Temperature and Top_p

Temperature

Temperature adjusts the probability distribution: Higher values (e.g., 1.0) make the output more creative and unpredictable, while lower values (e.g., 0.1) make it more deterministic and focused.

Top_p

Top_p (or nucleus sampling) limits the pool of words considered for the next token to those whose cumulative probability exceeds the value `p`, leading to more constrained and coherent but still varied responses.

Recommended Settings for current use case

For this use case, we would want to prioritize grounded, consistent answers over creativity.

- Prompt validator: `temperature=0`, `top_p=0.1` (strict; don’t change; we only want it to state the category and nothing else).
- Answer generation (recommended): `temperature 0.2–0.4`, `top_p 0.7–0.9` → factual with natural phrasing.
- Feels robotic? Nudge `temperature` up slightly (≤0.5) or `top_p` to 0.85–0.9.
- See drift/speculation? Lower `temperature` and/or `top_p`.
- Avoid `temperature > 0.7` or `top_p > 0.95` — raises hallucination risk even with retrieval.