How Temperature and Top_p affects AI responses.

By: Abdulazeez Bright Abu

Large language models (LLMs) generate text by predicting likely sequences of words based on input prompts. Two of the most important parameters that control the diversity and creativity of model outputs are 'temperature' and 'top_p'. Understanding these parameters allows users to better tailor LLM responses for their specific use case.

Temperature

The temperature parameter controls the amount of randomness in the model's output. A low temperature (near 0) makes the model more conservative, always picking the most likely next word. This results in more predictable, repetitive, and deterministic responses. A higher temperature (up to 1 or beyond) increases randomness, making the model more creative or diverse. However, very high values may cause incoherent or off-topic outputs.

Top_p (Nucleus Sampling)

Top_p controls diversity using a probabilistic approach. Instead of sampling from the full set of possible words, the model only considers the smallest set whose cumulative probability exceeds top_p (e.g., top_p=0.9 covers the top 90% most probable words). A low top_p value makes the output focused and conservative, while a higher top_p includes more possibilities for creative or unexpected responses. Adjusting top_p helps control the trade-off between accuracy and creativity.

Combined Effects

In practice, temperature and top_p both control randomness and creativity. They can be fine-tuned together: lowering both produces highly reliable and repetitive outputs, while increasing either can result in more surprising, diverse, or creative

text. The optimal values depend on the application context—customer support may favor lower values, while brainstorming or storytelling may benefit from higher ones.