Topic 7: Decoding Strategies in LLMs

1. An introductory guide by Hugging Face explaining greedy, beam search, top-k, and top-p sampling: https://huggingface.co/blog/how-to-generate
2. A blog post showing real-world implementation of speculative decoding and measured speedups: https://blog.vllm.ai/2024/10/17/spec-decode.html
3. NVIDIA's detailed article explaining speculative decoding on GPU hardware and the draft/target model paradigm: https://developer.nvidia.com/blog/tensorrt-llm-speculative-decoding-boosts-inference-throughput-by-up-to-3-6x
4. A concise write-up explaining speculative decoding: https://www.baseten.co/blog/a-quick-introduction-to-speculative-decoding/