EdSutra

Teaching guide

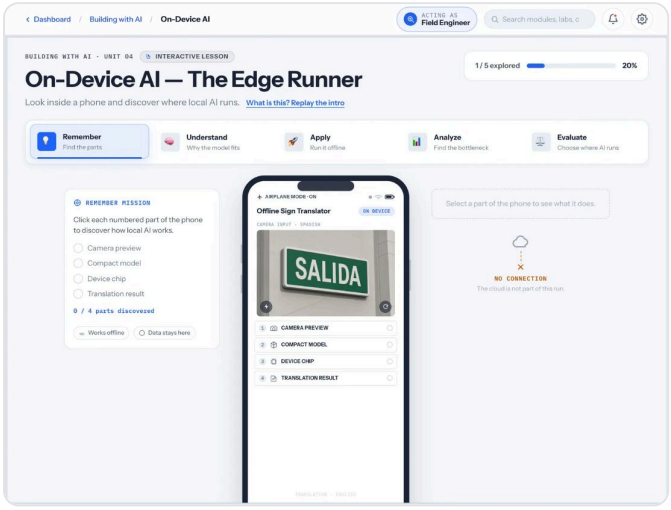
Module 06 · Building with AI · Unit 04 · Stage 3, Responsibility

On-Device AI

The Edge Runner

Grades 9–12 | 45 minutes | Role: **Field Engineer**

Grade 9: About age 14–15 · Year 10 in the UK · Class 9 in India



The big idea

Some AI runs on the device itself: the model is downloaded once, stored locally and run by the device's own chip, so nothing needs to be sent away. To fit limited memory, models are made smaller, which can cost quality. Where AI should run depends on the task, connection, privacy, memory and safety.

The learner's job

Learners are the **Field Engineer** at Kestrel Works · Field Unit. This has to run with the network off and the battery draining. Their task: make it work on the device. Say what you gave up to fit.

By the end, learners can say

- I can name the four parts of a local AI pipeline: camera, compact model, device chip, result
- I can explain why lower-precision models use less memory — and that this can affect quality
- I can run a simulated offline translation and identify what made it possible without a connection
- I can diagnose a failed local run as a memory, compute, or task-fit problem, and choose a valid repair
- I can choose and defend where an AI task should run — on-device, cloud, or hybrid — using named constraints

Before class

- One computer or tablet for each learner or pair, with EdSutra open (this Module is part of the full EdSutra studio).
- Open the studio once on each device before class, so the Unit still runs if the internet drops.
- Play the first two steps yourself, so you know where learners will click and how the precision control works.
- Draw a phone on the board with four boxes: camera, compact model, chip, result. Learners fill in each job during Remember.
- Pairs work well: one learner drives, the Field Engineer says what the smaller build gave up to fit.
- Best after Build Your AI and AI Pipelines.
- Units open in order, and Module 1 comes first. To start a class here, switch on Open every Unit in Settings → Teachers & parents.

Words to introduce

On-device (edge) AI: AI that runs on your phone or laptop, not on a distant server.

Inference: Using a trained model to produce an answer, like a translation.

AI memory budget: How much of the device's memory is free for running AI.

Precision (bits): How many bits store each number in the model; fewer bits, smaller file.

Quantization: Storing a model's numbers with less precision so it uses less memory.

Distillation: A small student model learns to imitate a larger teacher model.

Hybrid: Part of a task runs on the device and part in the cloud.

Running the lesson 45 minutes

Open 5 min

"No signal, no help, just the device. Can a phone still translate a sign? Where would the AI have to live?"

Before anyone clicks, ask: if we store a model's numbers with less detail, will the file get smaller? Will its answers change? Record a hands-up vote for each and come back to it in Understand.

Step	What learners do	Ask the class	Listen for
1. Remember 5 min	Click each numbered part of the phone (camera preview, compact model, device chip, translation result) to see how the sign SALIDA becomes EXIT. Then answer where the image was processed.	• Which part does the actual AI calculation? • How much of the image was sent to a server?	The model is already stored on the phone and its chip runs it. The run sends 0 KB of the image away.
2. Understand 9 min	Answer two predictions, then move the precision control from 32-bit down to 8-bit and 4-bit, and watch the model's size change against the phone's 2 GB AI memory budget.	• At which precision did the model first fit the budget? • Why might fewer bits change the answers? • Look at our vote. Were we right?	Fewer bits per number means a smaller file that can fit, but quality can change, depending on the model and task.
3. Apply 8 min	Choose a phone or tablet and a model (compact 0.8 GB or general 4.2 GB), keep airplane mode on, and run. Then pick the three things that made the offline run possible.	• What happened when you ran the general model? • Did airplane mode make the translation work? What did?	The 4.2 GB model cannot load. The compact one works: stored locally, fits the budget, run by the chip.
4. Analyze 9 min	Read the evidence table for each failed run (model doesn't fit, too slow, wrong task). Pick what stopped it and a repair that really fixes it. Learners need two of the three right.	• The medical answer came back fast. Why is it still a failure? • Why won't turning Wi-Fi on fix the memory case?	Memory, compute and task fit are different problems with different fixes. A fast, confident answer is not a safe one.
5. Evaluate 9 min	For at least two scenarios (hiking, broad research, private journal, medical symptoms), choose on-device, cloud or hybrid, watch the consequences change, tag two constraints and pick a defense that matches.	• Why is the cloud ruled out on the hiking trail? • In the medical case, does it matter where the AI runs?	Choose the least resource-heavy place that can do the task responsibly. For medical symptoms, a person is needed wherever it runs.

Steps follow Bloom's revised taxonomy, in the order the Unit presents them.

Deepen the thinking

Critical thinking	An app says its AI is "private because it runs on your phone". What would you check before believing that?
AI thinking	A compact model fits by using fewer bits. Which tasks would you trust a smaller model with, and which would you test carefully first?
Ethics and fairness	Many people own older or cheaper phones with less memory. Who might be left out if useful AI only runs on new devices, or only in the cloud?
Connect to life	When have you needed something to work with no signal, like on a trip? What would you want your phone to do offline?

Misconceptions to listen for

"On-device AI is just a smaller cloud model."
In Remember and Apply, the model file is already on the phone and its chip does the work. 0 KB of the image is sent anywhere.

"A smaller model is just as good."
In Understand, fewer bits make the file smaller, but output quality can change. Narrow tasks may stay strong; hard or sensitive ones need testing.

"Airplane mode is what makes offline AI work."
Airplane mode only proves the cloud is unreachable. The run works because the model is stored on the phone, fits its memory and runs on its chip.

In real life the card learners see at the end of the Unit

- Some phones can do offline translation, voice typing or photo search on the device, with no data sent away.
- Cars with lane warnings run their AI on board, because they can't wait for an internet signal.
- Downloading a language pack for offline translation is "download once, run offline" in action.

Try it this week. With a parent or teacher, find one phone feature that still works in airplane mode. What does that tell you about where it runs?

Watch out for. Smaller models aren't free: shrinking one can lower its quality on hard or sensitive tasks. Some jobs still need the cloud.

Talk about it. What would you rather keep on your own device instead of sending online, and why?

Exit ticket

1. Name the four parts of the local translation, in order.

2. Why does storing numbers with fewer bits help a model fit on a phone?

3. Why did the 4.2 GB model fail on the phone, and what fixed it?

Going further

For homework, use the "Try it this week" activity from the In real life card: with a parent or teacher, find one phone feature that still works in airplane mode, and say what that shows about where it runs.