✅ 20 MCQs on Temperature, Top-k, and Top-p (With Explanations)

**1. What does the temperature parameter control in language models?**

A) Length of output  
B) Randomness/creativity of responses  
C) Model accuracy  
D) Number of tokens

**Answer:** B  
**Explanation:** Temperature adjusts randomness — low temperature makes answers deterministic, high temperature increases creativity.

**2. If temperature = 0, what will likely happen?**

A) Output becomes very random  
B) Output is highly deterministic and repetitive  
C) Model refuses to answer  
D) Model generates longer outputs

**Answer:** B  
**Explanation:** At 0, the model always chooses the highest probability word → deterministic response.

**3. Increasing temperature above 1 generally makes outputs:**

A) More predictable  
B) Shorter  
C) More random and creative  
D) More accurate

**Answer:** C  
**Explanation:** Higher temperature increases sampling randomness → more diverse and imaginative text.

**4. Which parameter limits the model to choosing from the top K probable words?**

A) Temperature  
B) Top-p  
C) Top-k  
D) Max tokens

**Answer:** C  
**Explanation:** Top-k limits token selection to the k most likely words.

**5. If top-k = 1, the model will:**

A) Randomly select from many tokens  
B) Always choose the single most probable token  
C) Refuse to answer  
D) Generate longer text

**Answer:** B  
**Explanation:** With k=1, sampling is restricted to the most likely next word (greedy decoding).

**6. What does top-p (nucleus sampling) control?**

A) Chooses tokens until probability mass reaches p  
B) Always selects top k words  
C) Controls output length  
D) Reduces hallucinations

**Answer:** A  
**Explanation:** Top-p considers the smallest set of words whose combined probability ≥ p (e.g., p=0.9).

**7. If top-p = 1.0, what effect does it have?**

A) No filtering (all tokens considered)  
B) Only one token is chosen  
C) Same as temperature = 0  
D) Makes answers shorter

**Answer:** A  
**Explanation:** At p=1, no probability cutoff is applied; all tokens are eligible.

**8. Which of these best describes temperature vs top-k?**

A) Both control creativity in the same way  
B) Temperature scales randomness; top-k restricts choice to k tokens  
C) Temperature controls length; top-k controls speed  
D) Temperature increases accuracy; top-k increases hallucinations

**Answer:** B  
**Explanation:** Temperature changes distribution sharpness, while top-k restricts candidate pool.

**9. Which setting gives the most deterministic response?**

A) Temperature=1.0, top-k=50  
B) Temperature=0, top-k=1  
C) Temperature=0.9, top-p=0.95  
D) Temperature=2.0, top-k=100

**Answer:** B  
**Explanation:** Zero temperature + top-k=1 → always the most probable word → deterministic.

**10. Setting temperature too high (e.g., 2.0+) often leads to:**

A) More logical answers  
B) Repetitive answers  
C) Nonsensical or chaotic responses  
D) Shorter outputs

**Answer:** C  
**Explanation:** Very high temperature makes the model over-randomized, often producing nonsense.

**11. If top-k = 5, what does the model do?**

A) Chooses randomly from all tokens  
B) Picks only from the 5 most probable tokens  
C) Reduces output length to 5 tokens  
D) Generates 5 different answers

**Answer:** B  
**Explanation:** The next word will be chosen only from the top 5 probable candidates.

**12. Which setting balances creativity and reliability best?**

A) Temperature=0, top-k=1  
B) Temperature=2.0, top-p=1  
C) Temperature≈0.7, top-p≈0.9  
D) Temperature=0.1, top-k=500

**Answer:** C  
**Explanation:** Medium temperature (0.7) + nucleus sampling (p≈0.9) gives creative but still controlled outputs.

**13. What is the difference between top-k and top-p?**

A) Top-k fixes number of tokens; top-p fixes probability mass  
B) Top-k is random; top-p is deterministic  
C) Top-k is for training; top-p is for inference  
D) Both are identical

**Answer:** A  
**Explanation:** Top-k = fixed cutoff in count; Top-p = dynamic cutoff based on probability distribution.

**14. Which combination produces the most creative answers?**

A) Low temperature + top-k=1  
B) High temperature + top-p=0.95  
C) Temperature=0 + top-k=5  
D) Temperature=0.2 + top-p=0.5

**Answer:** B  
**Explanation:** High temperature + wide nucleus sampling → more diverse, creative outputs.

**15. What happens if temperature=0 but top-p=0.9?**

A) Randomness increases a lot  
B) Model still deterministic, since temperature=0 dominates  
C) Model refuses output  
D) Output always longer

**Answer:** B  
**Explanation:** Temperature=0 makes the model always pick the most probable word, regardless of top-p.

**16. Which setting helps prevent nonsensical hallucinations?**

A) Very high temperature  
B) Low temperature with moderate top-p  
C) No top-k filtering  
D) Random top-p

**Answer:** B  
**Explanation:** Controlled randomness (low T, medium p) avoids chaotic word selection.

**17. Why use top-p instead of top-k?**

A) Top-p adapts dynamically to probability distribution  
B) Top-p requires fewer tokens  
C) Top-p increases randomness  
D) Top-p ignores context

**Answer:** A  
**Explanation:** Unlike top-k, nucleus sampling adjusts dynamically to the probability mass.

**18. If top-k is too low (e.g., 2), what is the risk?**

A) Highly creative text  
B) Model outputs repetitive or boring text  
C) Very long answers  
D) Reduced accuracy

**Answer:** B  
**Explanation:** Limiting options too much reduces variety → repetitive responses.

**19. If temperature=1.0, top-p=0.8, what happens?**

A) Model picks from a smaller dynamic set of tokens (80% probability mass), with moderate randomness  
B) Output is always deterministic  
C) Only one token is allowed  
D) Same as greedy decoding

**Answer:** A  
**Explanation:** This is balanced sampling → creative but controlled.

**20. Which of the following is true about temperature, top-k, and top-p?**

A) All three control randomness in different ways  
B) Temperature controls length, top-k controls accuracy, top-p controls speed  
C) Only temperature affects creativity  
D) They cannot be used together

**Answer:** A  
**Explanation:** All three influence diversity/randomness — but by different mechanisms (scaling, cutoff by count, cutoff by probability mass).