How temperature and top-p affect AI responses

In AI models, temperature and top-p are two settings that control how predictable or creative the responses are. In my project, I hooked it up to streamlit, and in their UI, they feature these two options, while you get more options within the AWS platform. I think this shows that they are really important in determining the final responses. Temperature adjusts how confidently the model picks the next word: a low value (like 0.2) makes the output more focused, which is ideal when accuracy is the priority. A higher temperature (closer to 1.0) increases randomness, which can lead to more varied or creative responses, but also a higher chance of errors or unexpected output, so this would work well for creative writing prompts.

Top-p,works by narrowing down the choices the model considers the most likely options. For example, with a top-p of 0.9, the model will only sample from the smallest set of words whose probability is extremely high. A low top-p limits diversity and favors safe, high-confidence choices; a high top-p allows for more variation. Using them together allows you to control the precision & creativity of the response.

For me, in my project, adjusting the temperature and top-p varied the length, and factual nature of the language used in the responses. With a very low temperature & top p , the response came out a little robotic and simplistic, almost like through a translation.