

Temperature: low temperature gives more consistent, repetitive and predictable output while high temperature gives more creative, diverse and unexpected output.

Top_p: low top-p gives narrow and deterministic output, while high top-p gives wide, diverse and creative output

Example on usage:

Low Temperature and low Top-p are better for helping customers in chats, while high temperature and high Top-p are better for enhancing your storytelling.