ALDEYA
AI

A PLAIN-LANGUAGE GUIDE FOR K–12 EDUCATORS

AI 101

The chips, data centers, and devices behind the AI tools showing up in our classrooms.

1 What is AI, in plain terms?

AI refers to computer systems that learn patterns from large amounts of data and use those patterns to make predictions or generate content. The chatbots most of us encounter (ChatGPT, Gemini, Claude) are one type of AI. They don't think or understand the way people do. They predict what words are most likely to come next based on patterns in the text they were trained on. That's a real capability with real limitations, and it helps to keep both in view.

2 What is a large language model (LLM)?

An LLM is the type of AI behind most chatbots. "Large" refers to the enormous amount of text it learned from and the billions of internal settings it uses to make predictions. When you type a prompt (a message or request to the AI), the model generates a response one word at a time, choosing each word based on probability. This is why responses sound fluent but can also be confidently wrong.

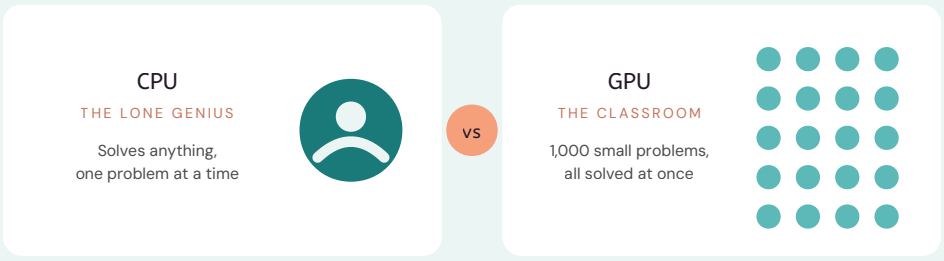
Most LLMs learn primarily from text gathered across the internet, so the biases, errors, and harmful content found online can show up in the patterns a model learns. Companies filter their training data and use human reviewers to shape how models respond, but no model is completely free of these issues. Some organizations also train on carefully curated, human-created data. In education, that can mean real student writing samples and teacher feedback, which helps a model handle classroom tasks more reliably. Where a model's training data came from is a fair question to ask of any tool you bring into your practice.

3 What does it mean to "train" an AI?

Training is the learning phase. Engineers feed a model billions of pages of text or images, and the model gradually adjusts itself to capture the patterns in that data. Training a large model takes weeks or months, thousands of specialized chips, and electricity costs that can run into the millions of dollars. It happens once (or occasionally, for new versions), and it happens long before you ever type a prompt.

4 Why does AI need special chips?

Picture a math test with 1,000 simple addition problems. A regular computer chip (a CPU) is like one very capable student who can solve any problem, but only one or two at a time. An AI chip (a GPU) is like a classroom of 1,000 students who each take one problem and work simultaneously. AI doesn't do one big hard task. It does billions of tiny math problems at once, so it needs the classroom, not the lone genius. Training a model like ChatGPT on regular chips would take decades instead of weeks.



Why AI runs on GPUs: billions of tiny math problems, worked in parallel.

PART 2 THE BUILDINGS

5 What is a data center?

If a chip is a classroom, a data center is the whole school building. It houses thousands of chips and provides everything they need: massive amounts of electricity, industrial cooling systems (the chips run hot), storage for the data, and ultra-fast internal connections so the chips can share work instantly. Data centers are large industrial facilities, and their energy and water demands have become a significant issue in the communities where they're built.

6 When I use a chatbot, am I using a data center?

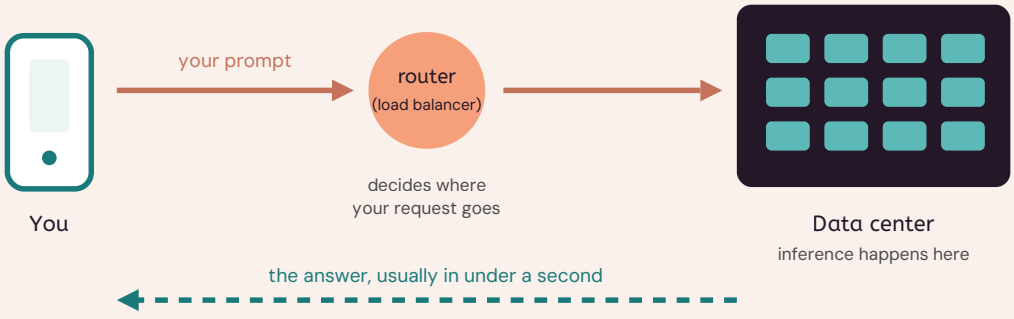
Yes. Every time you type a prompt, your request travels over the internet to a data center, gets processed by chips there, and the answer travels back. This is called inference, and it's distinct from training. Training is the occasional, massive effort to build the model. Inference is the constant, everyday work of answering billions of user prompts. Inference now accounts for the majority of AI's total energy use, because even though one query is small, billions of them add up.

7 Why are some data centers in remote areas and others near cities?

It comes down to what the building is for. Training centers are usually built in remote areas where land and electricity are cheap, because training doesn't need to be fast for any particular user. Inference centers are increasingly built near cities to reduce latency, the delay between hitting enter and getting a response. Think of remote centers as factories and urban centers as local distribution points. The split isn't absolute, but that's the general pattern.

8 Who decides where my question gets processed?

Not the AI itself. Before the model even sees your prompt, a piece of infrastructure called a router or load balancer decides where to send it. It considers the type of app, your location, and sometimes your account tier. A voice assistant needs a fast, nearby chip, and sometimes the work stays on the NPU in your phone instead of traveling anywhere at all. An essay generator can wait a few seconds, so that request may go to a cheaper data center farther away. Paid users often get routed to faster, closer servers than free users.



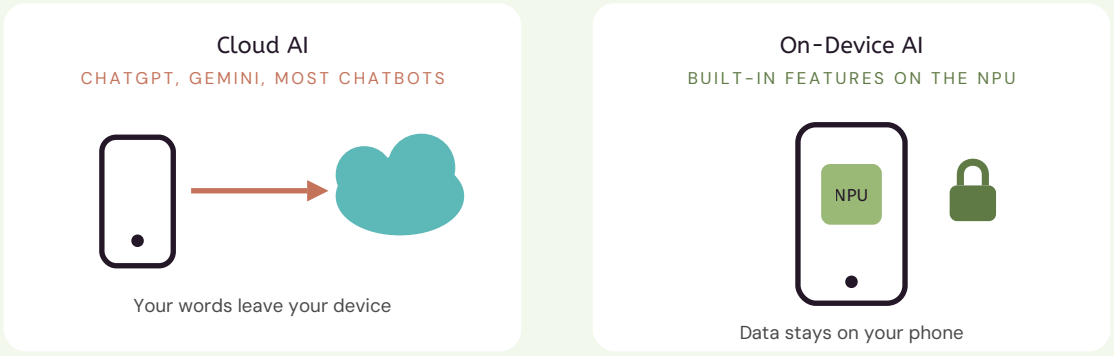
The journey of a prompt: every chatbot reply is a round trip to a building full of chips.

9 What is an NPU?

An NPU (Neural Processing Unit) is a small AI chip built directly into newer phones, tablets, and laptops. It handles AI tasks on the device itself, using very little battery. If a GPU is a sports car, an NPU is an efficient scooter. It can't train a giant model, but it can run smaller AI features locally, like summarizing a note or identifying a face in a photo, without sending anything to a data center.

10 If my phone has an NPU, does my data stay private?

Only if the app is built to use it. This is an important distinction for educators. Apps like ChatGPT send your prompts to the cloud (someone else's computers in a data center) regardless of what chip your phone has, because that's how they're designed. Built-in features from the phone maker (like on-device summarization) are more likely to stay local. A useful rule: if an app requires a login and shows a loading animation, your data is probably leaving your device. Warmth can be another clue. If your phone heats up while running an AI feature, the work is likely happening on the device itself, since cloud-based AI leaves your phone cool. This matters whenever student information or personal health details are involved.



The chip can only protect your data if the app is designed to use it.

11 What is a "core"?

A core is an independent worker inside a chip. Older chips had one core, so they did one thing at a time. Modern chips have many. Inside an NPU, each core contains thousands of tiny calculation units all working in parallel. So when your phone's spec sheet says "16-core Neural Engine," it means sixteen specialized workers, each with thousands of calculators on their desk.

PART 4 THE COSTS

12 Where does all the electricity come from?

This is one of the biggest open questions in tech right now. AI's energy appetite has grown so fast that companies are buying or reopening nuclear power plants and building massive battery installations to keep data centers running. In some regions, data centers compete with residential neighborhoods for power, which can affect local electricity prices. One reason companies are pushing AI onto phone NPUs is efficiency: work done on your device is work a power-hungry data center doesn't have to do.

13 What does this have to do with the water issues some towns near data centers are experiencing?

Heat is the connection. Thousands of chips running at once generate enormous heat, and many data centers cool them by evaporating water, which can consume millions of gallons a day at a single facility. That water often comes from the same municipal supply residents use, so in drought-prone or fast-growing areas, communities have raised real concerns about strain on local resources. Some newer facilities use closed-loop liquid cooling that recycles water instead of evaporating it, but the older approach is still common. When a data center is proposed nearby, water is usually one of the first questions residents ask, and for good reason.

KEY TERMS AT A GLANCE

LLM: large language model; the pattern-prediction engine behind most chatbots

CPU: the general-purpose chip in every computer; solves anything, one task at a time

GPU: the AI workhorse chip; thousands of small calculations at once

NPU: a small, efficient AI chip built into newer phones and laptops

Core: one independent worker inside a chip; modern chips have many

The cloud: someone else's computers in a data center, reached over the internet

Prompt: the message or request you type to an AI

Training: the learning phase, when a model absorbs patterns from massive datasets

Inference: the everyday phase, when a trained model answers your prompts

Latency: the delay between sending a prompt and getting a response

Load balancer: the traffic director that routes your prompt to a data center

Data center: the industrial building that houses thousands of AI chips

KEEP LEARNING

This guide pairs with Campana.

Campana is the free weekly newsletter on AI for K–12 teachers. One practical tip you can use Monday morning, one curated read, and a take from someone who has actually taught. Read past issues and subscribe at campana.aldeya.ai.

campana.aldeya.ai

ALDEYA
AI