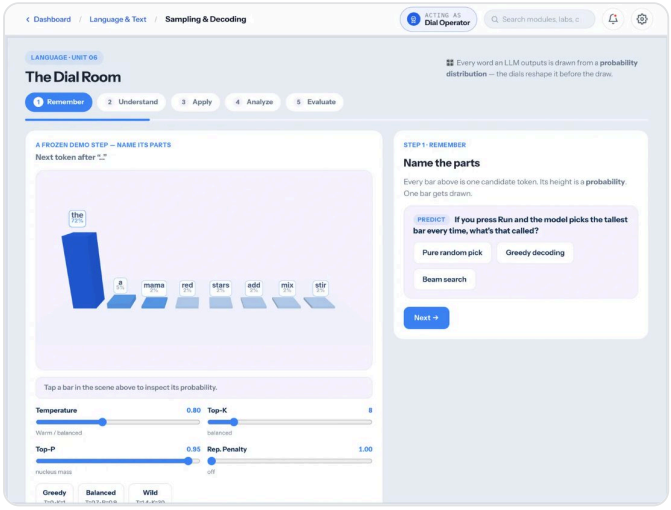
Module 03 · Language & Text · Unit 06 · Stage 2, Capabilities

Sampling & Decoding

The Dial Room

Grades 8–12 | 40 minutes | Role: **Dial Operator**

Grade 8: About age 13–14 · Year 9 in the UK · Class 8 in India



The big idea

A language model does not know the one right next word. It gives every possible next word a probability, and a sampling rule picks one. Dials like temperature, Top-K, Top-P and a repetition penalty reshape those odds before the pick. Careful settings give safe, repetitive text; loose settings give lively, riskier text. The job decides which is right.

The learner's job

Learners are the **Dial Operator** at Bureau of Plain Language · Dial Room. This machine can be careful or surprising. Somebody has to choose which, and for what. Their task: learn what each dial does. Match the setting to the job.

By the end, learners can say

- I can explain how language models generate text token by token
- I can describe what temperature controls and its effect on output diversity
- I can compare top-K and top-P (nucleus) cuts, and use a repetition penalty to break loops
- I can design sampling configurations for different use cases (creative vs. factual)

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 the dials are and what the Run button does.
- Draw a bar chart on the board with five words and their chances. You will flatten and sharpen it by hand during Understand.
- Best after Language & AI Intro and The Meaning Detector.
- 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

Token: One piece of text the model writes at a time; here, a whole word.

Probability distribution: The chance the model gives each possible next word; they add up to 100%.

Greedy decoding: Always picking the most likely next word. Same input, same output.

Temperature: A dial that sharpens (low) or flattens (high) the odds before the pick.

Top-K: Keep only the K most likely words, then pick among them.

Top-P: Keep the smallest group of top words whose chances add up to P.

Repetition penalty: Makes recently used words less likely, so the text stops looping.

Running the lesson 40 minutes

Open 4 min

"Ask a chatbot the same question twice and you can get two different answers. If it is the same machine, where does the difference come from?"

Before anyone clicks, ask: if we press Run three times with the same settings, will the machine write the same sentence each time? Take a vote, write it on the board, and come back to it in Apply.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	A frozen demo shows the candidate next words as bars; each bar's height is its probability. Learners tap bars to read them, then name the rule that always takes the tallest bar.	• What does the height of each bar tell you? • If the model always takes the tallest bar, what happens when you run it twice?	Each bar is a chance, not a certainty. Always taking the tallest is greedy decoding: same input, same output.
2. Understand 6 min	Learners drag the Temperature dial from low to high, watch the bars reshape live, then say what happens to them above 1.0.	• At low temperature, which word wins almost every time? • At high temperature, why do rare words start getting picked?	Low temperature sharpens the odds toward one word; high temperature flattens them, so unlikely words get a real chance.
3. Apply 8 min	Learners pick a mission and set the dials. Cold Start: three runs in a row come out exactly the same. Hot Take: three runs share no three-word phrase.	• Which dial did you move first, and why? • Look at our vote on the board. Does the machine always write the same sentence?	Only very sharp settings, like temperature near 0, repeat exactly. Otherwise each word is drawn, so runs differ.
4. Analyze 9 min	Learners predict which cut keeps a fixed number of words, Top-K or Top-P. Then, with Temperature at 0.4 or below, they raise the Repetition Penalty until no three-word phrase repeats.	• Why does Top-P keep fewer words when the model is confident? • Why not just raise the temperature to stop the loop?	Top-K keeps a fixed count; Top-P adapts. The penalty fixes loops without adding the randomness a hotter temperature brings.
5. Evaluate 9 min	Learners choose settings for a factual assistant, then hit the Goldilocks target: at least 10 words, 6 different words and no word twice in a row. A three-question quiz closes the Unit.	• Why does a factual assistant want low temperature and a small Top-P? • What would your Goldilocks settings be bad at?	A choice tied to the job and its cost: safe settings can be dull; lively settings can be wrong.

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

Deepen the thinking

Critical thinking	Two people ask a chatbot the same question and get different answers. What would you need to know before deciding which answer to trust?
AI thinking	The Dial Room's model learned from 30 short sentences. What would change if it learned from millions, and what would stay the same?
Ethics and fairness	A city uses an AI with a high temperature to write its safety notices, so they sound lively. Who could be harmed, and who should choose the setting?
Connect to life	Think of a music playlist on shuffle versus repeat. Where else do you choose between surprise and the same thing every time?

Misconceptions to listen for

"A higher temperature makes the AI more creative or smarter."
Temperature only reshapes the odds. In Understand the bars flatten; the model knows nothing new, so rare words, right or wrong, get picked more.

"Top-K and Top-P do the same thing."
Top-K always keeps exactly K words. Top-P keeps fewer when the model is confident and more when it is unsure, as Analyze shows.

"When the model repeats itself, it is broken."
Loops come naturally from sharp, low-temperature picks. In Analyze the Repetition Penalty makes recent words less likely and breaks the loop without adding randomness.

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

- Ask a chatbot the same question twice and answers may differ: each word is drawn from likely options.
- Some AI writing tools offer "creative" or "precise" modes, which can work a bit like a temperature dial.
- At a school raffle, a name with more tickets wins more often. AI picks next words much the same way.

Try it this week. Put word cards in a bag, with extra copies of likely words. Draw to build a silly sentence, then remove the rare words and draw again.

Watch out for. A higher temperature doesn't make AI smarter or more truthful; it only makes word picks less predictable. The right setting depends on the job.

Talk about it. For homework facts, would you want a predictable AI or a surprising one? What about story ideas?

Exit ticket

1. What does turning the temperature up do to the bars, and to the text?

2. Top-K and Top-P both cut words. Which one keeps a fixed number?

3. Choose settings for a factual helper and name one thing they cost.

Going further

In the AI Story Canvas Lab, learners build a choose-your-own-adventure story where each scene offers a few next steps. There the reader picks the path; in the Dial Room, weighted chance does. For homework, try the In real life card's word-bag activity.