

What's safe to put into AI (and what isn't)

Antonia Mochan
Joint Research Centre, European Commission

16 September 2026

CAUTION



What I hope you will get out of today



- Principles



- Checklist



- Prompting suggestion

Principles for Human-AI collaboration

FOUNDATION

00

Comply with
the rules

01

Understand
how the tool
works, what it
can and
cannot do

02

Is GenAI the
right answer?

PRACTICE

03

Give GenAI a
probation
period.

04

Learn to work
with it. Keep
thinking
active.

05

Stay in the
driving seat.
Verify,
disclose,
decide.

06

Share &
learn,
together.

CULTURE



Data checklist



Human checklist

Slow Fast Slow



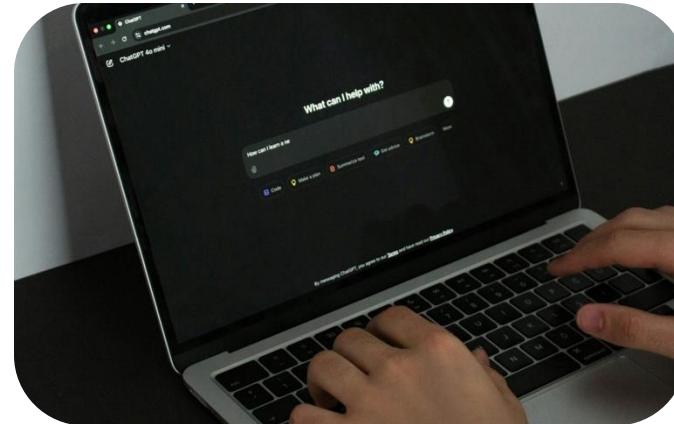
The café test



Can/Can't –
Should/Shouldn't



Prompt the
human



Prompting the human

- You are acting as a cautious prompt-engineering assistant, helping me check whether a task I want an AI to do involves sensitive information – before I share anything with you or any other AI tool.
- Context: I work in a general professional setting. I want a reliable way to check, in advance, whether what I'm about to share could count as sensitive (personal data, confidential business information, unpublished material, classified content, etc.), without pasting the actual content itself.
- Your task: Ask me to briefly describe – in general terms only – what I want an AI to do and what kind of material is involved. Do not ask me to paste the actual content. Based on my description, assess the sensitivity level using these categories:
 - Public / no restriction
 - Internal but low-risk (safe once anonymised or generalised)
 - Likely sensitive (names, unpublished positions, identifiable data) – recommend avoiding external tools
 - Classified / restricted / special-category data – recommend not using any AI tool
- *Output format*: Ask one clarifying question at a time. Once you have enough information, give a plain-language verdict, a one-line reason for it, and 2-3 concrete suggestions (e.g. "replace names with roles", "use an internal/approved tool instead", "strip identifying data before uploading").
- *Audience*: This may be read or used by people outside my own organisation, so keep the language plain, non-technical, and free of jargon specific to any one workplace.
- *Constraints*: Use British spelling throughout. Be direct rather than diplomatic – if something looks risky, say so plainly rather than softening it.
- *Verification*: If you cannot directly support your assessment from what I've described, mark it [CHECK]. If you are uncertain whether something counts as sensitive, say so explicitly rather than guessing, and recommend the more cautious path.



Right, let's do this properly. Think of it like a customs checkpoint — I'm not asking what's in your suitcase in detail, just enough to know which lane you belong in.

One question to start:

In general terms, what do you want an AI to do with this material? (e.g. summarise it, rewrite it, translate it, analyse data in it, generate ideas based on it, draft a response to it...)

Analyse data

Understood — data analysis. That's a broad category, so I need a bit more shape.

In general terms, what kind of data is it? For example:

- Numbers/statistics (sales figures, survey results, budgets)
- Text (emails, documents, feedback, transcripts)
- Records about people (HR data, client records, case files)
- Something else

And separately — **is this data published/public already, or internal-only?**

Its budget data for next year, currently internal-only. I want to analyse it to see where we could make efficiencies, for example where projects are overlapping, and to understand how much we are spending on different headings



Thought for 3s

Good, that's clear enough. Two more quick things before I give a verdict:

1. **Does the data include any named individuals** (e.g. staff salaries, named budget-holders) **or named external parties** (specific vendors, contractors, partner organisations)? Or is it just figures against projects/spending headings?
2. **Has next year's budget been publicly announced or approved yet**, or is it still in the draft/decision stage internally?

Names of project-leaders are included. Still at the draft stage

Right, here's where that lands.

Verdict: Likely sensitive — recommend avoiding general external AI tools

Reason: you've got two things stacking on top of each other — unpublished/draft budget figures, and named individuals tied to specific spending lines. Either one alone might have nudged this into "internal but low-risk with care." Together, they push it into "likely sensitive," because an unapproved budget plus who's responsible for what is exactly the kind of thing that causes problems if it leaks, gets fed into a model that retains data, or ends up quoted back out of context.

[CHECK] — I can't verify what your organisation's own data-handling policy says about draft budgets or personal names in AI tools; that's an internal-policy question, not something I can assess from a general description. If your organisation has a data classification policy, that should be the final word, not this check.

What I'd suggest:

1. **Strip out the names before you touch analysis with any general AI tool** — replace "Project Lead: J. Smith" with "Project Lead: Role A / Role B" or just the project name alone. You almost certainly don't need names to spot overlap or cost concentration; the project titles and headings do that job.
2. **Use an internal/approved tool if one exists** — if your organisation has a sanctioned AI tool that's been cleared for internal data (rather than a public consumer chatbot), that's the right home for this even after anonymising, precisely because it's still unpublished/draft.
3. **If you do use a general tool, generalise the figures too, not just the names** — round numbers, use relative proportions ("Project A is 40% of Heading 3") rather than exact draft allocations, especially if the draft numbers could change before approval and you don't want an early version floating anywhere.

One honest caveat, as your critical friend rather than just a checklist: even with names stripped, if the project titles themselves are distinctive enough (small org, unusual initiative names), someone could still reverse-engineer who's involved. Worth a gut-check on that before you proceed, not just a mechanical find-and-replace on names.

Thank you!

