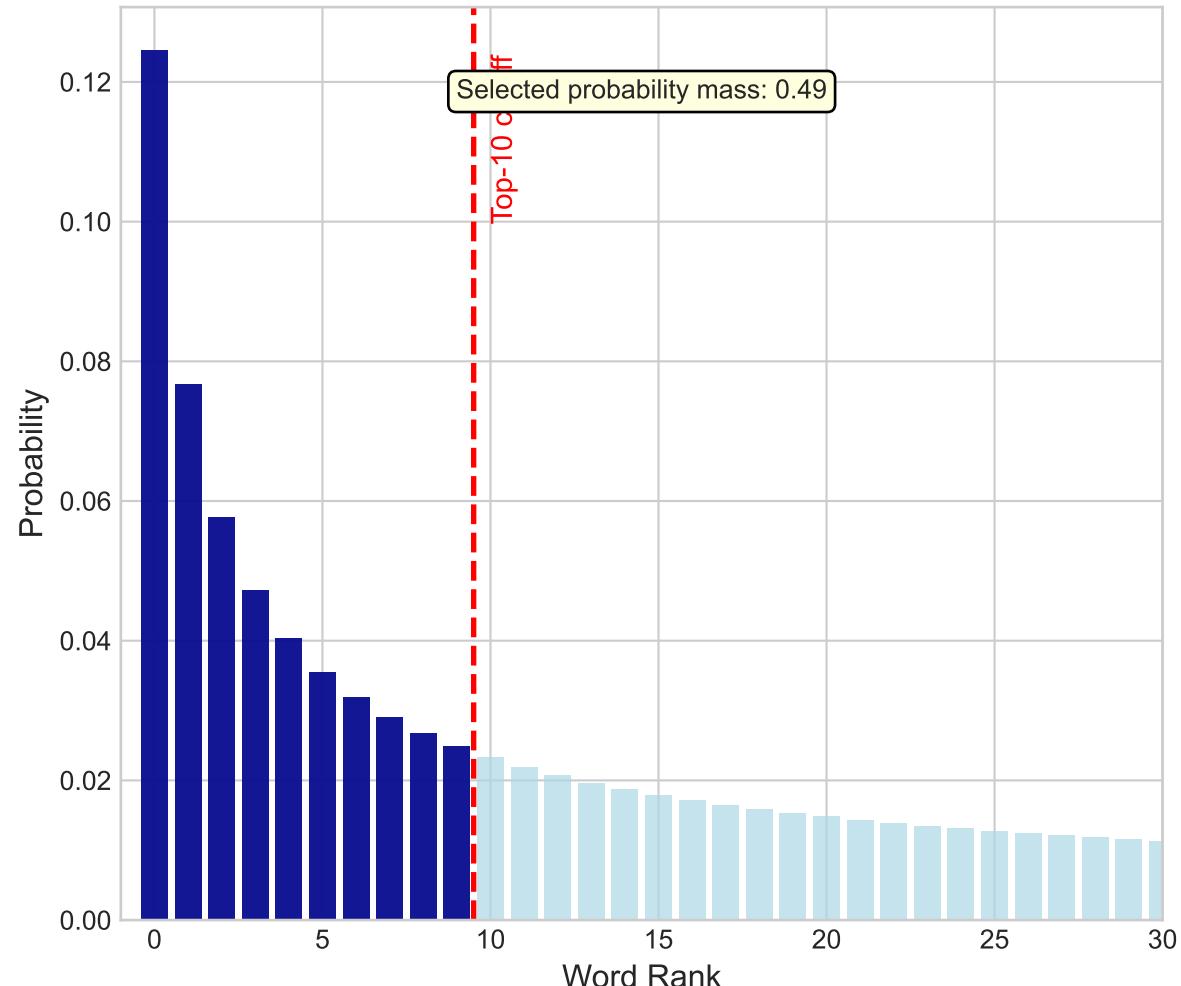
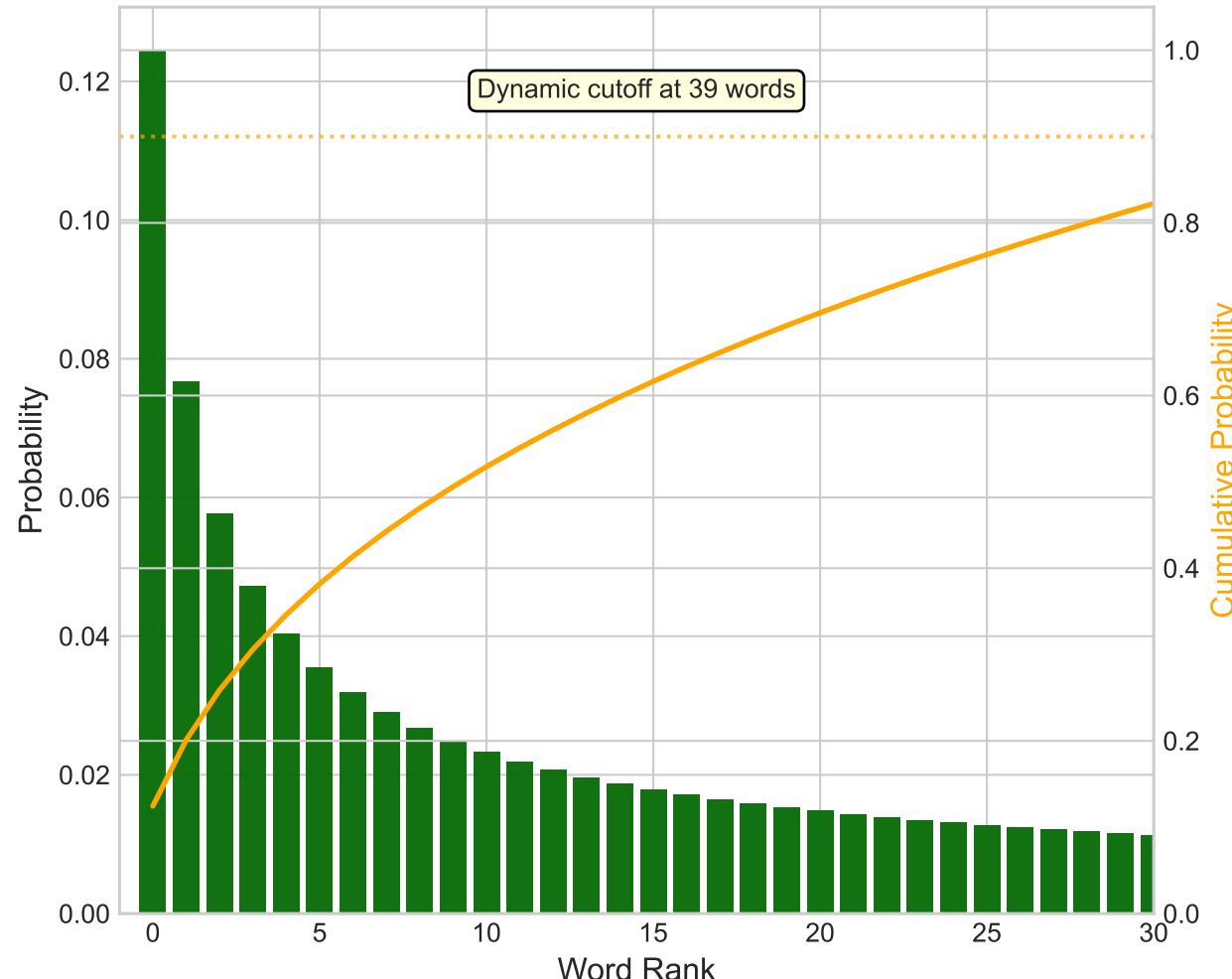


Top-k vs Top-p: Fixed vs Dynamic Vocabulary

Top-k Sampling (k=10)



Top-p (Nucleus) Sampling (p=0.9)



Top-p=0.9 cutoff