DAY 55 · Part 5 · Looking Inside

Can your phone run AI all on its own? — AI on your phone

Five-Minute AI Lessons · Oct 31, 2026

Your phone doesn't send it all out or keep it all in. It holds on to just the pieces it can handle.

In a nutshell

1. Some of it, yes. Phones these days have **a small model with about 3 billion numbers** inside, so even with no internet they can summarize, polish your writing and sort your notifications.

2. But big jobs are too much for it. Long writing or hard questions get sent to **the building you saw yesterday**, and the answer comes back. The phone model and the server model differ in size by tens to hundreds of times.

3. So phones today **split the work into pieces**. Whatever it can do, it finishes inside — it's faster, uses less electricity, and your text doesn't leave the phone.

Pieces on the phone vs pieces on a server

Pieces that run on the phone

· Small model, about 3 billion numbers
· Works right away, even without internet
· Your text never leaves the phone

→

Pieces that run on a server

· Big model, hundreds of billions of numbers
· Only works when connected
· Much better at hard jobs

The phone tries first and sends out only the pieces that are too much. Phones today mix the left and the right.

Key idea: it doesn't send everything, and it doesn't keep everything — it splits it into pieces

Does this job finish on the phone, or go out?

Internet off

What works in airplane mode

Read text in a… | Sum up a notifi… | Fix spelling | Suggest a short… | Write a long pi… | Search the web… | Solve hard math

The small model on the phone handles the four green ones by itself. They run even in airplane mode, and best of…

Internet on

Only heavy jobs make the trip

Read text in a… | Sum up a notifi… | Fix spelling | Suggest a short… | Write a long pi… | Search the web… | Solve hard math

The three orange ones leave the phone and go all the way to a data center and back. They do much better, but yo…

Handling one photo

Even one job splits into pieces

Find faces | Pull out text | Sharpen blurry… | Describe the sc… | Find similar ph…

Even one request doesn't go out whole. The phone tries first and sends out only the pieces that are too much. T…

Tap the buttons to switch situations. Green pieces finish on the phone, orange pieces leave the phone, and fade…

Try it yourself in the web lesson · next-token.net/en/learn/day-55/

Quick check

Just three questions

Q1

What's true about an AI model that runs on your phone?

① It's much bigger than server models
② It can't handle photos at all
③ It uses more electricity
④ It's small and runs without internet

Q2

What happens when a hard question gets sent to a server?

① It does better, but your text leaves the phone
② Your phone battery lasts longer
③ You don't need internet
④ The answer is always wrong

Q3

How do phones actually work today?

① All on the phone
② Split into pieces
③ All on a server
④ They do nothing

Know the why, too

Q1

④

The on-phone model released in 2024 had about 3 billion numbers. It's tens to hundreds of times smaller than server models, but it runs even in airplane mode.

Q2

①

The big model does much better. But your text makes a trip to that building you saw yesterday and back. What you get and what you give up come together.

Q3

②

You saw it in the third button of the experiment. Even handling one photo splits into pieces for the phone and pieces for the server. The phone tries first and sends out only...

Your phone doesn't send it all out or keep it all in.
It holds on to just the pieces it can handle.

Tomorrow: The AGI debate — will AI that can do what people do really arrive? Let's see why grown-ups argue about it.