Temperature controls how random or creative the AI's responses are.

* A low temperature that’s close to 0, makes the AI give more precise, predictable, and consistent answers — ideal for technical or factual queries like in in heavy machinery chatbot.
* A high temperature close to 1, allows the AI to produce more diverse, imaginative, and sometimes unexpected answers, which is better in open-ended conversations but may affect how factual or precise the responses are.

Top\_p which can also be referred to as nucleus sampling controls randomness differently by limiting the next word choices model’s can use to the most probable tokens whose combined probability exceeds a threshold *p*.

* A low top\_p will limit the AI to highly probable words, which will lead to a more conservative or controlled answers.
* A high top\_p allows a broader range of possibilities, increasing diversity and creativity in the chatbot responses.

In the chatbot, adjusting temperature and top\_p allowed me to balance between precise technical answers and more natural conversational flow. For the heavy machinery queries, a low temperature combined with a moderate top\_p was best for maintaining accuracy while keeping responses natural and understandable.