Temperature and top-p are key parameters that control the randomness and creativity of AI language model responses. Temperature determines how deterministic or random the model's token selection is: low values (near 0) make the model more predictable by favoring high-probability tokens, while higher values increase randomness and creativity by giving lower-probability tokens a greater chance of being selected. This allows users to control whether outputs are factual and consistent or imaginative and varied.

Top-p, also known as nucleus sampling, affects which subset of tokens the model considers when generating the next word. It filters tokens to include only those whose cumulative probability reaches a specified threshold (p). Low top-p values restrict the model to high-probability tokens, producing safe and coherent outputs, whereas higher top-p values allow a broader range of tokens, increasing diversity and creativity. Together, temperature and top-p let users balance predictability and novelty in AI-generated responses.