How temperature and top_p affect AI responses

Both temperature and top_p (nucleus sampling) are decoding hyperparameters that control the creativity, randomness, and determinism of the generated responses.

Temperature

The hyperparameter temperature controls randomness of token selection. It adjusts the sharpness of the probability distribution over possible next tokens. Its range typically varies from 0.0 to 2.0.

Zero temperature (0.0): Always picks the most likely word.
Low temperature (from 0.1 to 0.3): More deterministic and focused. Tends to pick the most likely word.
Medium temperature (from 0.7 to 1.0): Balances coherence and creativity.
High temperature (more than 1.0): More chaotic or creative. Risk of nonsense.

Examples:
Prompt: “The sky is…”

Temperature	Response
0.1	"The sky is blue."
1.0	"The sky is a canvas of pastel dreams, shifting moods like ocean tides."

top_p (nucleus sampling)

The hyperparameter top_p (nucleus sampling) limits token selection to a dynamic probability mass. Instead of picking from all tokens, it filters the vocabulary to the smallest set of tokens whose cumulative probability is ≥ top_p. Its range goes from 0.0 to 1.0.

Low top_p (0.1): Only very likely tokens considered. Safer, less creative.
High top_p (from 0.9 to 1.0): Wider pool. More diverse, creative outputs.

If a token distribution is:
"blue" (0.6), "grey" (0.2), "clear" (0.15), "stormy" (0.05)
Total probability = 0.6 + 0.2 + 0.15 + 0.05 = 1.0

top_p	Eligibility

0.6	Only "blue" is eligible. Because 0.6 ≥ top_p.
0.9	"blue" (0.6), "grey" (0.2), and "clear" (0.15) are eligible. Because 0.6 + 0.2 + 0.15 = 0.95, which is ≥ top_p. Notice that 0.6 + 0.2 = 0.8, which is < top_p.