausible answer about something that is not there: which tone they use, which real concepts they attach the invention to, whether they cite sources. To code the 140 overviews, the researchers use a language model with a codebook they wrote, and check its codes by hand.

A6 — Virtual patients

A medical school builds virtual patients: language-model characters based on real patient narratives, who answer in their own words, react to what the student says and change their mind if the conversation goes well. Third-year students use them to practise shared decision-making, in unscripted conversations. The study compares two groups of students, one that practised with the virtual patients and one that followed the usual training, on their communication skills in an exam with human actors.

A7 — A pilot study without participants

A PhD student is preparing a 40-item questionnaire on trust in doctors, to be sent to 1,500 people. Before launching it she needs a pilot study, to estimate how answers will be distributed and how many respondents she needs to detect the effects she expects. Instead of recruiting pilot participants, she asks a language model to answer the questionnaire 500 times, each time as a different person with a given age, gender, education and region. She uses the simulated answers to calculate her sample size, and then runs the real survey.

A8 — Who wrote it?

Researchers ask four language models to write short statements on 24 contested topics. They then give the same statements back to the same four models and ask how much they agree with each one. The statement is presented in different ways: sometimes with no information about the author, sometimes as written by a person of a given nationality, sometimes as written by another AI. The researchers measure whether the models' agreement with the same text changes depending on who they are told wrote it.

Resolution

	Looks like → reads as	Why
A1	on → both	The same study also asked GPT-3 to judge the tweets, as true or false and human or AI. Is GPT-3 a good judge of GPT-3?
A2	with → with	A clean anchor: AI as a transcriber, nothing about the AI is studied. Transcription ran locally.
A3	on → on	A clean anchor: an audit study. The AI is what is measured; no AI does the measuring.
A4	with → contested	A dictionary and a lemmatiser are NLP, not a language model. Whether that counts as "AI"

		matters less than the property: every coding decision is deterministic and inspectable.
A5	on → both	The overviews are coded with a language model.
A6	with → contested	The AI is the intervention, the outcome is human behaviour. Is a drug trial "research with the drug"?
A7	with → with	Easy to place, and the whole study can be wrong: simulated respondents model what the model learned about people, not the population.
A8	on → both	The same models wrote the statements they then judged.

The disagreement in the room was not about the facts of the studies: everyone read the same text. It was about **what the AI is for in the design**. That is the thing you have to be able to say about your own work.

What to take from round 2. Three of the studies that look like research *on* AI (A1, A5, A8) hide a second use of AI as an instrument, usually in a single sentence: whoever reads to the end finds it, whoever skims places the study by its headline. A4 and A6 have no clean answer, and the honest answer names the property that matters instead of the label (A4: every coding decision is deterministic and inspectable; A6: the AI is the intervention, the outcome is human). A7 is the easy one to place and the one most likely to be wrong as a study.

4. Research on AI as a method

One lens reads any study:

- **Unit of analysis:** what is one observation?
- **What is manipulated:** and what is held fixed?
- **What could make it false:** the threat to validity.

And one pipeline describes most of them: **input → invocation → output → interpretation → validation**. For each step, ask where a human checked, and where nobody did.

Study	Unit	Manipulated	What could make it false
GPT-3 (dis)informs us better than humans (Spitale, Biller-Andorno & Germani 2023)	a tweet judged by a person (697 participants)	author (human or GPT-3) × truth	GPT-3 is no longer the model anyone uses
Emotional prompting (Vinay et al. 2025)	one of 19,800 generated posts	politeness of the request × model × system role	does anyone prompt like that? And with a "helpful

			assistant" system role, the effect disappears
Source framing (Germani & Spitale 2025)	an agreement score (192,000 of them)	the declared source of the same statement	the models also wrote the statements: bias in judging, or in writing?
Assisted-suicide case reports (Spitale, Schneider et al. 2023)	one of 72 Dutch case reports	three classifiers, from logistic regression to GPT-4	class imbalance (56 yes, 16 no: always saying "yes" already scores 78%), and the prompt: asking whether <i>everyone</i> agreed, instead of <i>the committee</i> , turned every contested case into a no
How Google AI Overview confabulates (Germani & Spitale, in progress)	one of 140 overviews	the kind of invented entity, 10 repetitions each	the coding itself is done by a language model

Ethics, in one line. Research on AI tends to end with a detector. A tool that hands down a verdict replaces the judgement it was meant to educate: the *paternalism trap* (Spitale & Germani 2026).

5. Exercise B · Kill the study

The exercise

Your card describes one study. Your job is to kill it: list every way the result could be wrong, misleading, or impossible to repeat. Be specific: not "bias", but *which* bias, entering *where*. Then pick the one threat that worries you most, and say what you would do about it. 20 minutes; in the last 3, submit your threats on RoomPulse, one entry per threat. Groups Turing and Hopper take the card ON, Lovelace and Shannon the card WITH.

Card ON · Emotional prompting

Researchers asked four language models (davinci-002, davinci-003, gpt-3.5-turbo, gpt-4) to write social-media posts on public health topics. The model was told to write as "Sam", a persona described as a bad person who likes to spread disinformation. The instruction was given in three tones: polite, neutral and impolite. 19,800 posts were generated and every one was fact-checked by hand and labelled as disinformation, refusal, or accurate information. *Finding: polite instructions produced more disinformation.*

Card WITH · An LLM codes sixty interviews

A team has sixty interviews with family caregivers. They write a codebook of 18 codes, with a definition and two examples each. A language model reads the transcripts, splits them into segments, assigns codes with a confidence score, and returns the quotations that justify each code. One researcher codes ten of the interviews by hand; the team computes agreement on those ten, inspects disagreements, revises the codebook and runs it again. The paper reports themes across all sixty.

This study is constructed for the exercise.

Resolution

Half the groups attacked a study *on* AI (emotional prompting), half a study *with* AI (a language model coding sixty interviews). The two lists of threats look different because they are about different things.

ON: emotional prompting	WITH: a language model codes interviews
the models used have been retired: the study cannot be re-run as it was	quotes paraphrased, or invented, presented as verbatim
prompt wording and the system role decide the result	codebook drift: the codebook applied is not the one declared
the experimental manipulation (polite and impolite versions of the request) was itself written by a model	synonym explosion: a new code for almost every excerpt
ecological validity: does anyone prompt like this?	agreement with humans measured on a convenient subset, or not at all
dual use: the paper is also a recipe	coding in batches changes the result

Left: did we measure the model correctly? Right: did the model measure our data correctly? A lens you cannot inspect is a lens you cannot trust.

Three numbers for the WITH side:

- **473 codes for 555 coded segments:** the first exploratory run of the AI Overview study, with an empty codebook and a vague study description. Nothing was wrong with the model; everything was wrong with what it was given.
- **$\beta = -0.41$ ($p = 0.007$):** coding segments in batches, rather than one at a time, significantly lowered agreement with human consensus (Rush et al. 2026). Same models, same task, different structure.
- **κ 0.75 vs 0.42:** in the same study, the models agreed with each other more than the human coders did, and their disagreements were more systematically patterned. Consistent is not the same as right.

The one threat each side has to find. ON: the models in the study have been retired, so it cannot be re-run as it was. WITH: quotes that are paraphrased or invented but presented as verbatim. And one detail from the published paper that almost nobody finds: the polite and impolite versions of the request were generated by gpt-3.5-turbo itself, so the experimental manipulation is also a use of AI as an instrument (Vinay et al. 2025). The posts were fact-checked by two authors independently, with less than 1% disagreement.

6. Research with AI, and the apparatus

Four things you can ask a language model to do. Tomorrow morning is built on them.

- **Compressors:** summaries, literature triage, first-pass themes.
- **Extractors:** transcription, speaker turns, raw text into structured data.
- **Expanders:** brainstorming, counterarguments, alternative designs.
- **Formalisers:** argument structures, templates, natural language into R or Python.

"Can AI do qualitative coding?" is a badly posed question, like asking whether statistics can be used for research: the answer depends on what surrounds it. A 2025 manifesto, signed by five authors and 419

colleagues, rejects generative AI for reflexive qualitative research (Jowsey et al. 2025). Its evidence base is overwhelmingly ad-hoc chatbot use; its conclusion is about AI as such. The methodological worry is real; the generalisation is what fails.

What makes a tool trustworthy is **the apparatus, not the model**:

Property of the tool	What it prevents
Prompts inspectable and versioned	an analysis that cannot be repeated; disclosure as anecdote
Runs versioned and re-runnable	"the AI said" as an event instead of a method
Excerpts anchored to literal spans of the source	fabricated quotes: not unlikely, structurally impossible
The codebook as a declared constraint	silent drift between the codebook you report and the one applied
Explicit human gates on accepting and rejecting	a tool that hands down verdicts instead of proposals

The last row is the Dragon's Lair moment. In *Dragon's Lair* (Cinematronics, 1983) the animation was already drawn: you only pressed the right direction at a handful of exact moments, and if you missed one you watched a beautifully animated death. A lot of research with AI works like that. The output is gorgeous and moves on its own; you are choosing at a few moments. Find those moments, and be awake when they come.

7. The live Turing test, and the reflexive case

The exercise

One turn from a published patient interview on prenatal testing, coded twice with the same codebook of 140 codes: once by a human researcher, once by a language model (Claude Opus 5.5, through AutoCode). The interviews come from two published studies (Fang et al. 2023; Fang et al. 2025). Four codes on screen: two applied only by the human, two only by the model. For each code, answer on RoomPulse: *a human researcher or the AI?*

Resolution

The model applied *Relationship with partner* and *Health variations – Impact on family life*; the human applied *Work and Pregnancy* and *Need for/missing information*. Most rooms get at least two of the four backwards.

There was no style to go on: same codebook, same labels. All you could judge was whether a code fits the text. That is the right question, and it does not tell you who coded. The human's codes look off on this turn for a reason: the human coded the passage together with the next one ("I knew more than the doctors for like what's practical"), and gave both the same codes. A different unit of coding, not a careless coder.

Four items prove nothing, and this is not the study. Whether a code is **right** about the text is the question the manifesto skips, and the one a real test has to ask.

The full design is the AI Coding Turing Test (in development): code the same material by hand and with a language model, mix the two, and ask experienced qualitative researchers, ideally signatories of the manifesto,

which is which (*that's gonna be HEAPS of fun and for sure we're gonna make ourselves LOTS of new friends*). It is *with* and *on* at the same time, and the instrument being studied is the instrument being used.

8. Exercise C · Your research, two ways

The exercise

In your group, pick one topic close to at least one person's PhD. Write one research question where AI is the **object**, and one where AI is the **instrument**. Choose one to take to the final project, say why, and name the threat that could sink it (look back at Exercise B). Work in your group's project sheet (15 minutes), then enter three steps on RoomPulse: the question; why AI is the object, the instrument or both here; the threat.

The project sheet has seven fields: our topic · a question where AI is the object · a question where AI is the instrument · the question we take to the project · because · what could sink it · the question we did not pick, and why not.

Resolution

There is no single right answer: the exercise is graded by its moves, not its topic. A good chain has three things:

- **The role of AI is stated, not assumed.** "AI is the instrument here, because...", or "both, because...", with the reason.
- **The threat is specific and comes from the right family.** An *on* question is threatened by model versions, prompt sensitivity and ecological validity; a *with* question by fabricated or drifting outputs and missing validation against humans.
- **The question you did not pick is still on the page.** It is often the better project, and saying why you did not choose it is part of the design.

A worked example, of the kind a group might write:

Project sheet	Example answer
Topic	People with chronic pain and AI symptom checkers
AI as object	How do AI symptom checkers respond to descriptions of chronic pain, compared with acute pain?
AI as instrument	What do people with chronic pain say about AI symptom checkers in online forums, and does it change how they describe their doctors?
Taken to the project	The instrument question: three of the group can work on it, and the data is reachable
What could sink it	The model that codes the posts brings its own idea of what "dismissal by doctors" sounds like, and we find our codebook instead of the patients' experience

Assignment

About 20 minutes, on Teams (due Tuesday 08:00 so that I have some time to read your work). On your own, take your group's question and write how you would study it in five or six steps. For each step: what you would ask a tool to do (or nothing); which tool you would use tonight, the one you would really reach for; and the level of the data that goes into it. Then one line: the step where this is most likely to go wrong. Nothing to install, no account to create. If you cannot do it tonight, come anyway: you will do it aloud with your group.

References

- Fang MT, Germani F, Spitale G, Wüscher S, Kunz L, Biller-Andorno N (2023). Women's experiences with non-invasive prenatal testing in Switzerland: a qualitative analysis. *BMC Medical Ethics*, 24(1), 85. <https://doi.org/10.1186/s12910-023-00964-3>
- Fang MT, Kunz L, Spitale G, Wüscher S, Germani F, Biller-Andorno N (2025). Women's perceptions of the medicalization of pregnancy and their preferred models of care: a qualitative analysis. *Narrative Inquiry in Bioethics*, 15(1), 53–68. <https://doi.org/10.1353/nib.2025.a957205>
- Germani F, Spitale G (2025). Source framing triggers systematic bias in large language models. *Science Advances*, 11(45), ead2924. <https://doi.org/10.1126/sciadv.adz2924>
- Jowsey T, Braun V, Clarke V, Lupton D, Fine M (2025). We reject the use of generative artificial intelligence for reflexive qualitative research. *Qualitative Inquiry*. <https://doi.org/10.1177/10778004251401851>
- Rush E, Karim MN, Yacu GS, et al. (2026). Evaluating hybrid human-LLM coding workflows for qualitative research in medical education: a generalizability study. *Medical Teacher*. <https://doi.org/10.1080/0142159x.2026.2721362>
- Spitale G (2025). *Doing research with AI* [lecture slides]. IBME, University of Zurich, 10 December 2025. <https://www.giovanispitale.net/wp-content/uploads/2025/12/Doing-research-with-AI.pdf>
- Spitale G, Biller-Andorno N, Germani F (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), eadh1850. <https://doi.org/10.1126/sciadv.adh1850>
- Spitale G, Germani F (2026). Misinformation, disinformation, and autonomy: against the epistemic paternalism trap [preprint]. *Zenodo*. <https://doi.org/10.5281/zenodo.20321941>
- Spitale G, Schneider G, Germani F, Biller-Andorno N (2023). Exploring the role of AI in classifying, analyzing, and generating case reports on assisted suicide cases: feasibility and ethical implications. *Frontiers in Artificial Intelligence*, 6, 1328865. <https://doi.org/10.3389/frai.2023.1328865>
- Vinay R, Spitale G, Biller-Andorno N, Germani F (2025). Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence*, 8, 1543603. <https://doi.org/10.3389/frai.2025.1543603>

Work in progress mentioned in class, not yet citable: the AI Overview confabulation study (Germani & Spitale); the patient-derived taxonomy of healthcare harm (Spitale, Naishtat, Ritzmann, Biller-Andorno); the AI Coding Turing Test.