

Printable flashcards

Where is your progress saved? - In this browser, on this machine. It is not an account, and it does not leave the page.

Which sentence is the point of ?How to use this campus?? - One file, a track order, and a way to know what you have finished.

What is the course trying to produce? - Someone who can frame a task, choose a model and tools, evaluate the result, and hand the evidence to another person.

Which sentence is the point of ?What you are learning to do?? - The skill is judgment you can show, not a tour of product names.

Which field is the file another person can open later? - The artifact. The trace explains how you got there. The metric says how it scored.

Which sentence is the point of ?Keep proof for every exercise?? - Run, trace, metric, and artifact: four fields that make a lesson repeatable.

When is a tool allowed to send email on its own? - It is not. Sending waits for a person, even if the draft looks correct.

Which sentence is the point of ?Safety rules for every exercise?? - Private data stays local until you have a reason to move it, and some actions always wait for a person.

Is an API key required to study? - No. Add a local model or a hosted API only when an exercise needs a live call.

Which sentence is the point of ?What you need on your machine?? - A current browser is enough to study. A local model is optional and comes later.

Why does fine-tuning come after evals? - Because you need a measured failure of prompting and retrieval before training is a rational cost.

Which sentence is the point of ?How the tracks fit together?? - Models, then prompting, then a way to ground and act, then proof that it worked.

Why can a fluent answer still be false? - Because the model is continuing likely text. Likelihood is not a citation.

Which sentence is the point of ?A model predicts the next token?? - Fluency is the product of a probability model. It is not a source of record.

Why is 'about 0.75 words per token' not a budget? - It is an English-prose average, not a count for your model, your language, or your code.

Which sentence is the point of ?Tokens are the unit of cost and context?? - A token is a chunk of text, not a word, and every limit you hit is counted in tokens.

Who shares the context window? - The instructions, the retrieved or pasted sources, the conversation, and the tokens the model still has to write.

Which sentence is the point of ?The context window is one shared budget?? -

Instructions, retrieved text, the conversation, and the answer all spend the same token limit.

Does a low temperature stop hallucinations? - No. It reduces random variation. A likely-but-wrong token can still be the top choice.

Which sentence is the point of ?Temperature is a sampling control?? - Lower settings repeat likely tokens. Higher settings take more risks. Neither setting checks facts.

What do you score instead of the model's confidence? - Whether the answer matches a source, a tool result, or a known test case.

Which sentence is the point of ?Hallucination is fluent text without evidence?? - Catch

it with sources, tools, and tests. Do not catch it by asking the model if it is sure.

Where do tool results belong? - In a tool message, separate from the user's request, so the model can see what the tool returned.

Which sentence is the point of 'System, user, and tool messages do different jobs'? -

Put standing rules in the system message. Put the task in the user message. Put tool results in tool messages.

Why do you pin a model id? - Because a family name hides version changes that move quality, cost, and safety.

Which sentence is the point of 'Model families you will see in 2026 job posts'? - Hire and design for a role, then pin an id. A family name is not a version.

What should you try before you switch on a slower reasoning model? - A fast model on a fixed set of examples, after the prompt and output shape are clear.

Which sentence is the point of 'Choose a reasoning model or a fast chat model'? - Spend extra compute when the task needs a plan. Do not spend it on extraction you can score cheaply.

What makes a done-check real? - A person or a test can apply it to the output without asking the model whether it tried hard.

Which sentence is the point of 'Write the job, the input, and the done-check'? - A prompt is a job description. If you cannot tell whether the answer succeeded, the prompt is not finished.

Why validate in code if you asked for JSON? - Because the model can still wrap it in prose, drop a bracket, or invent a key. The schema mode reduces that. It does not replace the check.

Which sentence is the point of 'Lock the output shape'? - Ask for a structure, then validate it in code. Prose that looks like JSON is not JSON.

Why include a refusal example? - Because otherwise every example shows a successful extraction, and the model learns to fill gaps.

Which sentence is the point of 'Few-shot examples that do not leak'? - Examples teach the pattern. They are also data you might not have permission to send.

When is a fixed workflow the better design? - When you can name the steps in advance and test each one. Choice belongs in a loop only when the next action depends on what came back.

Which sentence is the point of 'Split a hard task into steps you can test'? - One prompt that plans, searches, calculates, and drafts is several jobs hiding in a paragraph.

What is the revision allowed to change? - Only the lines that failed the checklist.

Which sentence is the point of 'Critique the draft, then revise once'? - A second pass helps when the critique is a checklist, not a request to 'make it better'.

Why is the model not the security boundary? - Because it is trained to follow instructions it reads. Code has to restrict tools, validate arguments, and gate side effects.

Which sentence is the point of 'Prompt injection is untrusted text that pretends to be an order'? - Documents, web pages, and tool descriptions can carry instructions. Your system does not have to obey them.

Sources and go deeper

Vaswani and co-authors, Attention Is All You Need, 2017 (paper)

last checked 2026-10

<https://arxiv.org/abs/1706.03762>

RFC 9110, HTTP Semantics (spec)

last checked 2026-10

<https://www.rfc-editor.org/rfc/rfc9110>

RFC 8259, The JSON Data Interchange Format (spec)

last checked 2026-10

<https://www.rfc-editor.org/rfc/rfc8259>

Model Context Protocol specification, 2025-06-18 (spec)

last checked 2026-10

<https://modelcontextprotocol.io/specification/2025-06-18>

Lewis and co-authors, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020 (paper)

last checked 2026-10

<https://arxiv.org/abs/2005.11401>

Large language model - Wikipedia (wikipedia)

last checked 2026-10

https://en.wikipedia.org/wiki/Large_language_model

This is an independent, free, learning-only student project. It respects every company and vendor, and every trademark belongs to its owner. It is not affiliated with or endorsed by any of them. Nothing here is sold. The author is happy to volunteer.

cyphermonkeylabs.com

No email address is printed on this sheet.