Explanation of Temperature and Top P

Temperature

Temperature is a hyperparameter that controls the randomness of the model's output. It scales the logits (unnormalized probabilities) before applying the softmax function.

- **Low Temperature (e.g., 0.1 - 0.3)**: The model becomes more deterministic and confident. It tends to choose the most likely next token. This is useful for tasks requiring factual accuracy, code generation, or precise answers.
- **High Temperature (e.g., 0.7 - 1.0)**: The model becomes more creative and diverse. It flattens the probability distribution, allowing less likely tokens to be selected. This is useful for creative writing, brainstorming, or generating varied responses.

Top P (Nucleus Sampling)

Top P, or Nucleus Sampling, is an alternative to temperature for controlling randomness. Instead of sampling from the entire vocabulary, the model samples from the smallest set of top-ranked tokens whose cumulative probability exceeds the threshold P.

- **Low Top P (e.g., 0.1)**: The model considers only the very top tokens (the "nucleus" of probability). This restricts the output to high-confidence tokens, reducing the chance of nonsensical or irrelevant text.
- **High Top P (e.g., 0.9)**: The model considers a wider range of tokens, allowing for more diversity while still filtering out the very low-probability tail.

Interaction

Often, both parameters are used together. A common practice is to keep one fixed (e.g., Top P = 1.0) and adjust the other, or tune both to balance coherence and creativity. For a document query system, we typically want lower values (e.g., Temperature 0, Top P 0.9 or lower) to ensure the model sticks to the provided context and avoids hallucination.