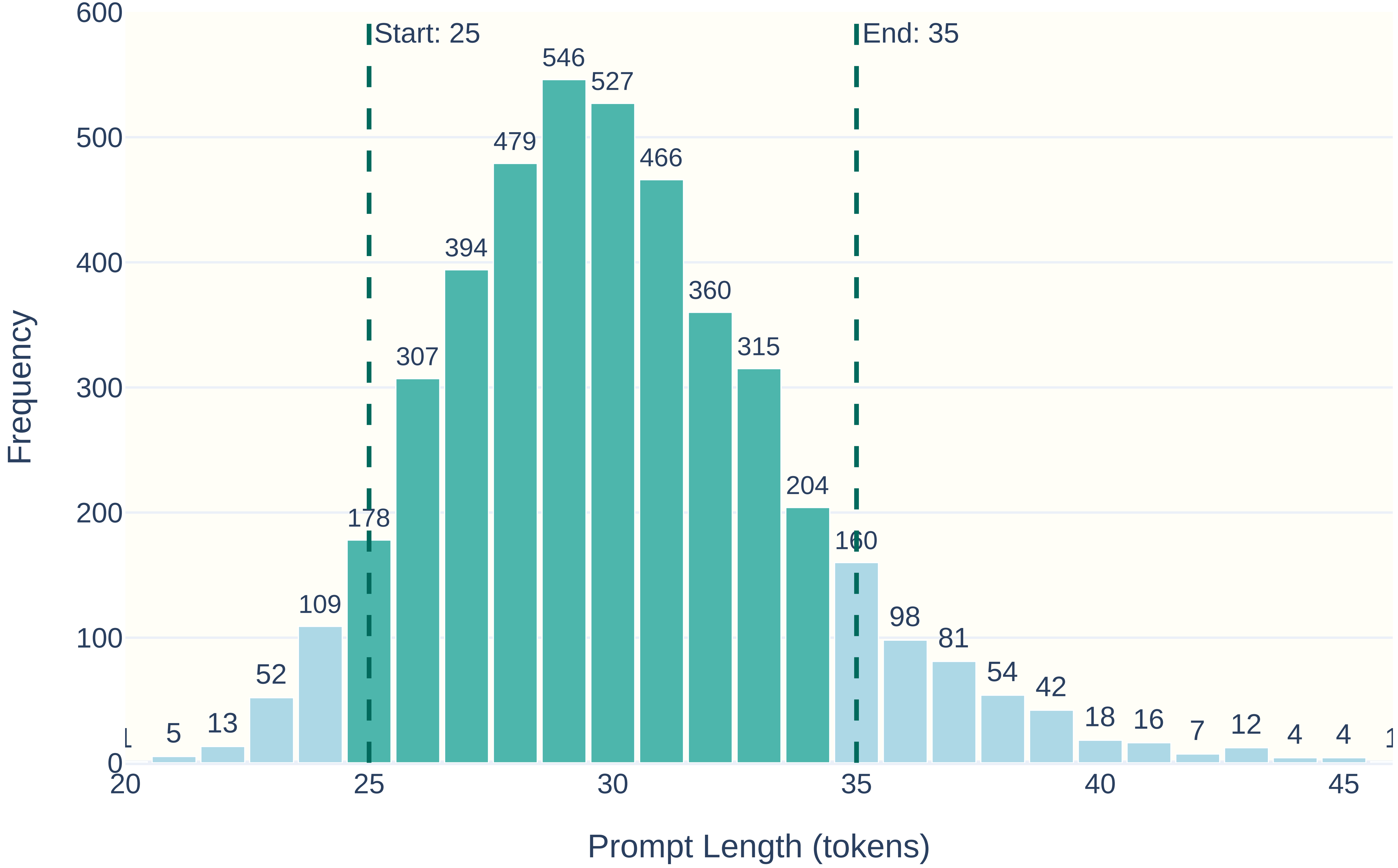
Llama-3-8B-Instruct on harmful dataset [25, 35) (3776/4453 samples, 84.8%)


Start: 25
End: 35
Frequency
Prompt Length (tokens)