TEACHER-FACING FRIDAY DESIGN

Why Models Behave Differently: Parameters, Variability, and Control

CAS Human-AI Collaboration | Module 1 | Unit 3

LITERACY	COLLABORATION MOVE	CONTACT TIME	CHALLENGE
Mechanism	**Instruction + mechanistic control**	**365 minutes**	**Parameter Control Experiment**

Position in Module

- **Unit:** 3
- **Topic:** Why Models Behave Differently: Parameters, Variability, and Control
- **Literacy perspective:** Mechanism
- **Collaboration level:** Instruction, with mechanistic control
- **Fundamentals focus:** Generative AI Fundamentals
- **Why this Friday matters:** Participants learn to stop treating model variability as inexplicable noise. They learn to hold a task stable, vary one generation condition, repeat trials, and use the resulting evidence to decide when variability is useful, when it is risky, and which controls are appropriate.

Literacy Perspective

- **Primary literacy:** Mechanism
- **Secondary literacy, if any:** None
- **What this literacy explains:** Mechanism Literacy explains why output depends not only on the visible request but also on tokenisation, attention, sampling, model conditions, and probabilistic next-token selection. It also explains why exact character operations and perfectly reproducible generation can remain fragile.
- **What participants should manipulate, observe, and interpret:** Participants keep the task and request stable while manipulating one visible generation or model condition. They observe within-condition variation, between-condition differences, constraint compliance, factual stability, and failure patterns, then interpret how much control the manipulated condition actually provides.

Collaboration Level

- **Level relation:** Instruction, with mechanistic control
- **Collaboration move being learned:** Participants extend structured instruction by specifying and testing the generation conditions under which the instruction is executed. They move from issuing a request once to designing a controlled, repeatable comparison.

- **What participants should be able to do differently after this Friday:** They should be able to choose one credible mechanism-level variable, predict its effect, run repeated trials, distinguish systematic effects from ordinary variation, and formulate a fit-for-purpose control rule without claiming more mechanistic certainty than their evidence supports.

Generative AI Fundamentals

This Friday should provide the practical mechanistic layer that explains why similar prompts can behave differently, and why some tasks are easy for LLMs while others remain fragile.

- Introduce tokens as the smallest units the model directly works with. The model does not see text exactly as humans see letters, words, jokes, or word parts; it sees sequences of tokens.
- **Explain tokenisation with a simplified byte-pair-style process:** start from small units such as characters or bytes, repeatedly merge frequent adjacent sequences from the training corpus, and build a vocabulary of reusable subword units. The important point is that tokens are learned from statistical patterns in text, not from explicit semantic rules.
- Use tokenisation to explain typo resilience. If `Andelfingen` is represented through partly overlapping pieces such as `and`, `elf`, and `ingen`, then a misspelling such as `Andlefingen` may still share enough token-level and contextual evidence to preserve meaning. Frequent typo variants in similar textual neighbourhoods can also make the model robust to spelling errors.
- Use tokenisation to explain character-level limits. Tasks that require exact letters, positions, reversals, word lengths, or crossword-like constraints can be fragile because the model is not internally operating on human-visible characters in a reliable rule-based way. If exact character inspection matters, the information or procedure must often be externalized through explicit decomposition, examples, lists, or tools.
- Introduce attention as the mechanism by which the model weighs relationships among tokens in the current sequence. This explains why surrounding context can change the interpretation of ambiguous tokens and why longer context can influence later output.
- Connect these mechanisms to variability and control. The model predicts a distribution over possible next tokens; generation settings such as temperature or sampling strategy influence how that distribution is used. This helps explain why the same prompt can produce different outputs and why controlled comparisons are needed.
- **Caveat to demonstrate:** the lowest unit the model directly processes is the token, not the human-visible character. Character-counting, letter-position, reversal, and word-length tasks therefore work against the model's native representation unless the task is decomposed, tool-supported, or covered by memorized textual patterns in training data.
- **Emphasize practical control rules:** if a task depends on exact characters, use explicit checks or tools; if reliability matters, use constraints, repeated runs, grounding, tests, or external verification; if creativity or exploration matters, controlled variability can be useful.

Recurring Agentic Coding Demonstration

Demonstrate the same bounded coding task under two model or reasoning conditions while keeping the repository state, task brief, target files, permissions, and test command stable. Run each condition more than once. Preserve the plan, diff, test result, and short run metadata for every trial, then compare planning depth, patch consistency, constraint compliance, edge-case handling, and reproducibility. The demonstration makes Mechanism Literacy visible without implying access to hidden model internals: participants infer from controlled behaviour. Transfer the same logic to any professional task in which a model, sampling setting, or reasoning mode can be varied while the task remains fixed.

Shared Classroom Experiment

Use one common task before participants transfer the method to their own contexts. This ensures that differences between groups can be discussed against the same input and evaluation criteria.

Default task:

> **Using a fixed briefing about a fictional organisation's internal knowledge-base pilot, produce a 160-word management note containing exactly three expected benefits, three risks, and one recommendation. Do not introduce facts that are absent from the briefing.**

Prepare a fixed experiment pack containing:

- a 400- to 600-word fictional briefing with several concrete facts and one explicit uncertainty
- one locked request used without wording changes across all main runs
- a scoring sheet for structural compliance, source fidelity, unsupported additions, and usefulness
- two configuration cards that differ on one condition only; the default is lower versus higher sampling variability
- a run log with space for configuration, run number, output, score, and unexpected behaviour
- three short tokenisation or character-level micro-tasks for demonstrating the representation caveat

Each group should run at least three trials per condition. The default comparison is a lower-variability condition against a higher-variability condition using the same model. If the available environment does not expose sampling settings, the teacher may substitute one visible model or reasoning condition, or use prepared output sets with authentic configuration metadata. Do not vary the prompt, model, and sampling settings simultaneously.

Friday Activity Plan

Indicative contact time is 365 minutes, excluding breaks and lunch. Teachers may adjust durations while preserving prediction, repeated trials, comparison, interpretation, and transfer.

Time	Activity	Concrete output
45 min	From text to probability distributions	Annotated mental model and two testable predictions
40 min	Representation caveat lab	Short caveat record with failure, repair, and control implication
55 min	Low-variability baseline	Three baseline outputs and completed score rows
55 min	Higher-variability comparison	Three target outputs and completed score rows
50 min	Separating signal from noise	Comparison table or plot, evidence-backed claim, and limitation
50 min	Designing fit-for-purpose control	Three short control policies with rationale
70 min	Challenge kickoff and first controlled run	Challenge protocol, prediction table, first run evidence, and replication plan

Detailed activity blocks

1

45 min **From text to probability distributions**

Content and manipulation. Connect tokens, attention, next-token distributions, sampling, and repeated generation. Distinguish observable model behaviour from inaccessible internal explanation.

Method and expected observation. Interactive input with prediction questions; participants annotate where the visible request ends and mechanism-level conditions begin.

Concrete output. Annotated mental model and two testable predictions

Tools and materials. Slides, token visualiser or prepared examples, prediction cards

2

40 min **Representation caveat lab**

Content and manipulation. Test character counting, letter positions, reversal, or unfamiliar compound words, then add explicit decomposition or a checking tool.

Method and expected observation. Pairs predict failures, run the micro-tasks, and identify which repair externalises work the model does not reliably perform at token level.

Concrete output. Short caveat record with failure, repair, and control implication

Tools and materials. Prepared micro-tasks, chat interface, optional simple script or spreadsheet

3

55 min

Low-variability baseline

Content and manipulation. Run the locked management-note task three times under the baseline condition. Keep all visible task inputs constant.

Method and expected observation. Small-group experiment; participants score every run before discussing impressions. Expected observation is that low variability may improve repeatability without guaranteeing correctness.

Concrete output. Three baseline outputs and completed score rows

Tools and materials. Shared experiment pack, controlled interface or notebook, run log

4

55 min

Higher-variability comparison

Content and manipulation. Repeat the same task three times after changing only the selected generation condition.

Method and expected observation. Small-group experiment with a second prediction; participants record both useful variation and losses of consistency or compliance.

Concrete output. Three target outputs and completed score rows

Tools and materials. Same environment, target configuration card, run log

5

50 min

Separating signal from noise

Content and manipulation. Compare within-condition spread and between-condition differences in structure, wording, unsupported claims, and usefulness.

Method and expected observation. Groups create a simple comparison display and challenge one causal claim made by another group.

Concrete output. Comparison table or plot, evidence-backed claim, and limitation

Tools and materials. Comparison sheet, spreadsheet or shared board

6

50 min

Designing fit-for-purpose control

Content and manipulation. Decide which configuration and verification strategy fits an exact task, an exploratory task, and a mixed task. Add repeated runs, constraints, tests, or external checks where appropriate.

Method and expected observation. Scenario design clinic; expected observation is that no single setting is best for every purpose.

Concrete output. Three short control policies with rationale

Tools and materials. Scenario cards, control-policy template

7 **70 min** **Challenge kickoff and first controlled run**

Content and manipulation. Select a context-relevant task, define one mechanism-level manipulation, specify constants and criteria, and capture the first comparable evidence.

Method and expected observation. Individual or paired coaching with peer review of variable isolation and feasibility.

Concrete output. Challenge protocol, prediction table, first run evidence, and replication plan

Tools and materials. Challenge template, run log, available model environment

Experiment Focus

- **Experiment focus:** the effect of one generation or model condition on variability and task control.
- **Primary manipulated condition:** lower versus higher sampling variability in the same model. Where temperature or an equivalent setting is unavailable, use one documented alternative such as model variant or reasoning mode.
- **What remains constant across comparison:** Keep the task, visible request, source material, model when possible, output limit, available tools, evaluation criteria, and execution environment stable.
- **Replication requirement:** Run at least three trials per condition so that one striking output is not mistaken for a stable effect.
- **Prediction requirement:** Before each condition, predict the expected direction of change in variability, compliance, factual stability, or usefulness and state why.

Observation Focus

- **Within-condition variability:** How different are repeated outputs produced under the same condition?
- **Between-condition effect:** Is the difference between conditions larger or more consistent than the variation within them?
- **Constraint compliance:** Does each output meet length, structure, and content requirements?
- **Factual stability:** Which claims remain stable, disappear, or become unsupported across runs?
- **Usefulness:** Does variation create genuinely useful alternatives or merely stylistic novelty?
- **Representation caveat:** Which exact-character tasks improve only after decomposition or tool support?
- **Evidence participants should capture:** locked request; configuration metadata; numbered outputs; completed scoring sheet; comparison table or plot; failed and repaired micro-task; and unexpected results.
- **Interpretation participants should be prepared to make:** Explain which observed differences are plausibly connected to the manipulated condition, where attribution remains uncertain, and which combination of settings and verification is appropriate for the selected task.

Challenge Kickoff Deliverables

Participants must leave Friday with:

- a selected task for which variability, consistency, or model condition plausibly matters

- one baseline and one target condition that differ on one documented mechanism-level variable
- a locked request and fixed source material or input
- a prediction table covering variability, compliance, factual stability, and usefulness
- a protocol draft identifying the task, intended output, manipulated condition, constants, run count, evaluation criteria, and caveat
- a replication plan with at least three trials per condition, or a justified alternative when runs are costly
- at least one captured run with configuration metadata and a completed observation row
- a plan for external verification when exactness or factual reliability matters

The kickoff should already reflect the challenge-focus logic:

- **Difficulty:** Similar prompts still produce different results when participants do not understand or control generation conditions.
- **Approach:** Hold the task stable, vary one generation or model condition, predict effects before execution, repeat trials, and interpret differences in accessible mechanistic terms.
- **Artefacts:** protocol document; locked request and input; prediction table; run log with configurations, repeated outputs, scores, and observations; selected excerpts or screenshots; concise interpretation; uncertainty statement; and one practical control rule.

Teacher Setup

Mandatory:

- one shared experiment pack with a fictional briefing, locked request, two one-variable configuration cards, and scoring criteria
- an environment that exposes at least one generation or model condition and allows repeated runs
- verified configuration metadata; teachers should know which settings the interface actually controls and which remain hidden
- tokenisation or character-level micro-tasks plus a decomposition or tool-supported repair
- a completed sample run log and comparison table
- protocol, prediction, configuration, scoring, and evidence-capture templates for the challenge kickoff
- demonstration setup for the recurring agentic coding case, if used
- a prepared-output fallback containing at least three authentic runs per condition so the experiment does not depend on live API or platform access

Configurable:

- exact model, notebook, playground, or interface
- **primary variable:** sampling setting, model variant, or reasoning condition; only one should change in the main comparison
- number of repetitions beyond the minimum
- **scoring method:** anchored qualitative ratings, simple counts, or a small descriptive plot
- work mode and group size
- domain of the fixed task and participants' transfer tasks
- scaffolding depth and peer-feedback format