Effects of *temperature* and *top_p*

The temperature and top_p parameters influence how random and creative responses are from a language model like Claude through Amazon Bedrock. The temperature value, which ranges from 0 to 1, sets the level of randomness in the model's predictions. A lower temperature, such as 0.1, leads to more focused, predictable, and deterministic responses. This setting works well for factual queries. On the other hand, a higher temperature, like 0.9 or 1.0, adds more randomness and variety to the outputs, allowing the model to be more creative and open-ended. This is helpful for brainstorming or storytelling.

The top_p parameter, also known as nucleus sampling, determines how the model selects by limiting the probability mass for the next token. When top_p is set to 0.9, the model chooses from the smallest group of possible next words that together have a cumulative probability of 90%. This helps strike a balance between coherence and variety. A lower top_p, for instance, 0.3, makes the model focus on more likely outputs. In contrast, a higher top_p, like 1.0, permits sampling from the entire probability distribution. These parameters together provide developers with control over the creativity, determinism, and risk of the model's responses in different situations.