SESSION 5 REFERENCE GUIDE

Session 5: Add Controls — Instructor Guide

AI + Research Level 2 — Supplementary Material

Concept: HYPERPARAMETERS

Space: Text Playground
Model: `distilgpt2` (~80MB, fast on free CPU)

Time Breakdown (2 hours)

0:00–0:05 — Show-and-Tell

- Ask: "Did anyone add a fourth model or find maximum disagreement?"
- Quick share of between-session experiments.

0:05–0:15 — The Hook: Same Prompt, Different Settings

- Open the finished Text Playground Space.
- Type: **"Once upon a time in a school where robots"**
- Generate with **temperature = 0.1** — predictable, safe output.
- Same prompt, **temperature = 1.5** — wild, chaotic output.
- Ask: "Same model, same prompt. What changed?"

0:15–0:50 — Live Build (35 minutes)

1. **Create new Space on Hugging Face** — "Text Playground"
2. **Write requirements.txt** — `transformers`, `torch`, `gradio`
3. **Build app.py step by step:**
   - Import libraries
   - Load `distilgpt2`
   - Write a basic generation function (just prompt → output)
   - Test it — "it works but we can't control it"

    - Add temperature slider — "now we have a knob"
    - Add top-p slider
    - Add max-length slider
    - Wire up Gradio with all inputs
    - Add examples
4. **Deploy and test**

Teaching moment when adding sliders: "Every slider we add is a hyperparameter. The model's weights are fixed — we're not changing what it knows, we're changing how it behaves."

0:50–1:20 — Systematic Experimentation (30 minutes)

Guide students through controlled experiments. Use the same prompt each time and change ONE slider:

Experiment 1: Temperature
- Prompt: "The secret ingredient in the recipe was"
- Temperature: 0.1, 0.5, 0.7, 1.0, 1.5, 2.0
- Observation: Low = repetitive/predictable. High = creative/chaotic. Very high = gibberish.

Experiment 2: Top-p
- Same prompt, temperature fixed at 0.7
- Top-p: 0.1, 0.5, 0.9, 1.0
- Observation: Low top-p = only the most likely words. High top-p = more variety.

Experiment 3: Max Length
- Same prompt, same temperature/top-p
- Max length: 20, 50, 100, 200
- Observation: Longer isn't always better. Models can lose coherence.

Key question: "If you change two sliders at once, can you tell which one caused the change?" (No — this is why scientists change one variable at a time.)

1:20–1:40 — Name the Concept: HYPERPARAMETERS (20 minutes)

- "Hyper" = above/beyond. These are parameters that sit above the model's learned parameters.
- The model has millions of internal parameters (weights) learned during training — we can't change those.
- Hyperparameters are the knobs WE control at runtime.

- Every AI tool you've ever used has these — ChatGPT, image generators, all of them. Most just hide the sliders.

Analogy: "A guitar has fixed properties — the wood, the strings, the shape. Those are like the model's weights. But you control how hard you strum, where you pick, whether you use a capo. Those are hyperparameters."

1:40–1:50 — Challenge: Best Settings for the Job (10 minutes)

Students find optimal settings for different tasks:

Task	Goal
Scary story opening	High creativity, medium length
Formal email to a teacher	Low creativity, controlled length
Funny random story	High creativity, longer length
News headline continuation	Low creativity, short length

"There's no single 'best' setting — it depends on what you're trying to do."

Frame distilgpt2's low quality as a feature: "This is a tiny model — about 80MB. GPT-4 is estimated to be 1,000x bigger. But the same controls work on both. You're learning the universal remote, not just one TV."

1:50–2:00 — Notebook Time

Share the Colab link in the Zoom chat.

Walk through together:
1. Run the setup cell (installs transformers, loads distilgpt2)
2. Run the first generation cell — everyone sees the same prompt, different output
3. "Now try changing the temperature value in your notebook"

Notebook skill being introduced: Using sliders/widgets in the code — changing parameter values and re-running cells to see different outputs.

Say: "The experiments in the notebook let you try all the combinations we didn't have time for. Find the best settings for a scary story — bring your recipe to next session."

Hyperparameter Reference (Instructor Knowledge)

Temperature

- Controls the randomness of token selection.
- Mathematically: divides the logits (raw scores) before softmax.
- Low temperature (0.1–0.3): Model almost always picks the highest-probability word. Output is repetitive and "safe."
- Medium temperature (0.5–0.8): Good balance. Most common for production use.
- High temperature (1.0–1.5): More random selections. Creative but less coherent.
- Very high (>1.5): Approaches uniform random selection. Often gibberish.

Top-p (Nucleus Sampling)

- Instead of considering all possible next words, only consider the smallest set whose cumulative probability exceeds p.
- Top-p = 0.1: Only the very top words are considered (maybe 1-3 words).
- Top-p = 0.9: Most words are considered, excluding only the very unlikely ones.
- Top-p = 1.0: All words are considered (no filtering).
- Works in combination with temperature — both affect randomness but in different ways.

Max Length

- Maximum number of tokens in the output (including the prompt).
- Tokens ≈ words but not exactly (roughly 1 token = 0.75 words for English).
- Model may stop before max_length if it generates an end-of-sequence token.
- Longer doesn't mean better — models can lose coherence in long generations.

Pre-Prepared Prompts

Prompt	Why It's Interesting
"Once upon a time in a school where robots"	Fantasy/narrative — shows how temperature affects storytelling
"The secret ingredient in the recipe was"	Concrete completion — low temp gives real ingredients, high temp gets weird
"Dear Principal, I am writing to request"	Formal register — shows how low temp maintains formality
"Breaking news: scientists discover that cats"	News style — fun to see how creativity slider changes "news"
"The haunted house at the end of the street"	Horror genre — high temp makes it creepier
"In the year 2050, schools will"	Future speculation — good for comparing constrained vs. wild outputs
"The most important rule of cooking is"	Instructional — low temp gives real advice, high temp gives absurd rules

Technical Notes

- **Model size:** distilgpt2 is ~80MB, loads fast on free CPU.
- **Generation speed:** Expect 5-15 seconds per generation on free CPU, depending on max_length.
- **Temperature floor:** Code uses `max(temperature, 0.01)` to avoid division by zero.
- **Token limit:** max_length is in tokens, not words. 100 tokens ≈ 75 words.

Concept Connections

- **Session 4:** Students compared models (different weights/training data).
- **Session 5 (this session):** Same model, different settings — hyperparameters.

- **Together:** Model choice AND hyperparameters both affect output. Evaluation has to account for both.