The Degeneration Problem: Model Repetition

Real Output from Greedy Decoding:

"The city of New York is a major city in the United States. The city is known for its diverse culture and the city has many tourist attractions. The city is also home to the city's financial district..."

Problem: "the city" appears 6 times in 4 sentences!

Why? Always picking argmax → same patterns repeated

Solution: Penalize tokens similar to recent context (Contrastive Search)