Temperature

"Temperature" controls the level of creativity in a model when it generates outputs. Temperature sets how likely the model is to choose less probable next words when generating text.

Effect of low vs high temperature

- Low Temperature (e.g. 0 or close to 0): The model is more deterministic. The model will likely choose the word with the highest probability. This results in output words and texts that are more predictable. This setting is good for tasks that need accuracy, e.g. summarisation, translation, and a question and answer chatbot.
- High Temperature (closer to 1 or higher): The model is more creative. The model gives less probable words a greater chance of being selected. With creativity, this might lead to unexpected or nonsense outputs, known as hallucination. This setting is good for brainstorming and creative writing.

Top_p

“Top_p” is short for “nucleus sampling”. What “top_p” involves is

1. The model searches through a list of words, to identify the most probable next word.
2. The model keeps adding words to a "candidate pool/list" until the sum of their probabilities reach the chosen top_p value (e.g., 0.9).
3. From only this “candidate pool” of words, the model picks the next word, and ignores all the less likely words outside the pool.

Effect of low vs high top_p

- Low top_p (e.g., 0.05) tells the model to focus on a smaller candidate pool of words, which are more probable. So the model outputs words and texts which are more predictable and focused.
- High top_p (e.g., 0.95) tells the model to focus on a larger candidate pool of words, which could be less likely to appear in usual text. So the model outputs words and texts which are more diverse and potentially more creative in its writing.

Usually people set either the ‘temperature’ or ‘top_p’, not both at the same time, because these 2 settings overlap in functionality with each other.