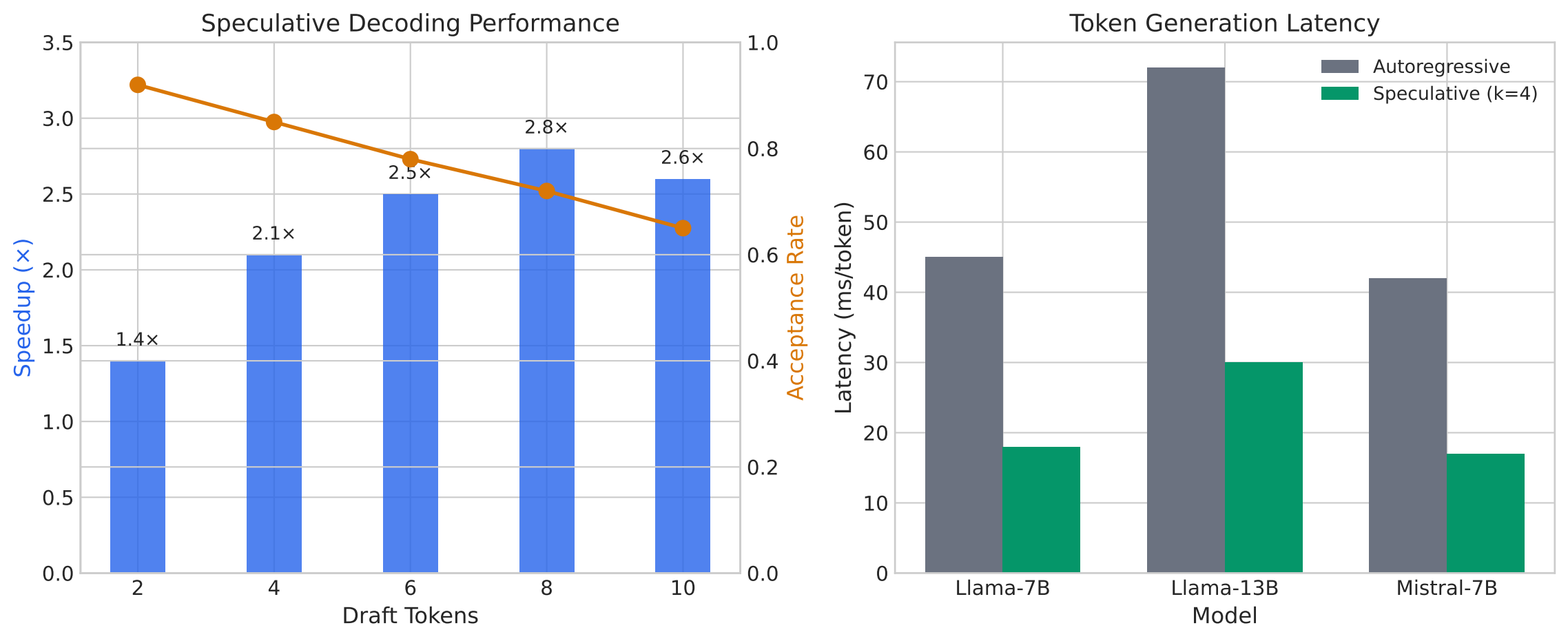
Speculative Decoding Performance
Speedup (×)
Acceptance Rate
Draft Tokens
1.4×
2.1×
2.5×
2.8×
2.6×
Token Generation Latency
Latency (ms/token)
Autoregressive
Speculative (k=4)
Llama-7B
Llama-13B
Mistral-7B
Model