

Homework — Safety, privacy, cost, and limitations

Estimated time: ~45 min · Tools needed: access to any AI chat (Claude, ChatGPT, etc.) and optionally a search engine · No code required

The seven risks covered in this lesson aren't abstract. They show up in real use. These three tasks give you direct experience with the three most common failure modes — hallucinations, privacy slip-ups, and outdated information — so they become instincts rather than things you have to remember.

Before you start

Have the Theory page open in another tab, particularly the Pre-flight Checklist and the Privacy Risk Grader. You may want to refer back to them during Task 3. Also open Claude.ai, ChatGPT, or any AI chat you use regularly.

Task 1 — Hallucination hunting

This task builds the "verify by default" habit. You're going to deliberately ask AI about things that are easy to get wrong, then check the answers against real sources.

Step A: Choose a domain you know something about. It could be your profession, a hobby, a historical period you've studied, a technical field, or even the city you grew up in. Pick something where you have enough background to recognise a wrong answer.

Step B: Ask the AI five specific factual questions in that domain. Make them specific — not "what is machine learning?" but "What paper first introduced the transformer architecture, who were the authors, and in what year?" Specific claims are where hallucinations hide. Write the AI's answers down (or copy them).

Step C: Verify each answer against an independent source. Use a search engine, Wikipedia, a textbook, or an official source — not another AI. Mark each answer: correct, partially correct, wrong, or unverifiable.

Step D: For any incorrect or suspicious answer, try again with the prompt *"If you're not sure about any part of this, please say so."* Compare the second set of answers.

Reflect and write down:

- Which domain did you choose? Why did you pick it?
- How many of the five answers were fully correct? How many had errors? Were any fabricated?
- Did adding "if you're not sure, say so" change anything? Did the AI flag more uncertainty the second time?
- What type of claim was most likely to be wrong — dates, names, numbers, qualitative descriptions?
- What habit will you apply to your daily AI use based on what you found?

✓ **Self-check: You should have verified five specific AI claims against real sources and identified at least one error or instance of hallucination. If all five were perfect, try a more obscure or niche domain — hallucinations are much more common outside popular knowledge.**

Task 2 — Privacy audit of your own AI use

This task asks you to look back at how you've been using AI and audit it against the privacy rules from this lesson. It's not about catching yourself doing something wrong — it's about building a conscious habit going forward.

Step A: Think back over the last 2–4 weeks of AI use (or longer, if you're new). What kinds of information have you pasted into AI chats? Go back through your actual chat history if you can. Categorise what you've shared:

- Public or general knowledge
- Personal but non-sensitive (your name, city, job title)
- Sensitive personal data (health details, finances, relationships)

- Work or company information
- Other people's information
- Credentials or tokens (hopefully none)

Step B: For each category you've used, use the Privacy Risk Grader in the Theory page to assess whether your tool choice matched the data sensitivity. Were there any mismatches?

Step C: Pick one real task you regularly use AI for — drafting emails, summarising documents, generating code, answering questions — and write a short "privacy rule" for that task. For example: "When I use AI to help draft client emails, I will not include client names, account numbers, or contract details. I will replace them with placeholders like [CLIENT] and [AMOUNT]."

Reflect and write down:

- What data categories have you shared with AI tools? Were any of them higher-risk than you'd realised?
- Were there any mismatches between the sensitivity of your data and the tool you were using?
- Write your personal privacy rule for at least one regular AI task.
- If you handle professionally sensitive data (health, legal, financial, company secrets), do you have a local model set up for those tasks? If not, what would you need to do to set one up?

✓ **Self-check: You should have at least one written privacy rule specific to your own AI use. If your audit turned up no concerns at all, either your use has been very conservative (good) or you haven't looked closely enough (try again with your actual chat history).**

Task 3 — The cutoff test and professional-limits reflection

This task combines the "doesn't know now" and "not a professional" limitations into one practical exercise.

Part A: Testing the knowledge cutoff

Choose a topic where things have changed recently. Good options:

- A software library or framework (e.g., "What's the latest stable version of React, and what did the most recent major release add?")
- A regulatory or legal change in your field or country
- A recent scientific finding or product launch
- Current prices of something (a popular product, an API service you use)

Step A: Ask your AI the question without giving it any current information. Record its answer.

Step B: Search for the current, correct answer from an official source (the library's docs, a government website, a reputable news article).

Step C: Paste the current information into a new prompt and ask the same question, this time providing the real data: *"Here is the current information: [paste text]. Based on this, can you answer [question]?"* Compare the quality of both responses.

Part B: The professional-limits reflection

Think of a real question you've asked AI (or might ask) in one of these areas: health, law, finance, or another professional domain.

Step D: Reframe the question as "help me understand and prepare" rather than "tell me what to do." For example, if the original question was *"Should I take this medication?"*, reframe it as *"What are the standard considerations a doctor weighs when prescribing this medication? What questions should I ask my doctor at my appointment?"*

Step E: Compare the original question (if you ran it) with the reframed version. What's different about the output?

Reflect and write down:

- What topic did you choose for the cutoff test? How out of date was the AI's unaided answer?
- Did providing the current information in the prompt improve the response? How?
- What is your rule for when to use a search-enabled tool vs. a standard AI chat?

- What professional question did you choose for Part B? How did reframing it change the nature of the answer?
- How will you use AI for professional-domain questions going forward?

✓ **Self-check:** You should have one concrete example of the knowledge cutoff causing a wrong or outdated answer, and one example of reframing a professional question from "tell me what to do" to "help me understand and prepare." Both should feel immediately applicable to how you use AI.

What good looks like after this lesson

After completing these three tasks, you should be doing three things automatically: verifying specific AI claims before acting on them, thinking about data sensitivity before pasting anything into a chat, and reaching for a search-enabled tool or a real professional whenever currency or expertise is non-negotiable. These three habits will protect you from the vast majority of AI-related mistakes.