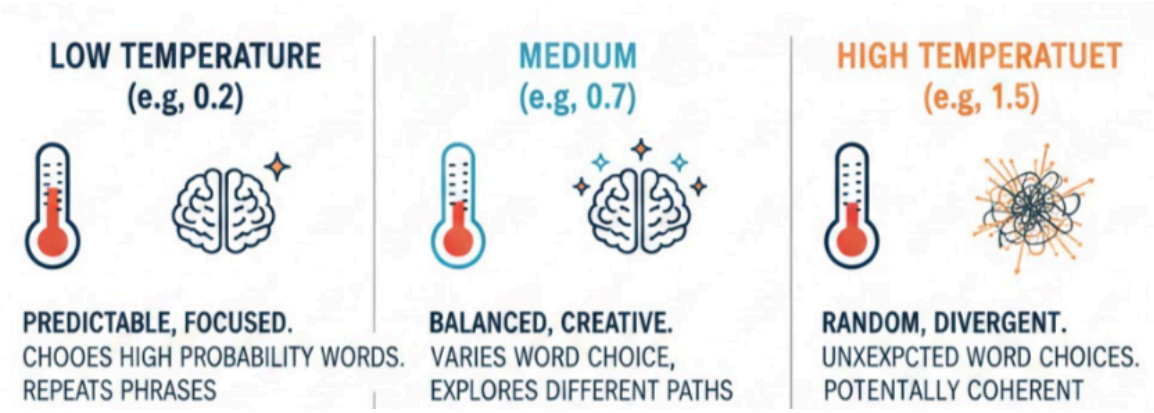
LLM Generation Parameters: Temperature and Top_P

🔥 Temperature

Controls how random or creative the model's responses are.

Temperature adjusts the sharpness of the model's probability distribution. Lower values make the model strongly prefer the highest-probability tokens, while higher values flatten the distribution so more tokens have similar probability.

- A **low temperature** (0.0–0.3) makes the model **more predictable** and **precise**. It picks the most likely next words.
- A **medium temperature** (.5–0.7) adds some variety without being too chaotic.
- A **high temperature** (0.7–1) makes the model **more creative**, sometimes **unpredictable**, and willing to take risks.



🎯 Top-P (Nucleus Sampling)

Controls how many possible next-word options the model considers.

Top-P limits token selection to a dynamic set of top-probability tokens whose cumulative probability adds up to P. Top-P defines a **probability threshold.**

- **Small Top-P (e.g., 0.3)** model chooses from only the most likely tokens → *more deterministic.*
- **Large Top-P (e.g., 0.9–1.0)** → model chooses from a broader set → *more diversity and creativity.*