Affect of temperature:

In AI models, particularly in natural language models, **temperature** controls the randomness of responses.

- **Low Temperature (e.g., 0.2)**: The model becomes more deterministic, choosing high-probability words more often. This is ideal for factual or predictable responses.
  Example: If asked, "What's the capital of France?" the model will consistently answer "Paris."

- **High Temperature (e.g., 0.8 or higher)**: The model generates more diverse and creative responses, as it selects from a wider range of possibilities. However, this can sometimes lead to less coherent or less factual outputs.
  Example: If prompted to "Write a poem about the ocean," the model may produce varied, imaginative verses each time.

Key takeaway: Use *low temperatures for precise and factual tasks, and high temperatures for creative and exploratory tasks*.

Affect of top_p:

In AI models, **top_p** (nucleus sampling) determines the probability threshold for choosing the next word, focusing on a subset of the most likely options.

- **Low top-p (e.g., 0.3)**: The model considers only the top 30% of probable outcomes, leading to more focused and predictable responses.
  Example: If asked, "What is the largest mammal?" the model is likely to consistently reply, "The blue whale."

- **High top_p (e.g., 0.9 or higher)**: The model considers a broader range of possible outcomes, including less likely ones, making responses more diverse and creative.
  Example: If prompted, "Tell me an interesting fact about space," it might generate varied answers like "Neutron stars can spin 600 times per second" or "There's a giant cloud of alcohol in the Milky Way."

Key takeaway: Use *low top_p for accuracy and consistency, and high top_p for creativity and variety in responses*.