

Configuration shapes model output

Generation settings influence behavior at inference time.

Settings influence

- Variation between responses
- Response length
- Where generation stops
- How consistently the model follows an assigned role

Key message

Configuration changes model behavior without retraining the model.

Temperature controls variation

“Temperature changes how strongly the model favors high-probability tokens.”

Lower temperature

More conservative choices

More consistent wording

Useful for focused tasks

Higher temperature

More varied choices

More unexpected wording

Useful for idea generation

Typical task fit

1. Extraction: low temperature
2. Factual summarization: low to moderate temperature
3. Creative generation: moderate to high temperature

Remember

Temperature changes variation rather than factual knowledge.

Top-p limits the candidate pool

Top-p uses nucleus sampling.

The model

- Ranks possible next tokens by probability.
- Builds the smallest candidate set whose cumulative probability reaches top-p.
- Selects from that candidate set.

Effect

- Lower top-p: Restricts generation to a narrower candidate set
- Higher top-p: Allows a broader range of possible tokens

Exam focus

Temperature changes randomness across token probabilities. Top-p limits selection to a cumulative probability range.

Maximum output tokens and stop conditions

Generation can end at a natural completion, a stop sequence, or the token limit.

Maximum output tokens

- Sets an upper limit on generated output
- Helps control latency and cost
- Can truncate an unfinished response

Example stop sequence

END_OF_SUMMARY

Stop conditions

- Boundary: End generation when the model reaches a specified sequence or condition
- Workflows: Useful for structured boundaries and multi-part workflows
- Caution: Require careful selection to avoid stopping too early

Remember

A token limit is a ceiling, not a required length, and it does not map exactly to a word limit.

Determinism and creativity form a spectrum

Requirement	Preferred configuration
Repeatable structured extraction	Conservative sampling and strict output instructions
Faithful summarization	Conservative sampling and a clear length constraint
Brainstorming alternatives	More permissive sampling
Marketing concepts	Moderate variation with explicit brand constraints

Callout

Configuration can improve consistency, but generated output remains probabilistic.

System instructions define persistent behavior

The application supplies system instructions separately from the user's request.

Example

"You are a customer-support assistant. Answer only from the supplied product documentation. If the documentation does not contain the answer, say that you cannot verify it. Return no more than three short paragraphs."

System instructions can define

- The model's role
- Behavioral boundaries
- Required response style
- Rules for using supplied information
- Conditions for escalating or declining a request

Exam trap

A system instruction guides model behavior. It does not train or permanently modify the foundation model.

Settings by task

Task	Temperature or top-p	Output limit	Instructions
Extraction	Conservative	Enough for the schema	Specify fields, format, and missing-value behavior
Summarization	Conservative to moderate	Match the desired summary length	Preserve important facts and avoid adding unsupported details
Creative generation	Moderate to high	Allow room for alternatives	Define audience, tone, constraints, and desired format

Practical sequence

- 1. Define the required output.
- 2. Start with conservative settings.
- 3. Evaluate representative examples.
- 4. Adjust one parameter at a time.
- 5. Validate the result against task-specific criteria.

Remember

There is no universally correct temperature or top-p value.

Demonstration: conservative versus creative

Run one prompt twice with the same model and system instructions. Change only the sampling configuration.

What students should observe

- Differences in wording
- Differences in originality
- Differences in structure
- Whether either response violates the prompt's constraints
- Whether higher variation improves the result for this task

Key message

Use a model that exposes temperature or comparable sampling controls. Interfaces and supported parameters vary by model.

Demonstration prompt

The prompt permits creative variation while keeping objective format constraints.

User prompt · “Write a launch announcement for a new voice-support feature. Include a headline, a description of no more than 80 words, and a one-sentence call to action.”

System instruction

“You are a marketing writer for an enterprise software company. Write clearly for IT professionals. Avoid exaggerated claims. Follow the requested format exactly.”

Why this prompt works

- It permits meaningful creative variation.
- It contains objective format constraints.
- It avoids reliance on external facts.
- Both responses can be evaluated against the same criteria.

Keep constant

Use the same model, system instruction, and user prompt on both runs.

Demonstration setup

Open the model playground in Microsoft Foundry and keep every variable except sampling unchanged.

Setup

- Select an available general-purpose language model that supports temperature controls.
- Start a new chat session.
- Enter the system instruction and user prompt.
- Set the maximum output to approximately 150 to 200 tokens.

Between runs

- Keep the model, prompt, system instruction, and output limit unchanged.
- Record the model name and model version for reproducibility.
- If a seed parameter is exposed, leave it at its default.
- The demonstration concerns practical variation rather than guaranteed reproducibility.

Instructor note

If the interface exposes a seed parameter, leave it at its default.

Run 1: conservative configuration

Apply a low-variation configuration, then submit the same prompt.

Suggested configuration

- Temperature: 0.2
- Top-p: 1.0
- Maximum output tokens: 180
- Stop sequence: none

Steps

- Apply the conservative configuration.
- Submit the prompt.
- Save or copy the response.
- Run it a second time if time permits.
- Note any changes between repeated responses.

Expected characteristics

Direct headline, familiar wording, close format adherence, and limited variation between runs.

Run 2: creative configuration

Begin a new chat. Change only the temperature.

Suggested configuration

- Temperature: 1.0
- Top-p: 1.0
- Maximum output tokens: 180
- Stop sequence: none

Steps

- Begin a new chat to avoid carrying over the first response.
- Keep the model, system instruction, and user prompt unchanged.
- Change only the temperature.
- Submit the prompt and save the response.
- Repeat once if time permits.

Expected characteristics

More varied headlines, less predictable phrasing, greater differences between runs, and possible weaker constraint adherence.

Compare the responses

Evaluation question	Conservative run	Creative run	Notes
Did it include all three required elements?			
Did it stay within 80 words?			
Was the wording clear?			
Was the headline distinctive?			
Did it avoid unsupported claims?			
Would you publish it without editing?			
Discussion questions			
1. Which response followed the instructions more closely?			
2. Which response produced the stronger headline?			
3. Did increased variation improve the useful output?			
4. Which configuration would suit a regulated support message?			
5. Which configuration would suit an early brainstorming session?			
Evaluate the task			
Higher variation is useful only when it improves the result for this task.			

Optional top-p extension

Repeat the conservative run while changing top-p and holding temperature constant.

Suggested configuration

- Temperature: 1.0
- Top-p: 0.2
- Maximum output tokens: 180

Compare with

- Creative run: Temperature 1.0 and top-p 1.0
- This run: Temperature 1.0 and top-p 0.2

Teaching point

A lower top-p restricts the candidate token pool. Change top-p while holding temperature constant.