Temperature

In AI, "temperature" is a parameter that controls the randomness of an output, like text generated by a language model. A low temperature (closer to 0) makes the AI more focused and deterministic, resulting in predictable and accurate responses. A high temperature (closer to 1 or higher) introduces more randomness, leading to more creative, varied, and potentially less predictable or surprising outputs. The ideal setting depends on the task; low is for factual tasks, while high is for creative tasks.

How temperature works

- **Low temperature (e.g., 0.1–0.5):**

  The model is more likely to choose the most probable words or tokens, leading to a more conservative and consistent output.

- **High temperature (e.g., 1.0–2.0):**

  The model considers a wider range of less probable words or tokens, resulting in more diverse, creative, and unexpected outputs.

Use cases for different temperatures

- **Low temperature:**

  Best for tasks that require accuracy and consistency, such as summarizing data, answering factual questions, or generating code.

Understanding Top-p sampling

Top-p sampling dynamically adjusts the number of words considered for each prediction based on the probability distribution. It aims to strike a balance between maintaining the coherence of high-probability choices and allowing for diversity in the generated text.

Key aspects of Top-p sampling include:

Probability Threshold: Uses a cumulative probability (p) as the cutoff for word selection.

Dynamic Vocabulary: The number of words considered varies for each prediction.

Tail Cutting: Effectively eliminates low-probability words from consideration.

Adaptability: Adjusts to the confidence of the model in different contexts.

Balancing Act: Seeks to balance between quality and diversity in generated text.