Cheat Sheet: LLM Decoding Strategies

Decoding Strategy	What It Is	Deterministic or Stochastic	Advantages	Disadvantages
Greedy Decoding	Always picks the most probable next token.	• Deterministic	• Efficient; Fast.	• Can be repetitive; Predictable.
Multinomial Sampling	Selects the next token based on probability distribution.	• Stochastic	• Promotes diversity; Encourages creativity.	• Can produce irrelevant outputs; Unpredictable.
Top-k Sampling	Narrows choices to top-k tokens, then samples based on probability.	• Stochastic	• Balances predictability and diversity.	• Risk of repetition; Finding optimal k can be challenging.
Top-p (Nucleus) Sampling	Chooses from a subset of tokens whose combined probability reaches or exceeds a threshold p	• Stochastic	• Reduces randomness; Focuses on plausible outputs.	• Complexity in choosing p; May exclude relevant but less probable tokens.
Beam Search	Explores multiple high-probability paths simultaneously.	• Deterministic	• Produces coherent text; Reduces nonsensical outputs.	• Computationally intensive; Slower generation times.
Beam Search with Multinomial Sampling	Combines beam and multinomial sampling for token selection within beams.	• Stochastic	• Coherent yet creative outputs; Enhances diversity.	• Increased complexity; Slower and resource-intensive.
Contrastive Search	Maximizes differences between top candidate sequences.	• Deterministic	• Boosts diversity; Reduces redundancy.	• Higher processing demands; May affect predictability/ coherence.