

AI Tools for Research

(Leveraging) AI for Research Purposes, Class 3

Giovanni Spitale

A standalone summary of the morning, 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. Products change. What you ask them to do doesn't.

The list of tools changes every few weeks. In December 2025 the general-purpose assistants were ChatGPT, Microsoft Copilot, Google Gemini and Anthropic Claude. By 29 September 2026 the UZH roadmap listed Copilot Studio (pilot, 11/2026), ChatGPT (11/2026, slipped from 09/2026) and an LLM Proxy (12/2026). What you ask these tools to do has not changed since the first chatbot. There are four things, and each comes with its own worry:

Use	Examples	The worry
Compress	summaries, literature triage, first-pass themes	what got left out, and you will never know
Extract	transcription, structured data from text	silent errors in data you will treat as clean
Expand	brainstorming, counterarguments, alternative designs	plausibility mistaken for evidence
Formalise	structures, templates, words into code	code that runs and is wrong

Most people spend most of their AI time on writing and summarising. Extract and formalise are the other half, and the half where research workflows gain the most.

2. Five families of tools

Family	Examples	What it is for
General-purpose assistants	ChatGPT, Claude, Gemini, Copilot	80% of anyone's AI time
Research-focused assistants	NotebookLM, Elicit, SciSpace	grounded in sources you give it, or that it finds
Local, privacy-first	LM Studio, Ollama, GPT4All, noScribe	anything above "internal"
Pipelines via code and API	Python and an API, notebooks	the same operation on 20,000

		items
Interoperability	MCP	tools that talk to each other, and to a model, without copy-paste (section 5)

3. Your data has a level. Every tool has a ceiling.

The decision is a match between the two. UZH classifies information on four levels: **öffentlich** (public), **intern** (internal), **vertraulich** (confidential) and **geheim** (secret). Three rules do most of the work:

- Unmarked means at least internal.
- Unpublished means at least internal, your own manuscript included.
- One sensitive item pulls its whole folder up.

Two sentences to remember: **free AI tools are for public data only. For secret data there is no AI tool at all.**

The doors: where the UZH tools are, and how to get in (state of 29 September 2026; the one link worth saving is the ZI overview page, because it changes):

Tool	Data it may receive	How to get in
M365 Copilot Chat (Basic)	public, internal	m365.cloud.microsoft/chat, UZH login; the shield icon means the protected version
M365 Copilot (Premium)	up to confidential, web search off	licensed per person, through your unit
DeepL Pro Advanced	up to confidential	UZH licence
GitHub Copilot	public, internal	UZH licence; Merkblatt on the ZI page
LM Studio + Mistral	up to confidential	installed on your device; nothing leaves it
ScienceCloud / ScienceCluster	up to confidential / internal	Central IT, for research workloads
Free AI tools	public only	—

Check yourself

Your unpublished manuscript: you want feedback on its structure. Which of these are you allowed to use? Free ChatGPT · M365 Copilot Chat (Basic) · M365 Copilot (Premium) · LM Studio on your device.

Answer. Copilot Basic, Copilot Premium and LM Studio. An unpublished manuscript is internal; all three reach internal. Free ChatGPT stops at public.

4. Exercise D · Tool triage

The exercise

On your own (3 minutes): for each of the six tasks below, what are you asking the tool to do? Compress, extract, expand, formalise, or search (one or more).

In your group (13 minutes): for each task, fill in: **use → tool family → one concrete tool → the main risk → the level of the data, and whether your tool is allowed there.**

D1. You are starting a new paper on a question nobody in your group has worked on. You want to know what has been published in the last five years. *Data: public.*

D2. You are doing a meta-analysis and need sample sizes, effect sizes and confidence intervals from 40 published PDFs. *Data: public.*

D3. You have 200 excerpts from interviews you conducted for your own study, with consent for research use. You want a first pass of codes before coding by hand. *Data: confidential; secret if the interviews were part of a clinical study.*

D4. Your survey platform exports 3,000 rows of pseudonymised answers, with inconsistent date formats, free text in numeric fields and three different spellings of every canton. You need a clean dataset and you are not good at R. *Data: confidential.*

D5. You have 30 hours of recorded interviews with nurses about their working conditions, for a study on staff shortages. You need transcripts. *Data: confidential.*

D6. Your manuscript is ready to submit and has 74 references, some added in a hurry. You want to be sure every one exists and is cited correctly. *Data: the references are public; the manuscript is internal.*

Resolution

	Use	A tool that fits	Main risk	Data level · allowed?
D1	search + compress	a research assistant with search (Elicit, Semantic Scholar) or Copilot Chat	fabricated or misattributed references: nothing counts until checked	public: any tool, free ones included
D2	extract	a research assistant (Elicit) or a pipeline via API	silent extraction errors in a table you will trust: check a sample by hand	public: any tool
D3	extract + compress	a local model (LM Studio + Mistral) or Copilot Premium, web search off	the obvious cloud tool is not allowed, AutoCode included	confidential: local or Premium only; secret if clinical: no AI tool
D4	formalise	Copilot Chat or GitHub Copilot, given only the schema	sending the data itself; code that runs and is wrong	the data is confidential, the schema is not: column names and five made-up rows, code run locally
D5	extract	a local transcriber: noScribe, or Whisper on	errors in names and jargon: read the transcript	confidential: on-device is

		your machine	against the audio	allowed, the cloud is not
D6	extract + search	doi.org, Crossref, a reference manager (Zotero)	the checker can hallucinate too; DOIs resolve or they don't	references public, draft internal: paste the list, not the manuscript

The tools are examples that fit, not the only right answer. The data-level column is the one that is not negotiable.

Three of the six have a trap worth remembering:

- **D3, your own interviews.** The obvious tool is a cloud model, and your data is confidential or secret. Not even AutoCode, which I built for exactly this task, is allowed: it sends text to a cloud model, on a server that is not UZH's. The rule applies to my tools too. What remains: a local model (LM Studio), or coding by hand with the AI kept away from the raw data.
- **D4, send the schema, not the data.** The model does not need your data to write code for your data. It needs the shape: the column names and five rows you made up. Take the code it writes and run it where the data lives. This is the most useful technique of the morning.
- **D5, the safe tool is also free.** The interviews are confidential, so the cloud is out, but a local transcriber (noScribe, or Whisper on your own machine) does the job without the audio ever leaving it. Not every ceiling is a loss. (Had they been interviews with patients in a medical study, the data would be secret, and there would be no AI tool at all, local ones included.)

For **D6**: DOIs resolve or they do not, so check them against Crossref or doi.org, not against another chatbot. Paste the reference list, not the manuscript.

5. Tools at work

Four stops, and at every stop the same three questions: **which use? At what level does the data flow? What makes it trustworthy, or not?** The third question is the apparatus table from Class 2, applied.

Stop 1 · A general assistant against a grounded one (use: compress)

On 9 October 2026, two questions went to a free chatbot with web search off (DeepSeek) and to NotebookLM loaded with three papers you met in Class 2 (Spitale et al. 2023; Vinay et al. 2025; Germani & Spitale 2025):

1. *Is AI-generated disinformation more convincing than human-written disinformation, and does the way you ask a model change how much it produces? Give the numbers and the sources.* 2. *Does this also hold for German-language social media?*

	Free chatbot (DeepSeek, search off)	NotebookLM (three papers)
Question 1	Cited the GPT-3 study in the wrong journal (<i>BMJ Open</i> 2024), with a DOI that resolves to nothing. Never found the emotional-prompting study; answered with jailbreak papers instead. Two more references in	The right journal, and every number checks against the papers: recognition 0.92 vs 0.89 (P = 0.0032); compliance 77% → 94% → 28% depending on the tone of the request.

	the wrong journals.	
Question 2	Invented a German literature: replications at LMU, TU Munich and Zurich, a German special issue, Fraunhofer reports, effect sizes of 0.2–0.5.	<i>"Your sources do not directly evaluate German-language social media."</i> Then it quoted the paper's own limitation: English-speaking participants.

Grounding is a property of the tool, not of the model. A grounded tool can say *"not in my sources"*. A free one never runs out of answers.

MCP · what it is, and why a researcher should care

Until recently, connecting a model to a tool meant custom glue for every pair, or you copying and pasting between windows. **MCP (Model Context Protocol)** is one standard plug: any tool written for it works with any application that speaks it. Anthropic introduced it on 25 November 2024; it has since been adopted by ChatGPT, Gemini, Microsoft Copilot, Visual Studio Code and others, and since December 2025 it belongs to the Agentic AI Foundation, under the Linux Foundation.

How it works: **you** talk to a **host** (the application, with the model), which connects to **servers** (the tools: search, coding, references). The model never sees a tool's code, only a description of what it does, and decides from that description when to call it. The host can ask you to confirm before anything happens. A *tool* is something the model decides to use; a *resource* is something you put on the table: the protocol encodes who has the initiative.

Why it matters:

- **Workflows stop being copy-paste.** The search tool, the coding tool and the reference manager become steps in one flow, with you at the joints instead of at the clipboard.
- **A connector is a door to your data.** Whatever a server reads goes where the model runs. If it reads your interview folder, your interview folder travels. Level and ceiling apply to connectors too.
- **The description is a prompt.** The model trusts what a tool says about itself. Connect servers you would trust with the data, not every server that exists.

Is it safe? As safe as the servers you connect and the data you let them read: the same question as the bootcamp.

Stop 2 · Contrarian (use: search + compress)

A claim is asked in a chat; the model calls Contrarian's tools through MCP, searching for the evidence **for and against** the claim, and returns a graded verdict; the public page (contrarian.borant.eu) shows what it did. The opposite claim is searched by default. The method is public, the trace is not. Contrarian, ArguMap and LSSR are tools I built: they are here not because they are better than commercial ones, but because I can open them up and show you the inside.

Stop 3 · ArguMap (use: formalise)

The paternalism-trap preprint (Spitale & Germani 2026), mapped as an argument: the claim, the premises, which step supports, attacks or qualifies which, and where the argument is weakest. The preprint is public, so any tool is allowed. The model proposes the map; whether it is the argument is the authors' call.

Stop 4 · LSSR, a living scoping review (use: the whole pipeline)

A review in progress on the move from opt-in to opt-out in organ donation: query, screening and extraction, with the model's votes next to the humans'. AI that votes but does not decide.

6. Exercise E · Grade the output

The exercise

Three AI tools were given the same question: *what is the typical level of agreement between human coders and large language models in qualitative coding? Give the values reported in the literature, the study each comes from, and full references with DOIs.*

	Generated by	Stands for
Output 1	DeepSeek, free, web search off	a free chatbot answering from memory
Output 2	Claude Sonnet, web search on	a chatbot with search
Output 3	Claude Sonnet, allowed to read only five open-access papers	a grounded notebook tool

All three were generated on 7 October 2026 and are printed verbatim in the **output pack** (on the course page). Each group gets a sheet of seven items, three from Output 1, two from Output 2 and two from Output 3, and codes each claim with its source:

Code	Meaning
A	the source does not exist (an identifier pointing to an unrelated paper counts)
B	it exists, but the reference is wrong
C	it exists, but does not say what it is cited for
D	it exists and supports the claim

One person per output. Check with doi.org, Crossref, PubMed, Europe PMC, arXiv and Google Scholar, all public and with no account. For C and D, **open the paper and find the number**: a title that looks right is not a check. 27 minutes, then 4 to agree on your seven codes and your most convincing error, which one person reports aloud.

Resolution

	References	Claims	Where it fails
Output 1 · no search	7 of 12 do not exist	0 of 12 supported, even where the reference is perfect	everywhere
Output 2 · web search	all correct	3 of 6 wrong: twice, a model agreeing with itself presented as	the claims

		agreement with humans	
Output 3 · five papers	all correct	29 of 32 correct	the synthesis: generalisations its five papers do not support

Group Turing

	Item	Code	What's going on
1	O1 · Xiao, Zhang & Xu 2023	A	The arXiv id leads to another paper ("Generative AI at Work"). No such study.
2	O1 · Gilardi et al. 2023	C	Perfect reference. But the benchmark is trained annotators, not experts, and the ranges are not in the paper.
3	O1 · Zhang, Zhang, Zhang & Zhang 2024	A	".12345" is a placeholder: the id is a physics paper.
4	O2 · Bijker et al. 2024	C	Numbers verbatim, but they are ChatGPT agreeing with itself (ten runs), not with humans.
5	O2 · Borse et al. 2025	D	In the abstract, verbatim.
6	O3 · summary: "70–99%", "consistently high"	C	Only two of the five papers report human–LLM percent agreement. The range is the answer's.
7	O3 · Balt et al. 2025	D	Results, verbatim: accuracy 0.84 (SD 0.09), κ 0.68.

Group Lovelace

	Item	Code	What's going on
1	O1 · Chew et al. 2023	C	Real paper, but GPT-3.5 only, and Gwet's AC1: no GPT-4, no kappa.
2	O1 · Zhu et al. 2023	A	The id is a French image-processing paper. No such study.
3	O1 · Meng, Zhang, Zhang & Yang 2023	A	Placeholder id: a black-hole physics paper.
4	O2 · Multi-LLM, arXiv:2512.20352	C	Agreement between runs of the same model. The paper: "we cannot claim ... human-AI agreement".
5	O2 · Sakaguchi et al. 2025	D	In the abstract, verbatim.
6	O3 · van den Berkmortel: "quotation unitization"	C	Not in the paper. The authors stress interpretive divergence; the answer reverses their emphasis.
7	O3 · Hill: 93.5%, κ 0.34, AC1 0.93	D	Table 1. The paper's own abstract says 92.5: the source disagrees with itself.

Group Hopper

	Item	Code	What's going on
1	O1 · Törnberg 2023	C	Real paper: one binary task, Krippendorff's α , no kappa.
2	O1 · Deng et al. 2023	A	The id is a mathematics paper ("On uniquely packable trees").
3	O1 · Sanders et al. 2024	C	The title exists, but by Ajunwa, Cardozo Law Review 2020: a law article, no LLMs, no coding.
4	O2 · Liu et al. 2025: κ 0.72–0.95	C	Holds for Study 2 only. In Study 1, zero-shot, it falls below 0.2.
5	O2 · Bijker et al. 2024	C	Numbers verbatim, but they are ChatGPT agreeing with itself (ten runs), not with humans.
6	O3 · Hill: AC1 "because" κ is attenuated	C	No such reason in the paper: AC1 was one of five pre-planned metrics.
7	O3 · summary: κ "moderate to substantial (0.34–0.68)"	C	On the scale Balt itself uses, 0.34 is "fair". Per-code values go down to –0.05.

Group Shannon

	Item	Code	What's going on
1	O1 · Bano, Zowghi, Zhang & Zhang 2024	C	The real paper has other co-authors and is a vision paper: no data, no kappa.
2	O1 · Padiyath et al. 2024	A	Placeholder id (a Floer homology paper). Real researcher, the paper does not exist.
3	O1 · Gebreegziabher et al. 2024, CHI	A	Right CHI prefix, resolves to a real paper about drones. The best fabrication in the pack.
4	O2 · Multi-LLM, arXiv:2512.20352	C	Agreement between runs of the same model. The paper: "we cannot claim ... human-AI agreement".
5	O2 · Liu et al. 2025: κ 0.72–0.95	C	Holds for Study 2 only. In Study 1, zero-shot, it falls below 0.2.
6	O3 · Hill: per-code κ from –0.05 to 0.54	D?	The text says exactly this, but Table 2 has 0.78. The paper contradicts itself: unease is the right answer.
7	O3 · Leas: "9331/10,000" in the results text	C	The results text says 9931. 9331 is a typo in Table 2.

The deviation that matters. Where you coded D but the key says C, you checked the number and not what it measures. Bijker and the Multi-LLM paper are the clearest cases: the numbers are in the papers, but they measure a model agreeing with itself.

It told you, at the bottom. The last line of Output 1: *"Several DOIs and arXiv identifiers above are illustrative/approximate based on the cited works' typical locations."* Twelve confident paragraphs, then this.

The same question, one switch. The same chatbot (DeepSeek), the same prompt, the same day, with web search off and on: 12 references from 2023–2024, 7 of them nonexistent, against 6 references from 2025–2026, all real but none with authors or DOI. Not one reference appears in both lists. One toggle and you get a different literature, and nothing in the answer tells you which one you got. Which version would have ended up in your paper?

The error moved.

- A reference that exists is not a reference that supports the claim.
- The grounded answer got its facts right and its synthesis wrong: grounding removes invention, it does not stop a model generalising beyond its sources. And somebody chose the sources.
- From the reference, to the claim, to the summary: your check has to move with it.

7. Why it matters

Two real documents from peer review, shown without comment before the break. A reviewer report that, halfway through a sentence, stops to say it might be confusing the manuscript with *"the previous manuscript about solar energy"*. And an author's response to reviewers that kept the chatbot's line *"The modified version in the canvas reflects all these changes"*. This is what happens downstream when nobody checks, on both sides of peer review.

8. Exercise F · The death scenes

The exercise

Take your group's question from Monday and put last night's individual versions side by side: where did you disagree? (4 minutes; whoever did not do the assignment says theirs aloud.)

First, two minutes on the **Our project** slide of the canvas: your question, whether AI is the object or the instrument, the data and its level, the threat. Then build **the worst workflow you can make plausible** for your question, starting from the tools you wrote down last night. The canvas has six phases (study design, literature, data collection, analysis, writing, dissemination) and, for each, the rows: **step · use · tool · data level · allowed? · human gate · log needed? (yes/no)**. The last row is **what this breaks, and what it costs**: every failure needs it. The writing column stays grey: writing is the subject of the afternoon session.

On the rescue slide of the canvas, **the rescue**: the two or three worst failures, and for each one *the change* (tool, data level, human gate: what you do instead) and *what the change costs* (time, model quality, feasibility). 22 minutes. On Wednesday you present the disaster first, then the rescue.

Resolution

There is no single right disaster. A good one has three properties:

- **Plausible.** Every failure is something researchers really do. *"The AI writes the thesis and we send it to Nature"* breaks no specific row, and nobody would do it.
- **Labelled.** Each failure says which row it breaks (data level, a tool where it is not allowed, a missing human gate, an ignored threat) and what it costs (a data breach, results that cannot be reproduced, references that do not support the claims).

- **Built from your own habits.** The most convincing disaster starts from the tool you would really reach for tonight.

A good rescue fixes the worst failures, not all of them: tools chosen for a reason, data kept at its level, a human gate wherever the disaster had none, and an honest word on what the fix costs.

A worked example, continuing the question from the Class 2 handout (*what do people with chronic pain say about AI symptom checkers in online forums?*):

Phase	The disaster	What it breaks, and what it costs
Literature	a free chatbot lists 20 "key studies"; they go straight into the introduction	references never checked: some will not exist (Exercise E, Output 1)
Data collection	5,000 forum posts scraped with usernames and pasted into a free chatbot	public in origin, but personal health data once collected: the data level is wrong for the tool, and the usernames make people findable
Analysis	the model codes all posts in one batch with no codebook, and its themes become the results	synonym explosion and batching effects; no human gate, so the model's idea of "dismissal by doctors" becomes the finding
Dissemination	the paper quotes posts verbatim	a verbatim quote from a public forum can be searched back to its author

And its rescue:

The failure	The change	What the change costs
unchecked references	every reference checked against Crossref or doi.org; claims checked in the paper	a day of work
health data in a free tool	usernames removed at collection; coding with a local model (LM Studio) or an allowed UZH tool	a weaker model, and setup time
batch coding, no codebook	a declared codebook, posts coded one by one, a human coding a sample and comparing	slower, and the agreement may disappoint

...

References

- Anthropic (2024). Introducing the Model Context Protocol, 25 November 2024.
<https://www.anthropic.com/news/model-context-protocol>

- Anthropic (2025). Donating the Model Context Protocol and establishing the Agentic AI Foundation, 9 December 2025. <https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation>
- Balt E, Salmi S, Bhulai S, et al. (2025). Deductively coding psychosocial autopsy interview data using a few-shot learning large language model. *Frontiers in Public Health*, 13, 1512537. <https://doi.org/10.3389/fpubh.2025.1512537>
- 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.ad2924>
- Gilardi F, Alizadeh M, Kubli M (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *PNAS*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Hill C, Dahil A, Simpson G, et al. (2026). Large language models for thematic analysis in healthcare research: a blinded mixed-methods comparison with human analysts. *PLOS Digital Health*, 5(4), e0001189. <https://doi.org/10.1371/journal.pdig.0001189>
- Leas EC, Ayers JW, Desai N, et al. (2024). Using large language models to support content analysis: a case study of ChatGPT for adverse event detection. *JMIR*, 26, e52499. <https://doi.org/10.2196/52499>
- Spitale G (2025). *Doing research with AI* [lecture slides]. IBME, University of Zurich, 10 December 2025. <https://www.giovannispitale.net/wp-content/uploads/2025/12/Doing-research-with-AI.pdf>
- Spitale G (2026). *AI governance* [bootcamp slides]. DSI Excellence Bootcamp, University of Zurich, 3 September 2026. <https://www.giovannispitale.net/ai-governance-dsi-excellence-bootcamp/>
- 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>
- UZH (2021). Weisung zur Klassifizierung von Informationen (in force since 1 September 2023). rud.uzh.ch
- UZH Zentrale Informatik (2026). Übersicht KI-Tools und Services, version of 29 September 2026. zi.uzh.ch
- van den Berkmortel B, Westra D, Gifford R, van de Baan FC (2026). Analyzing nurse sentiment in a Dutch regional newspaper during the COVID-19 pandemic: comparative content analysis of GPT-5, Gemini 2.5 Pro, and human coding. *JMIR*, 28, e98374. <https://doi.org/10.2196/98374>
- 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>

Exercise E: the three outputs are in the output pack; five of the papers above (Balt, Gilardi, Hill, Leas, van den Berkmortel) are the sources Output 3 was restricted to.