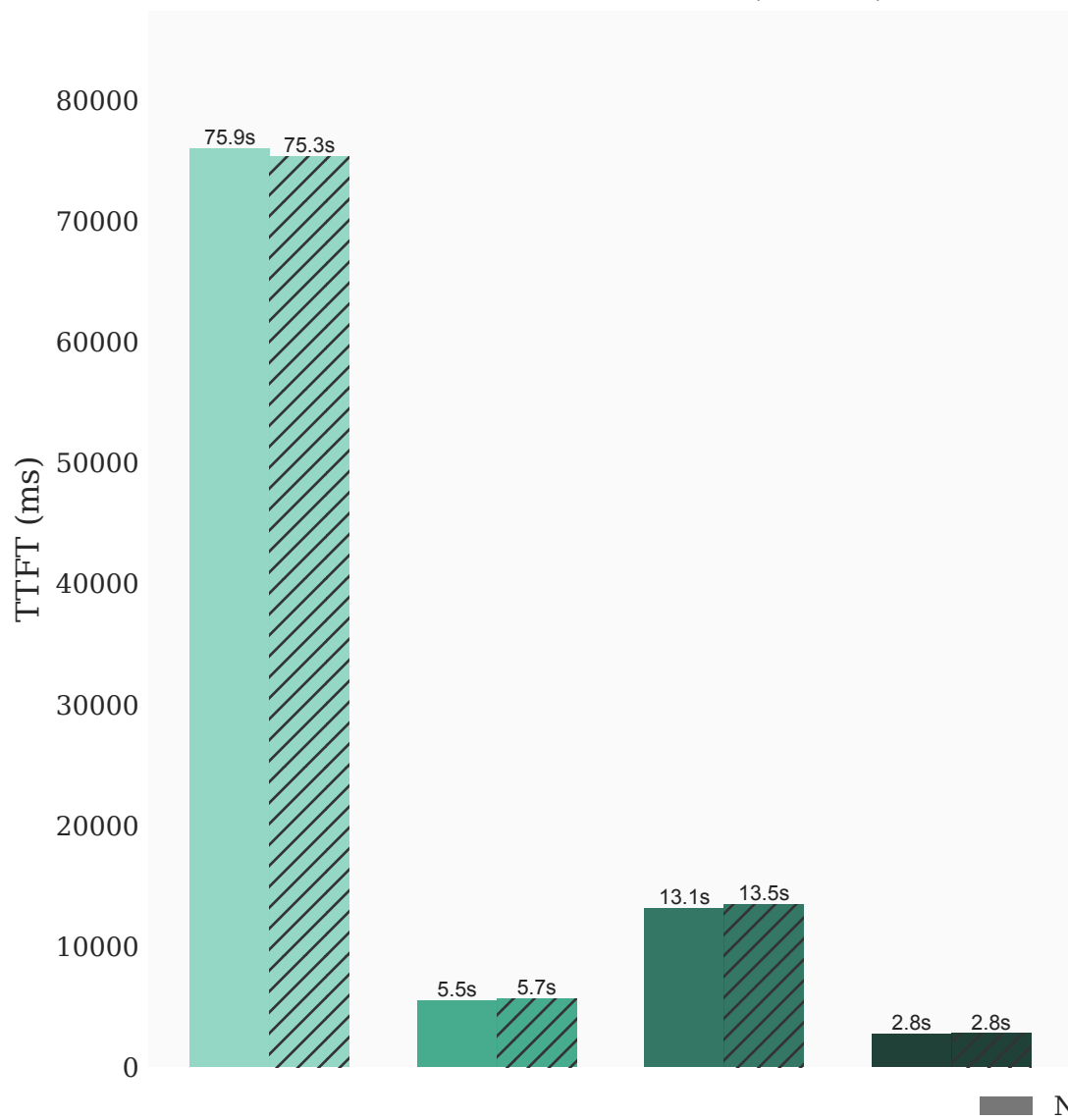
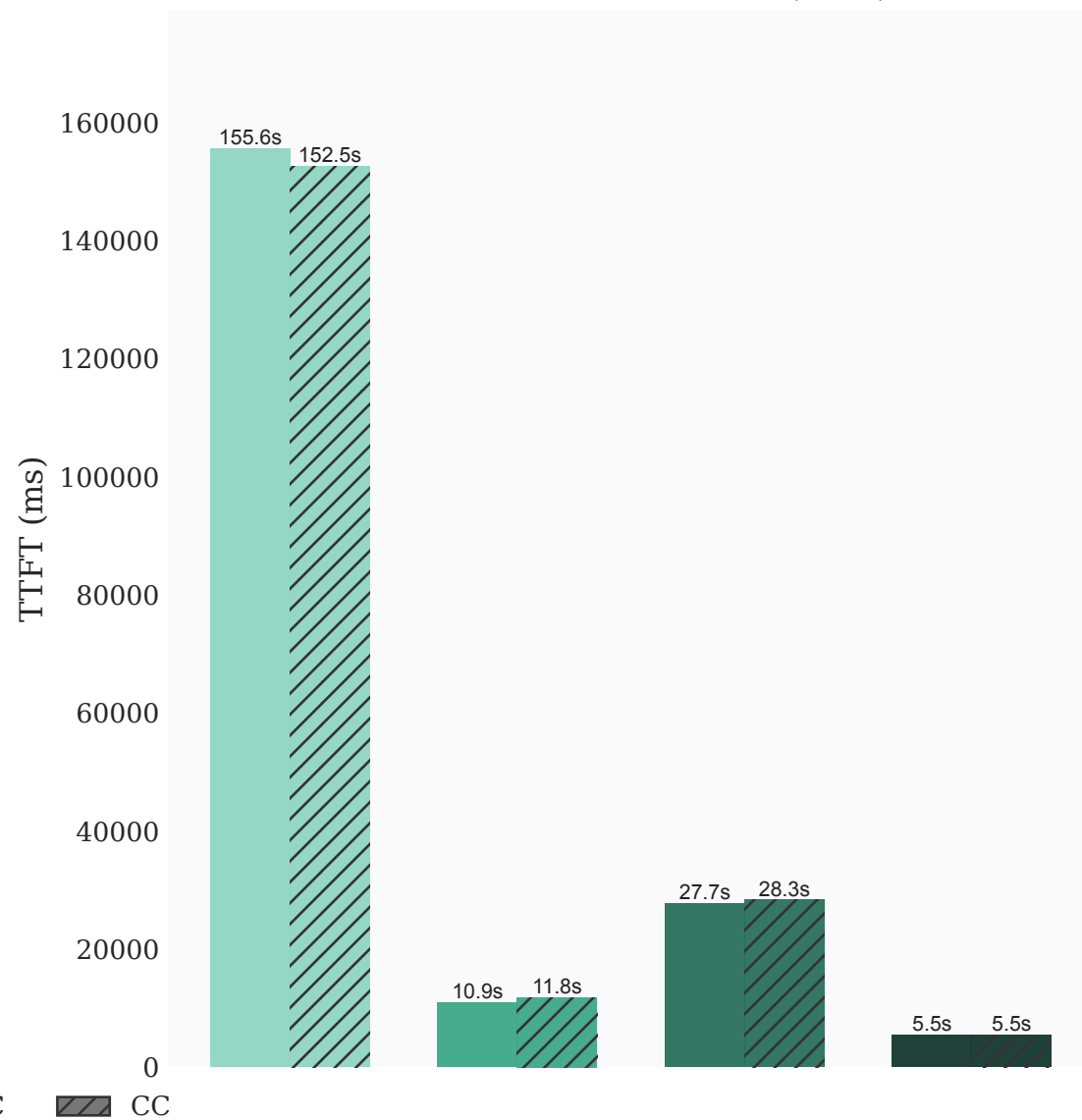


Random (1500 \Rightarrow 250) (100 Request Rate)

Time to First Token (Mean)



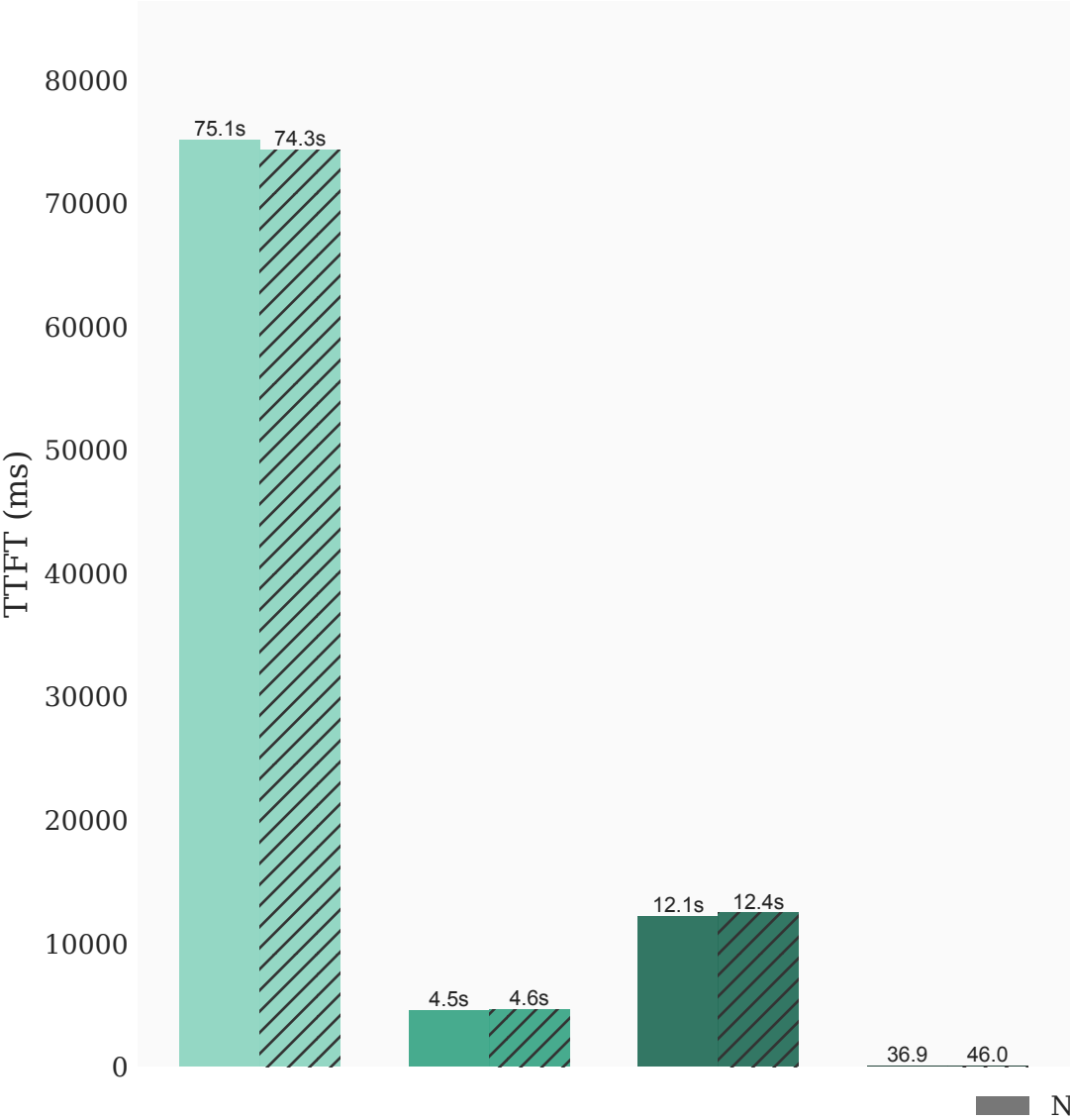
Time to First Token (P99)



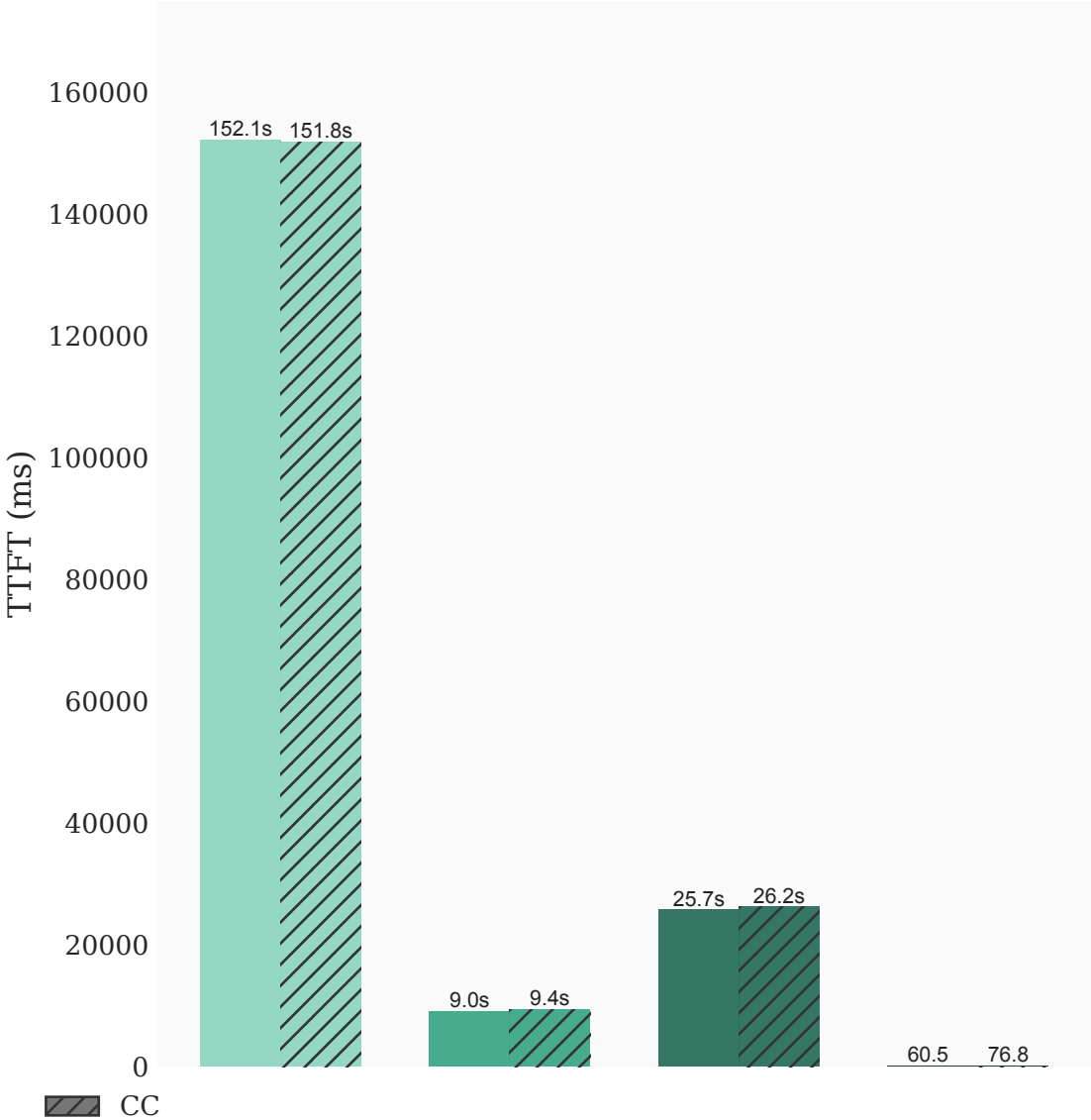
LLama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Random (1500 \Rightarrow 250) (50 Request Rate)

Time to First Token (Mean)



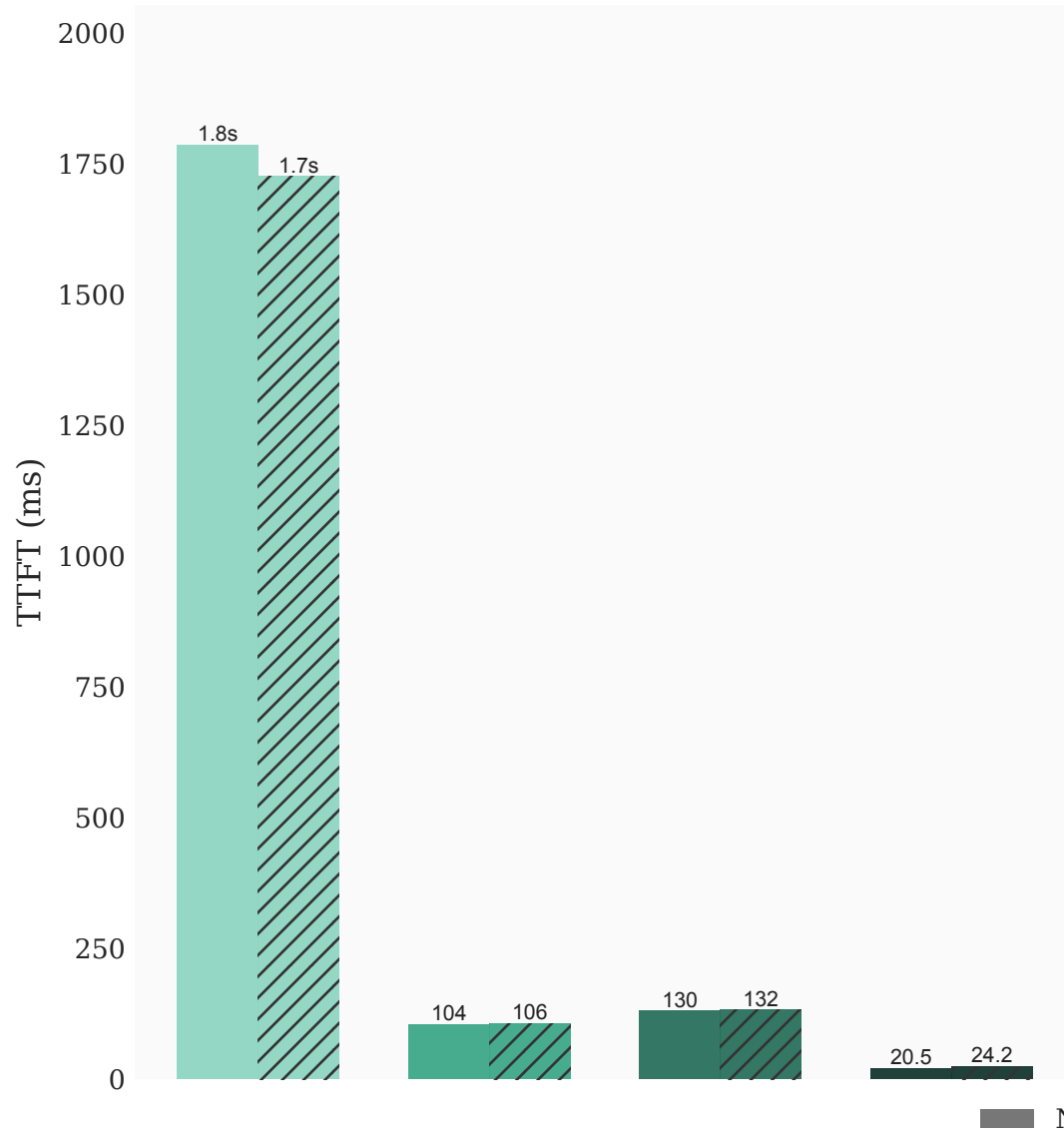
Time to First Token (P99)



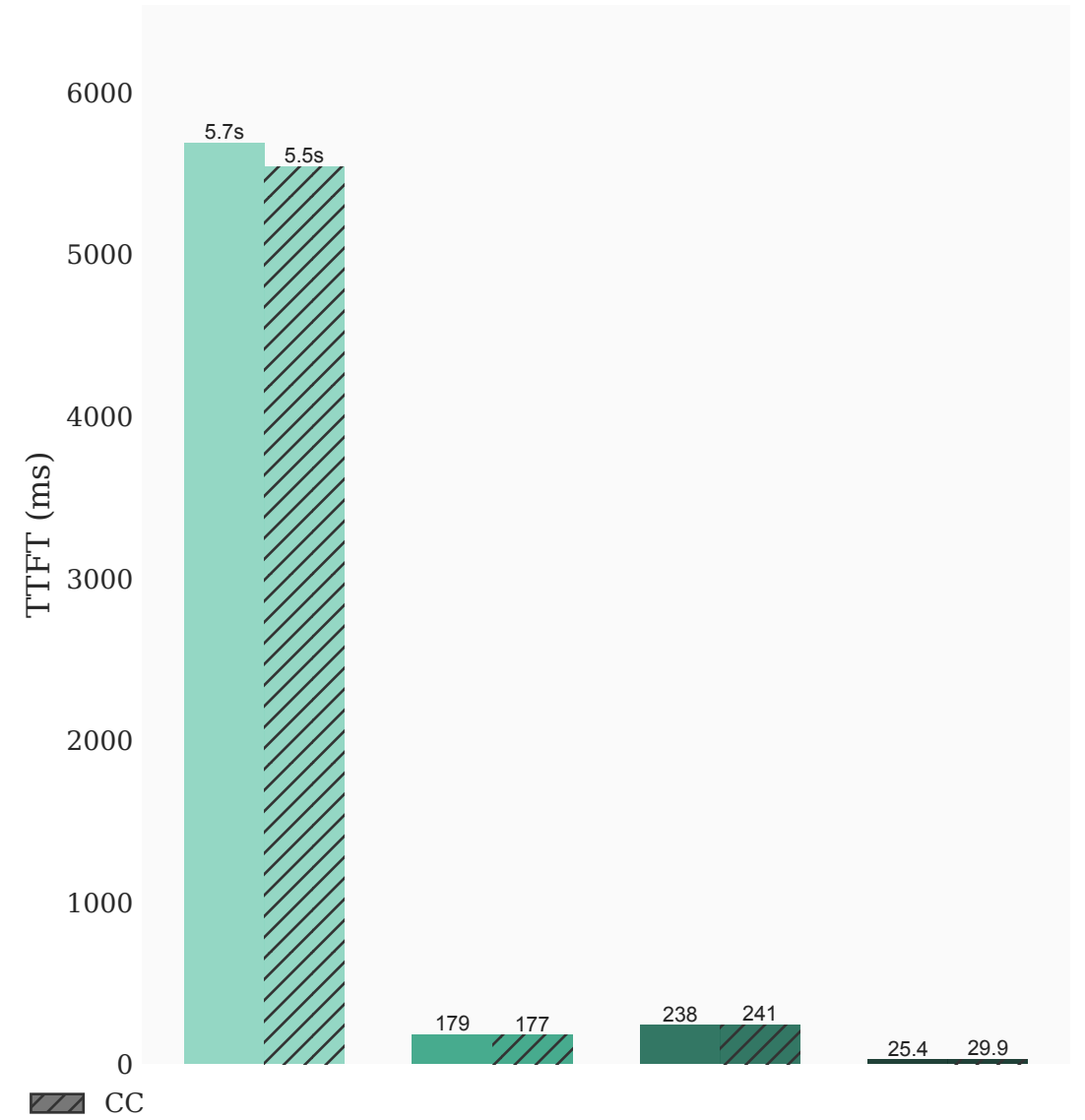
LLama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Random (1500 \Rightarrow 250) (Single Request)

Time to First Token (Mean)



Time to First Token (P99)

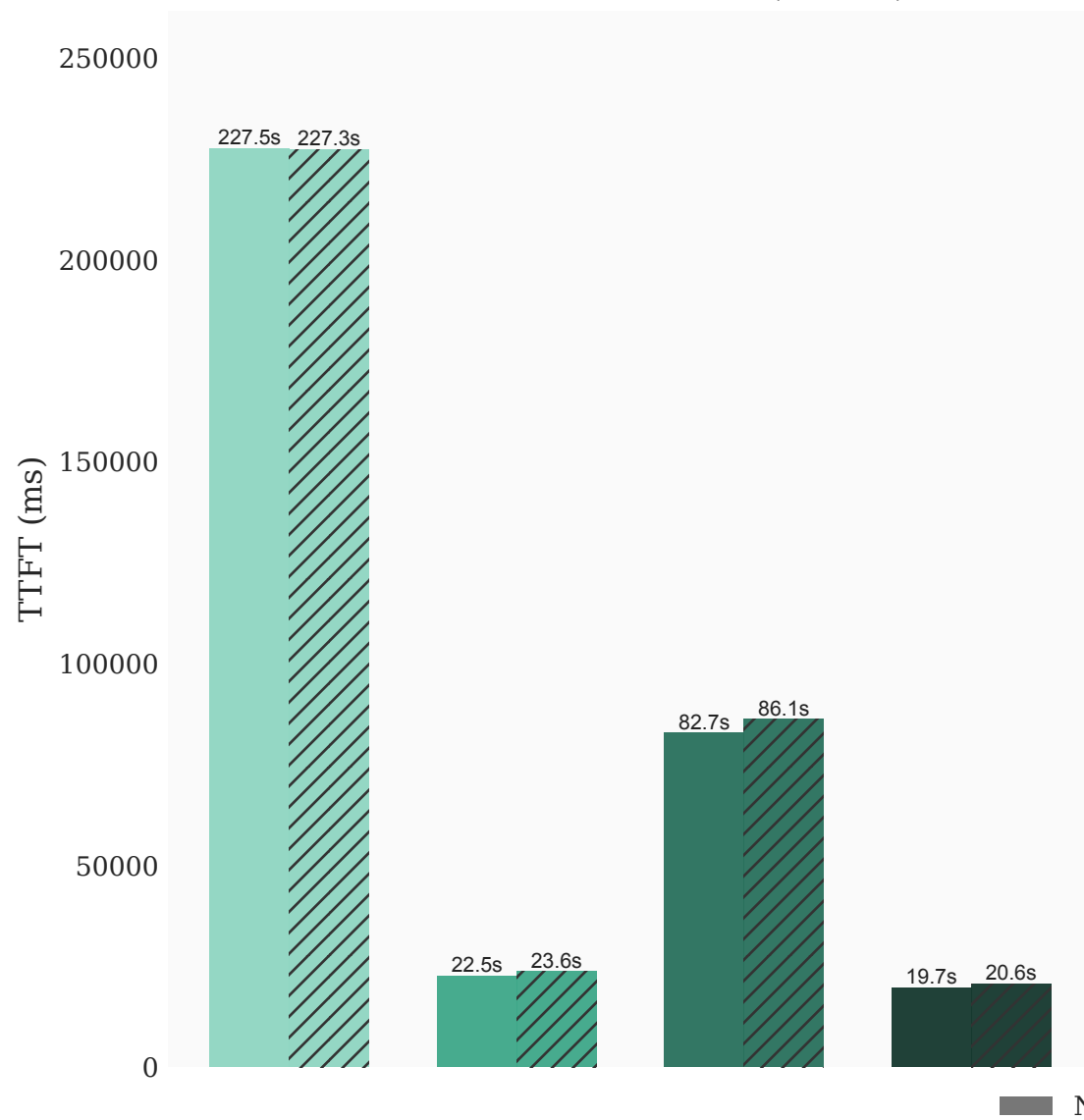


■ No CC ▨ CC

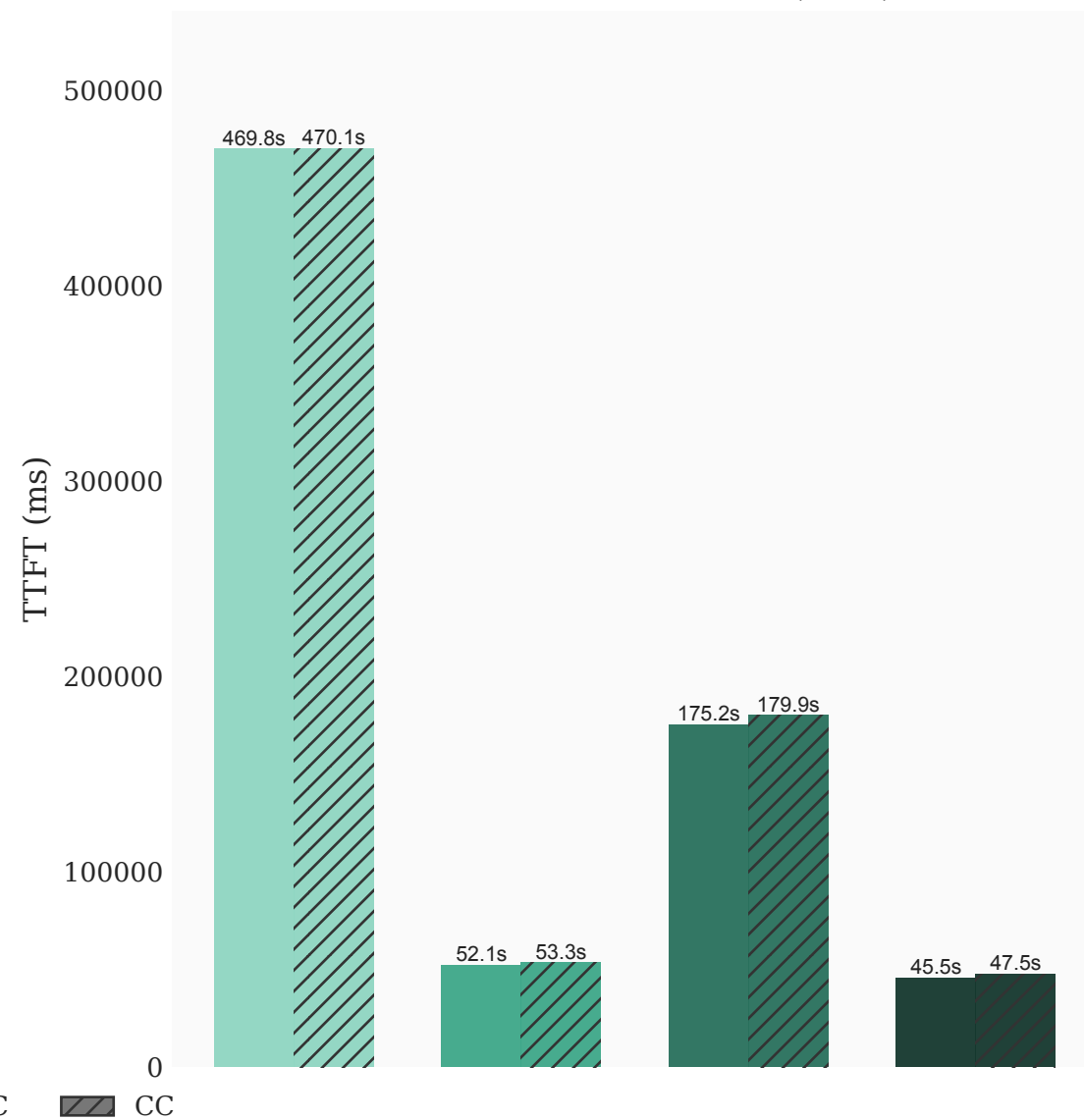
■ Llama 3.3 70B Int4 ■ GPT OSS 120B ■ Mistral 3.1 24B ■ Llama 3.1 8B

Random (4000 \Rightarrow 1000) (100 Request Rate)

Time to First Token (Mean)



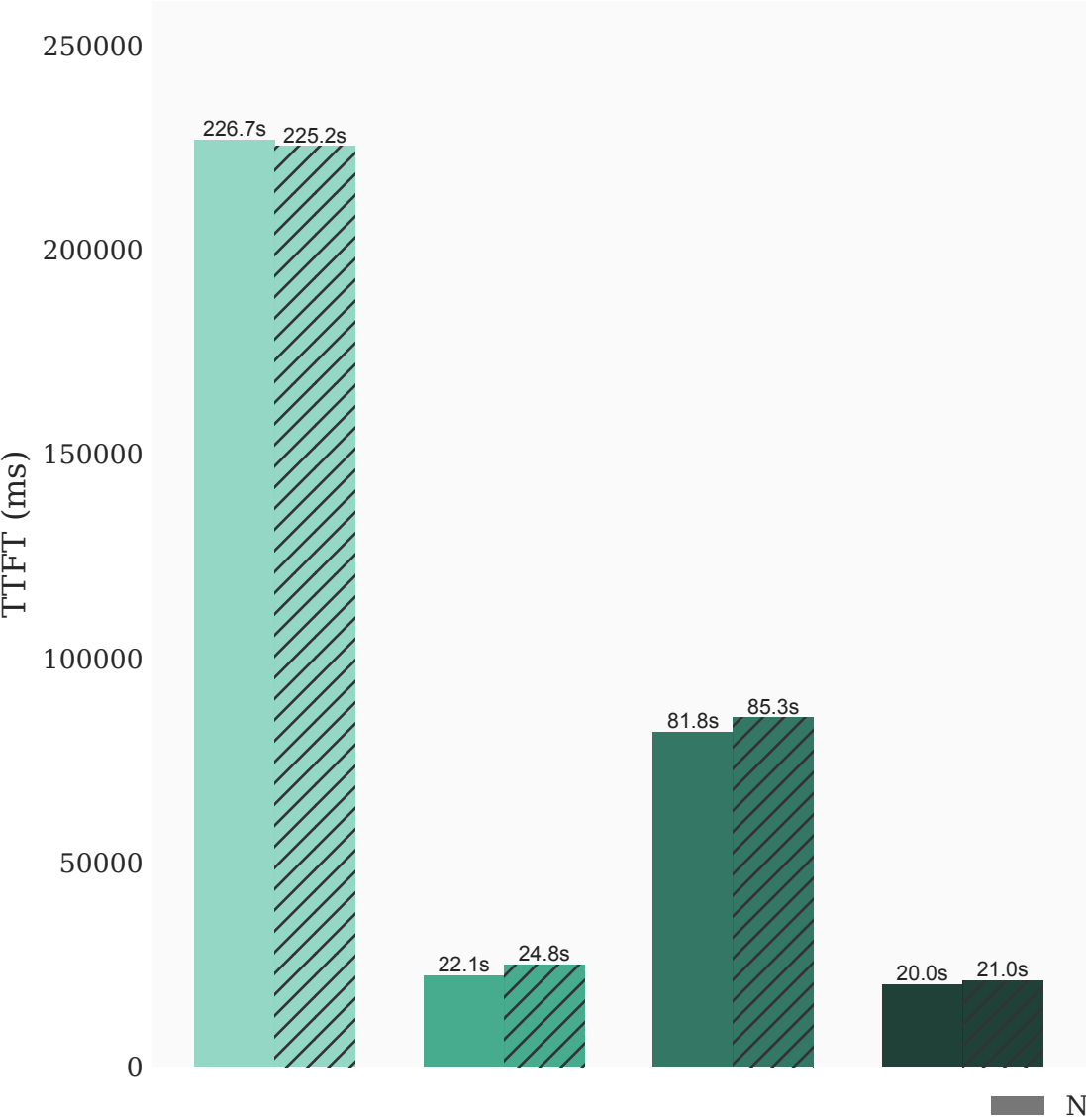
Time to First Token (P99)



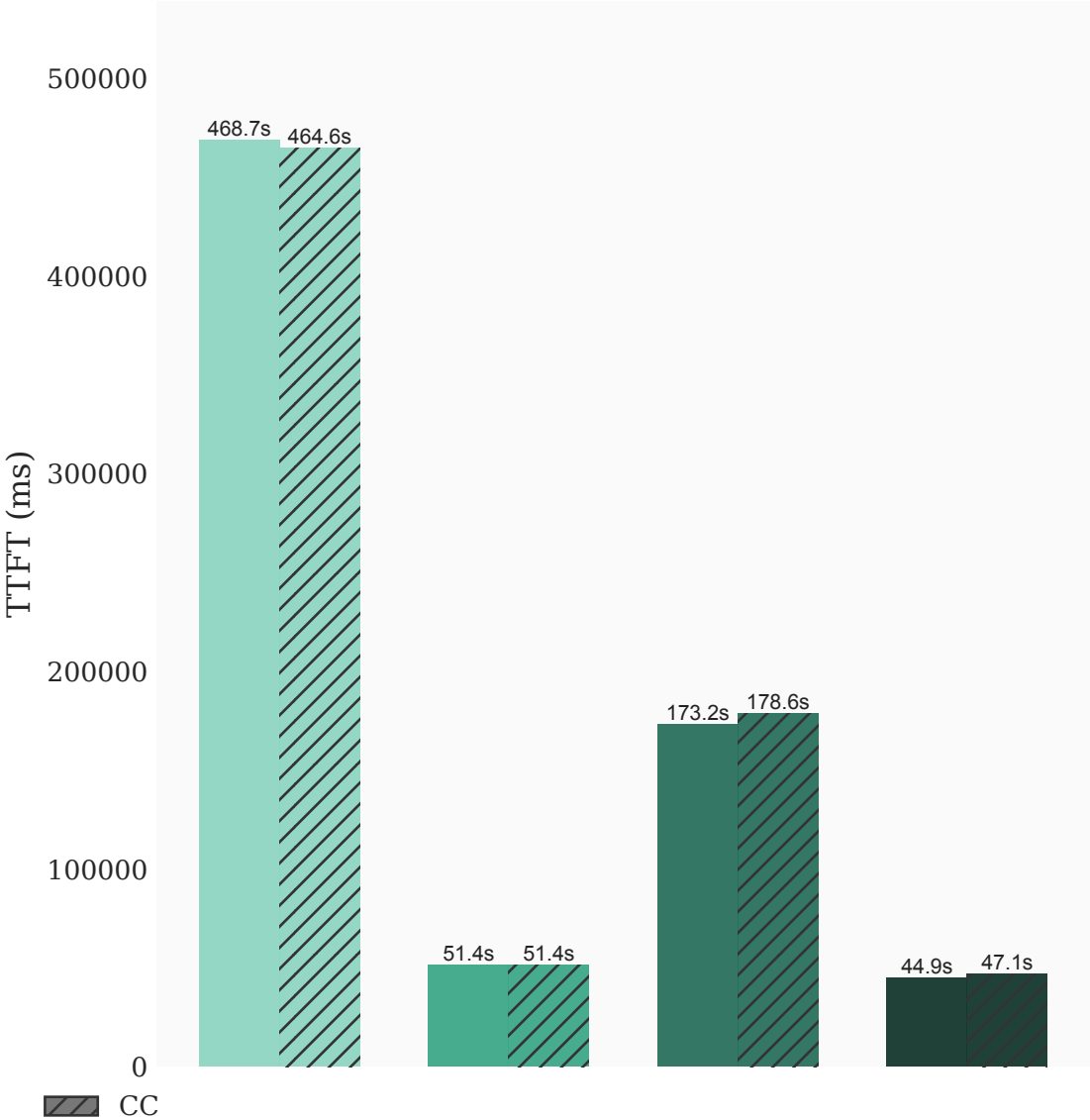
Llama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Random (4000 \Rightarrow 1000) (50 Request Rate)

Time to First Token (Mean)



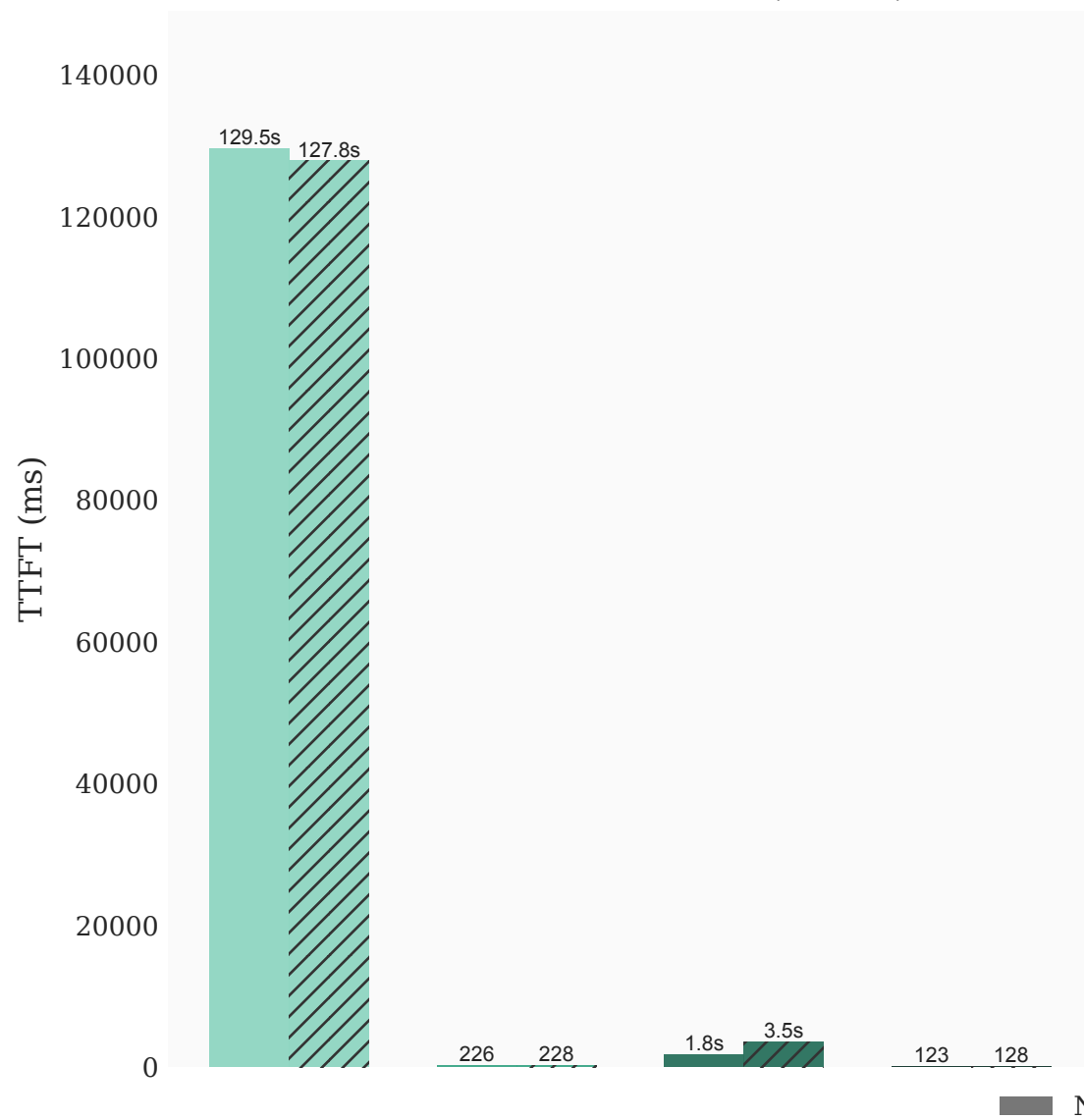
Time to First Token (P99)



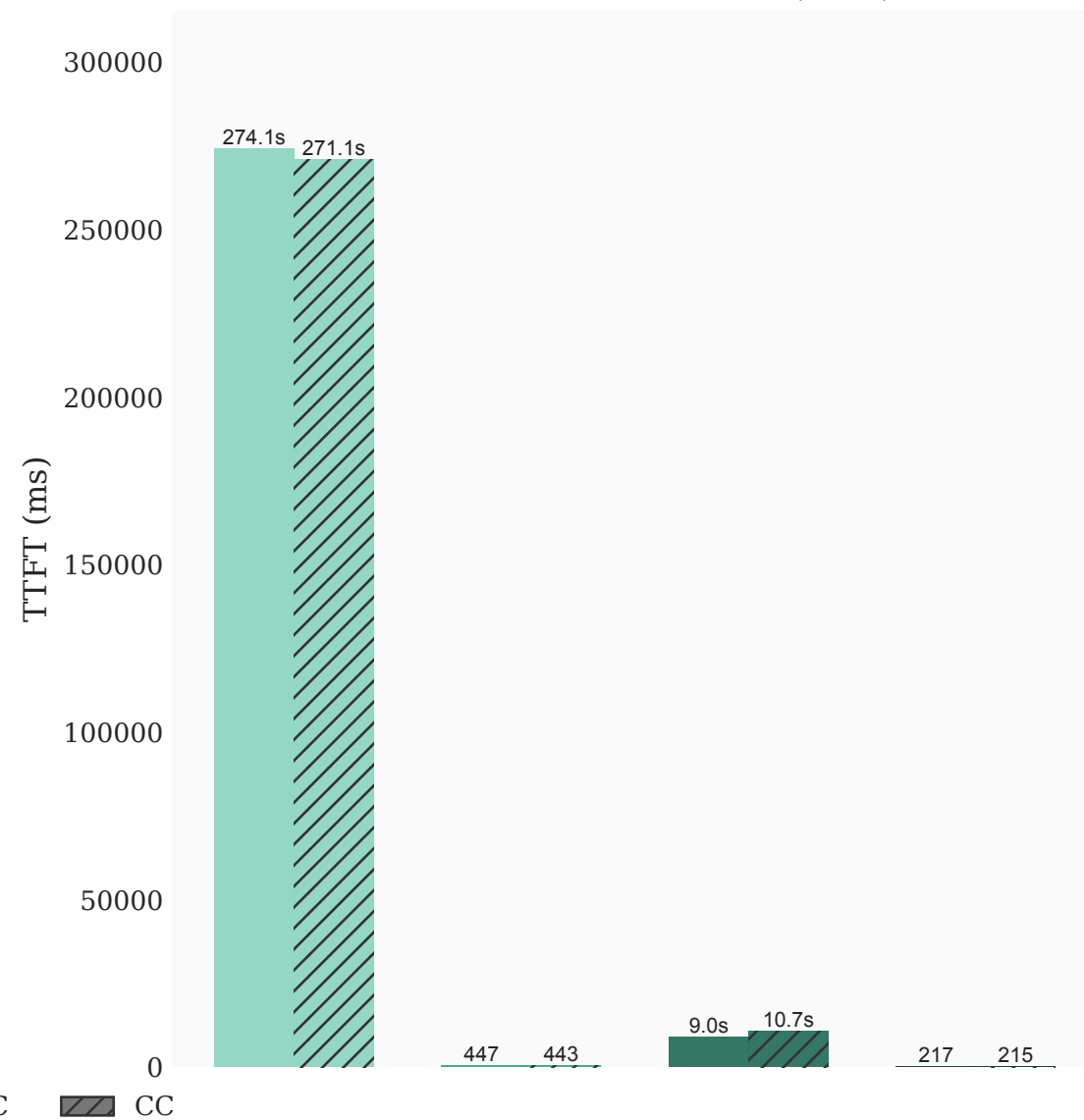
Llama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B Llama 3.1 8B

Random (4000 \Rightarrow 1000) (Single Request)

Time to First Token (Mean)



Time to First Token (P99)

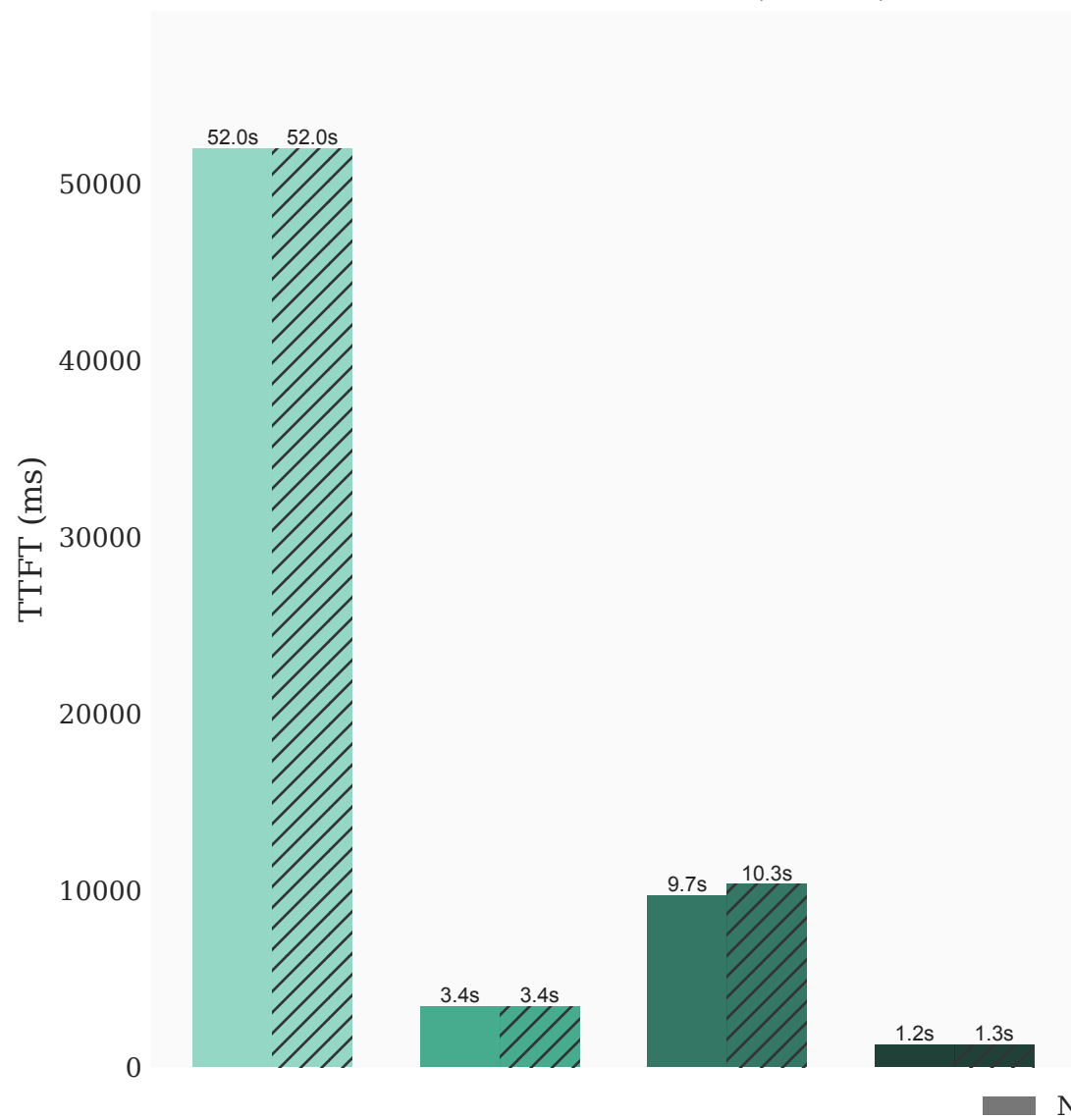


■ No CC ■ CC

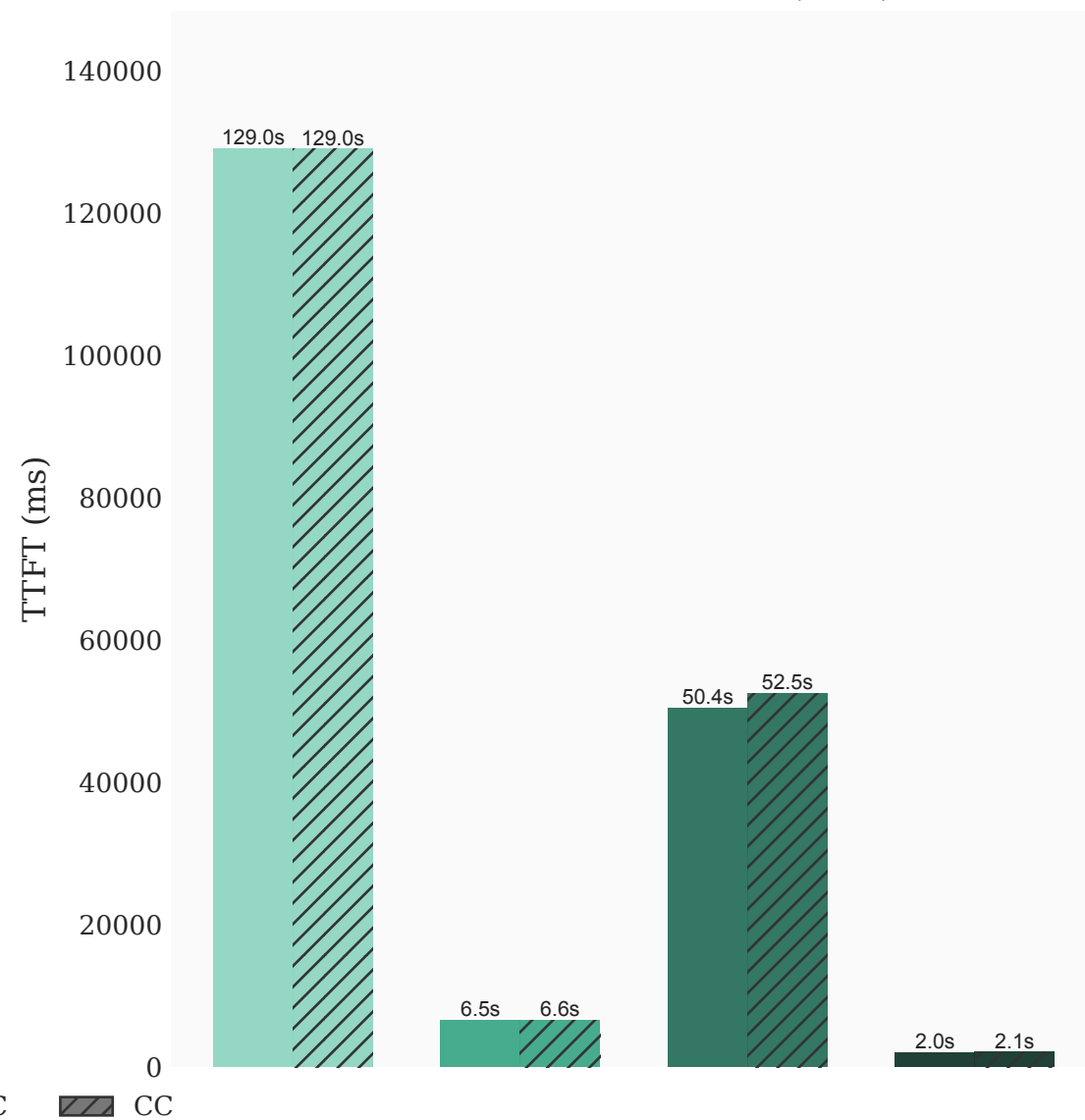
■ Llama 3.3 70B Int4 ■ GPT OSS 120B ■ Mistral 3.1 24B ■ Llama 3.1 8B

Random (1000 \Rightarrow 1000) (100 Request Rate)

Time to First Token (Mean)



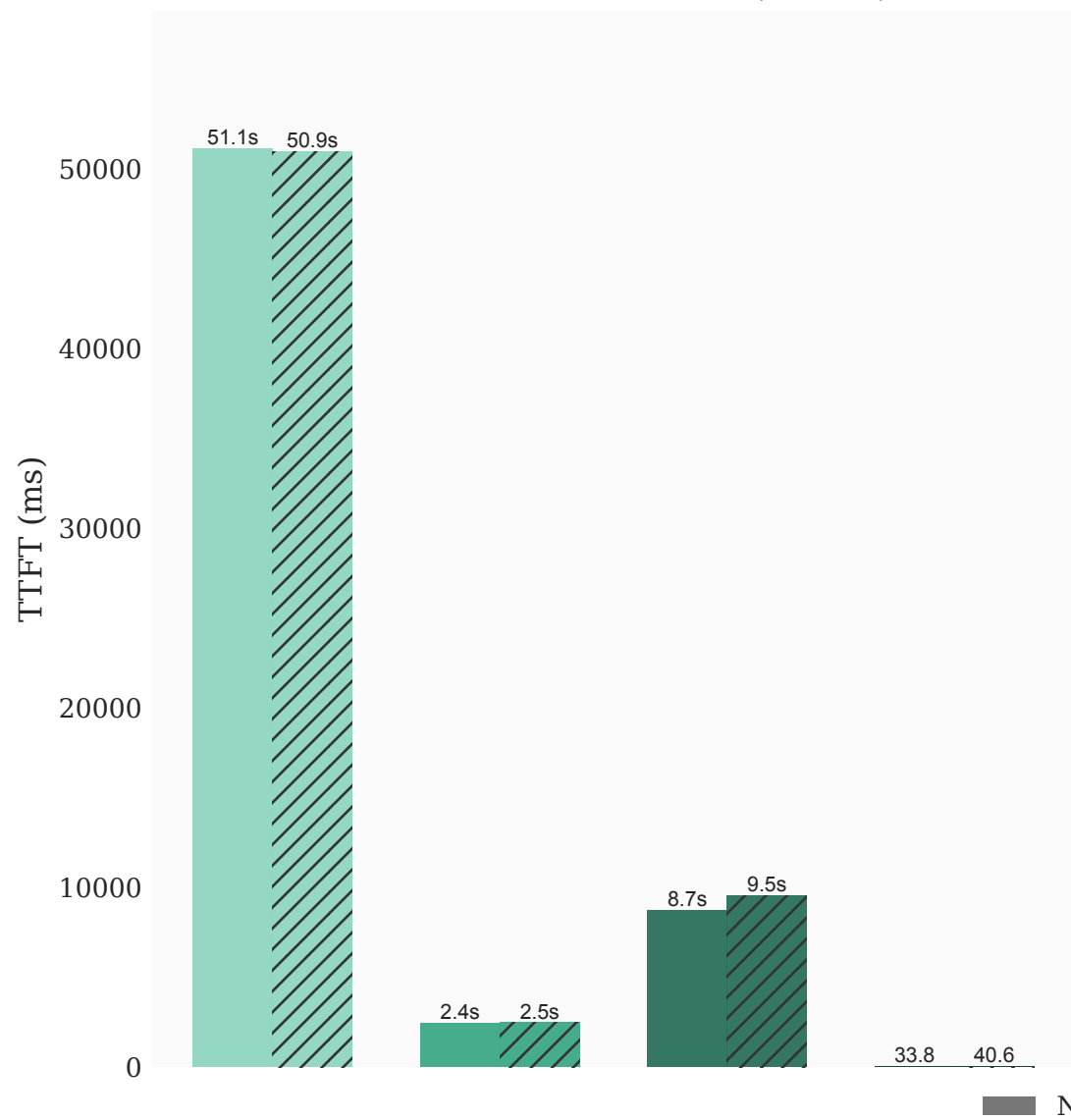
Time to First Token (P99)



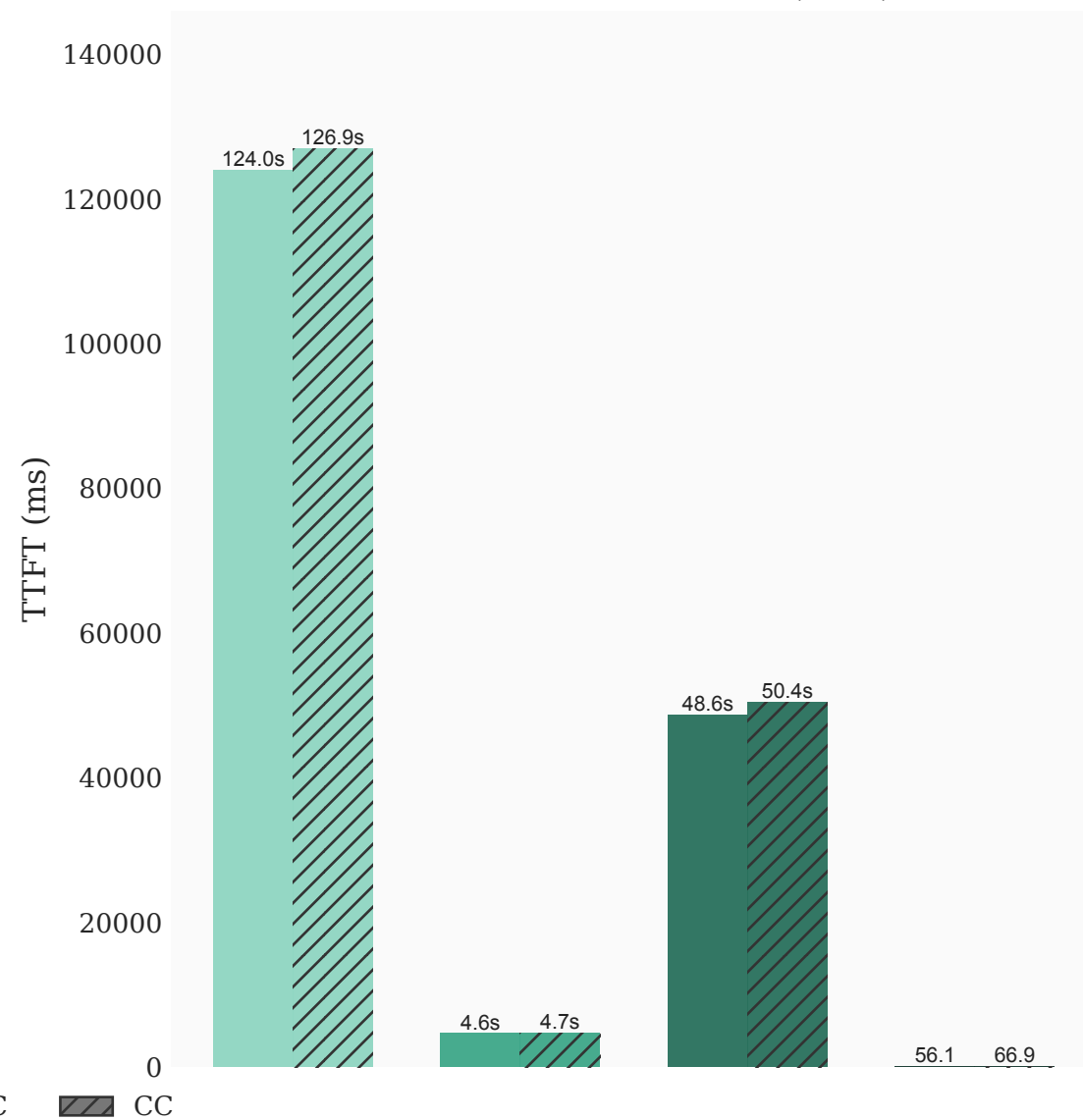
LLama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Random (1000 \Rightarrow 1000) (50 Request Rate)

Time to First Token (Mean)



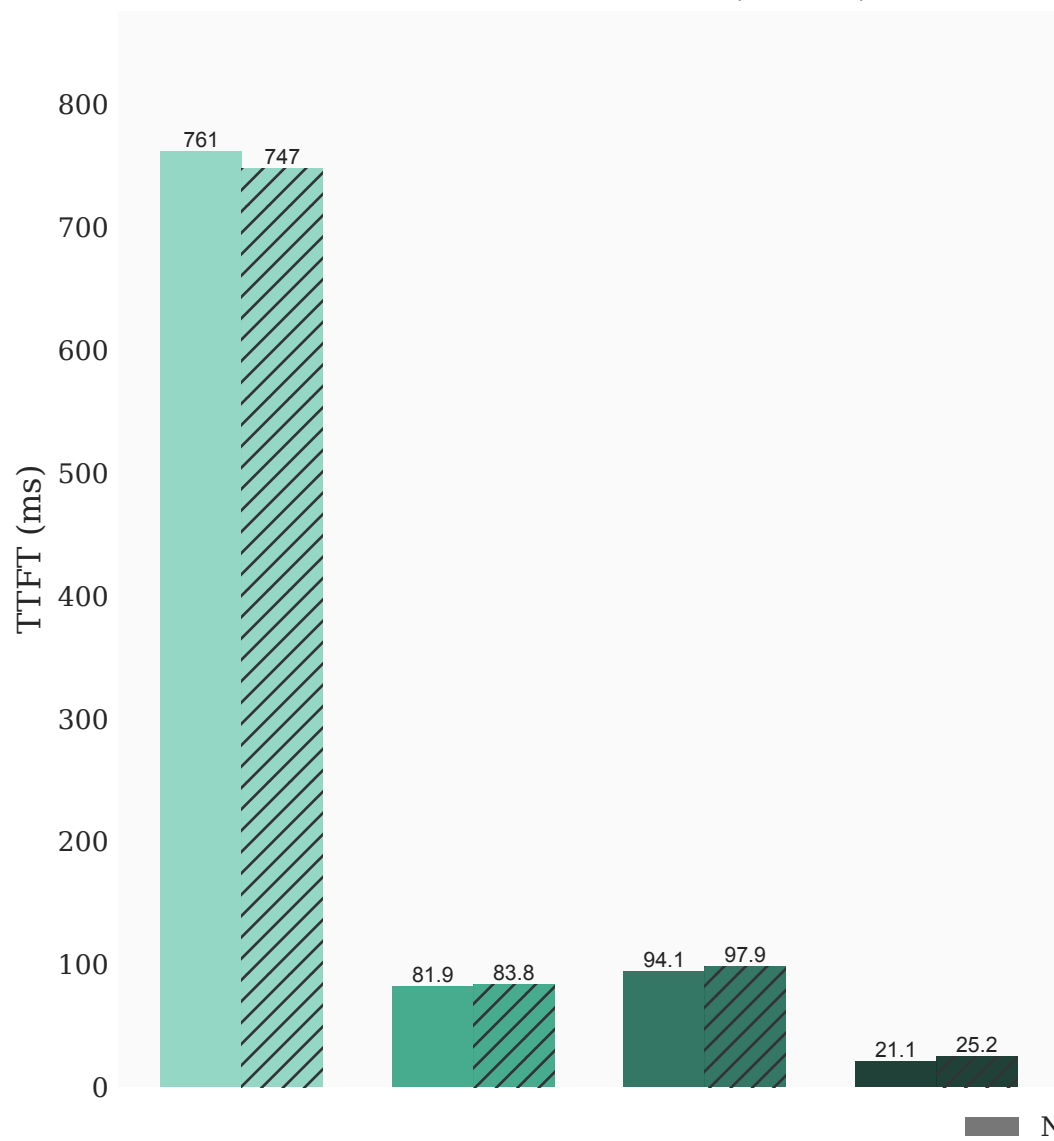
Time to First Token (P99)



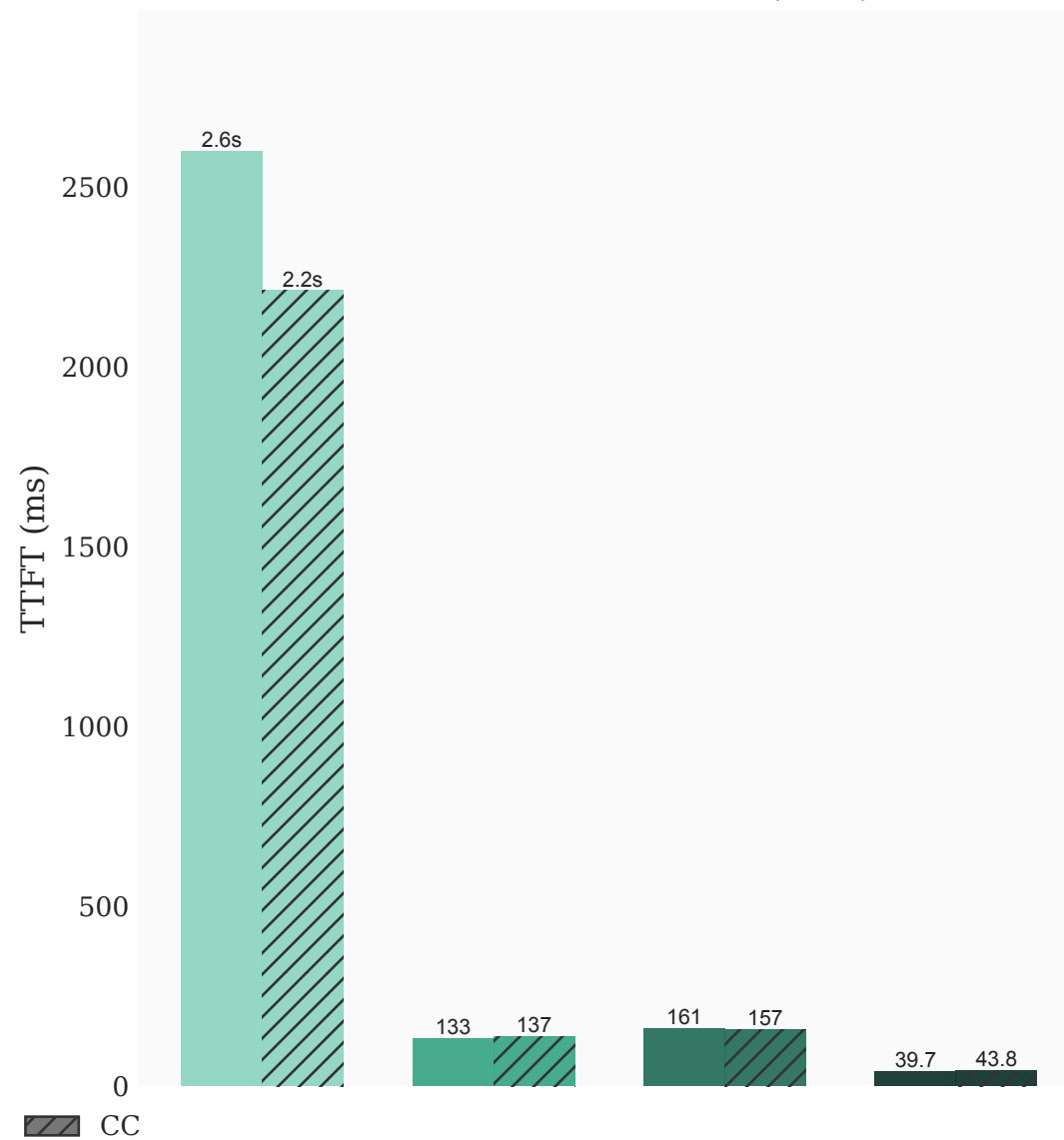
LLama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Random (1000 \Rightarrow 1000) (Single Request)

Time to First Token (Mean)



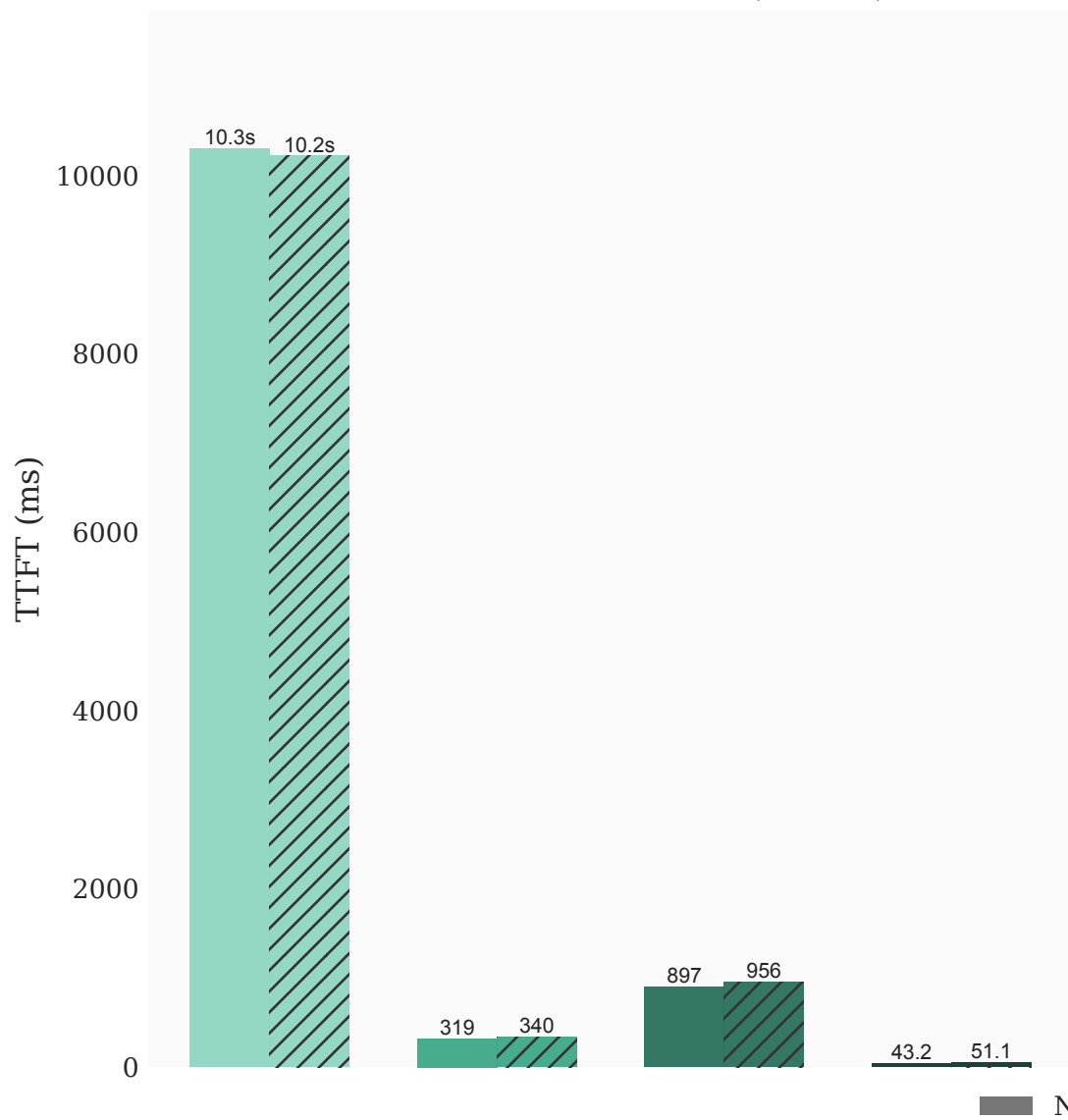
Time to First Token (P99)



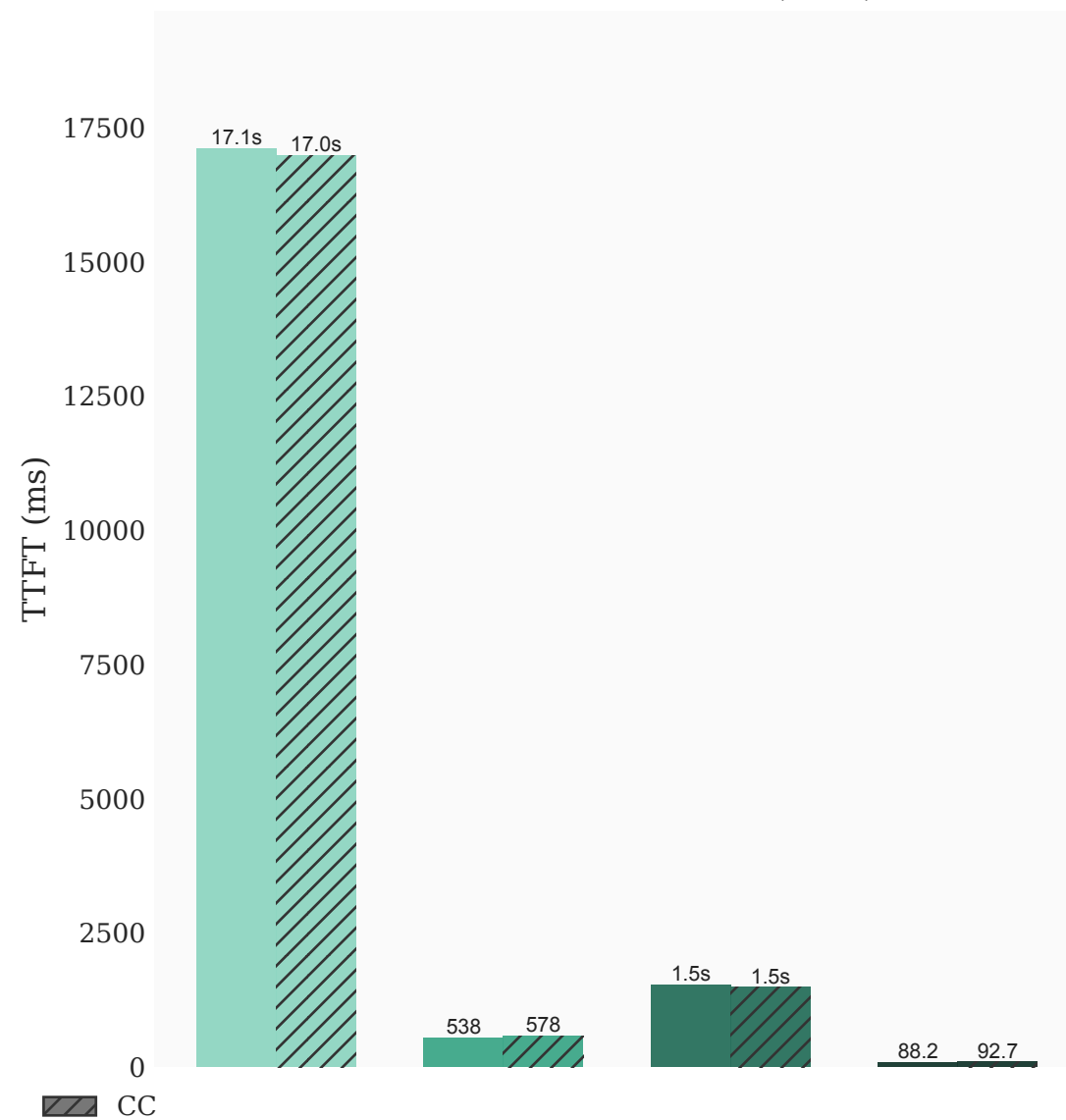
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

ShareGPT (100 Request Rate)

Time to First Token (Mean)



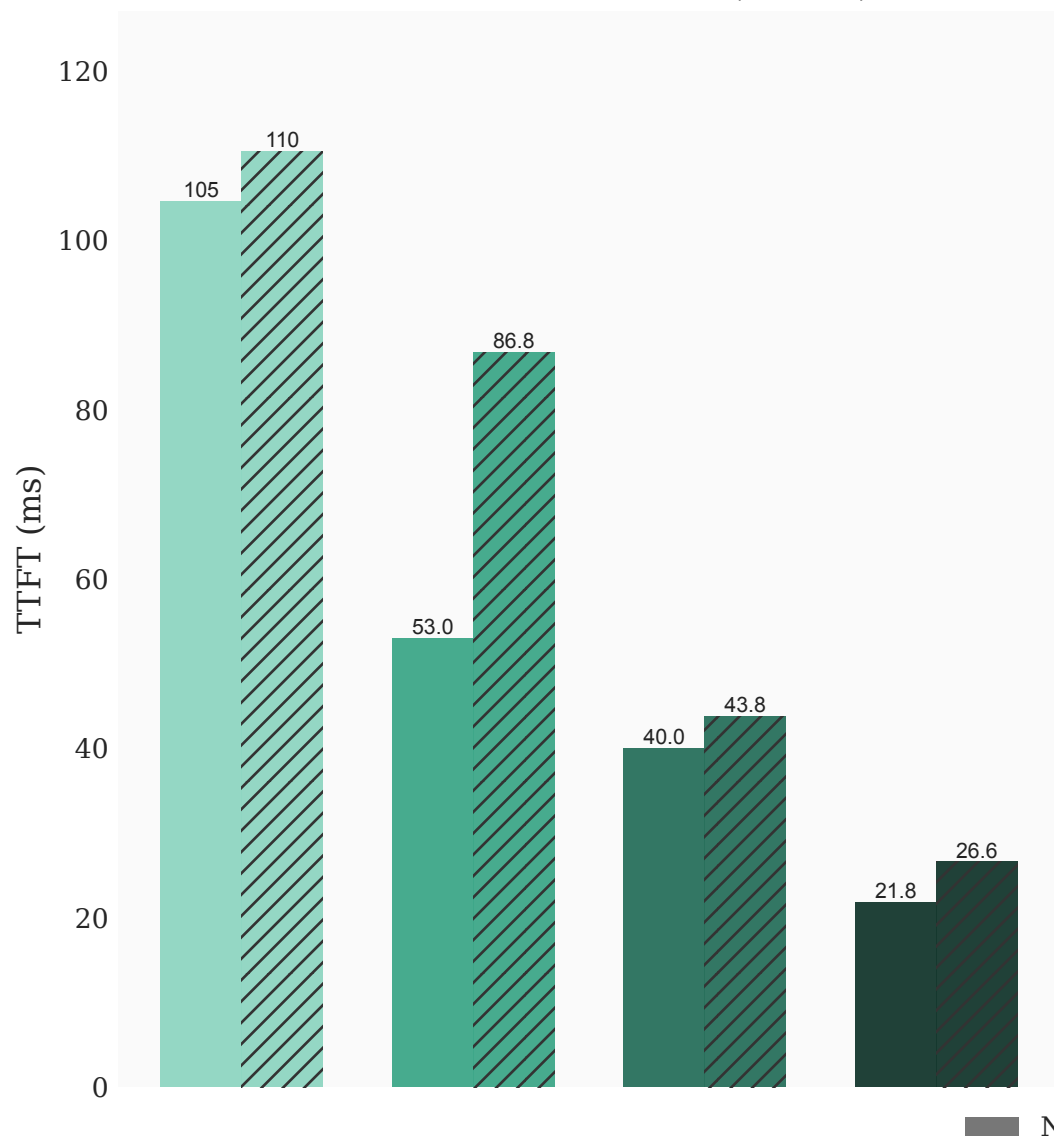
Time to First Token (P99)



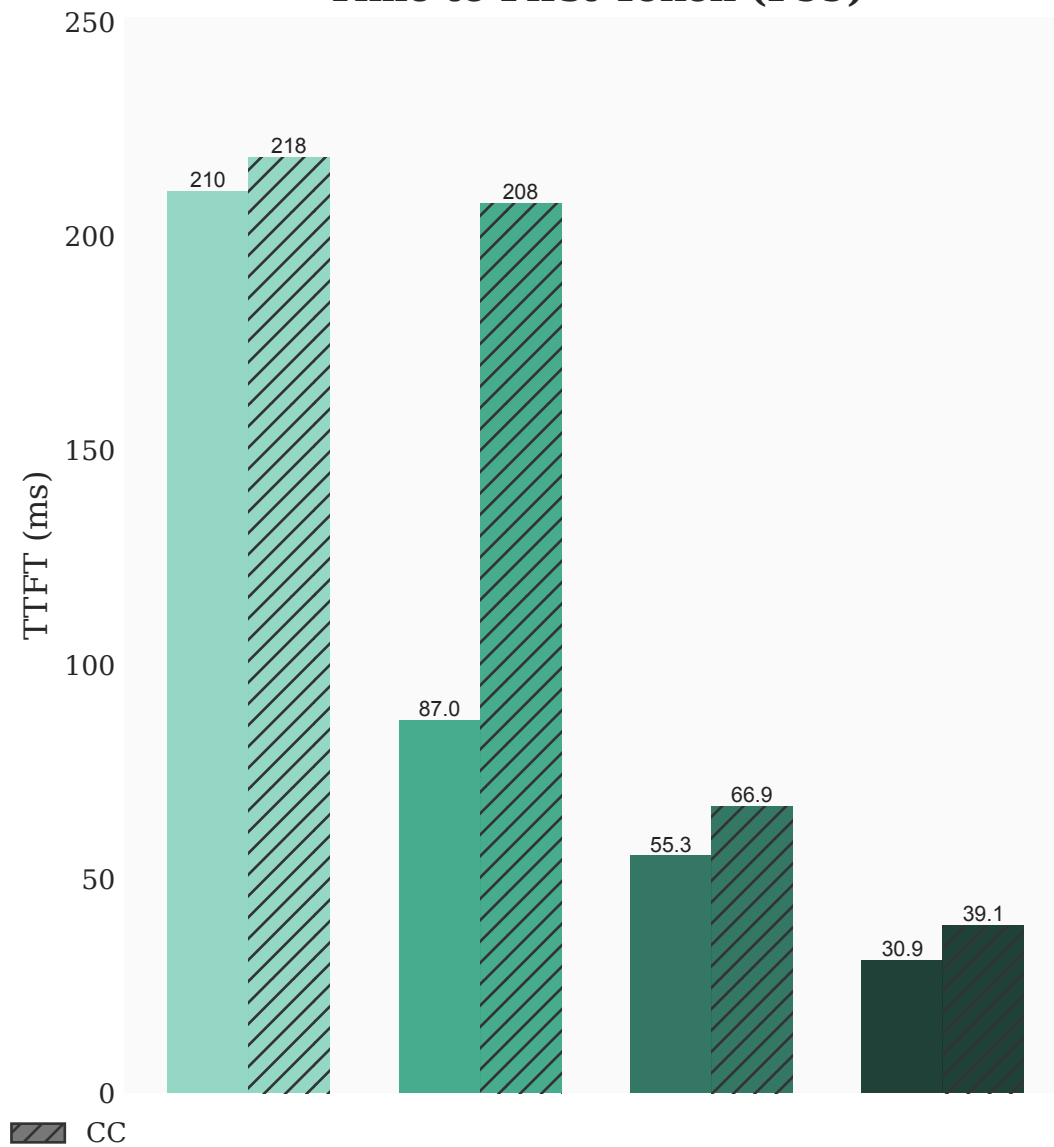
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

ShareGPT (50 Request Rate)

Time to First Token (Mean)



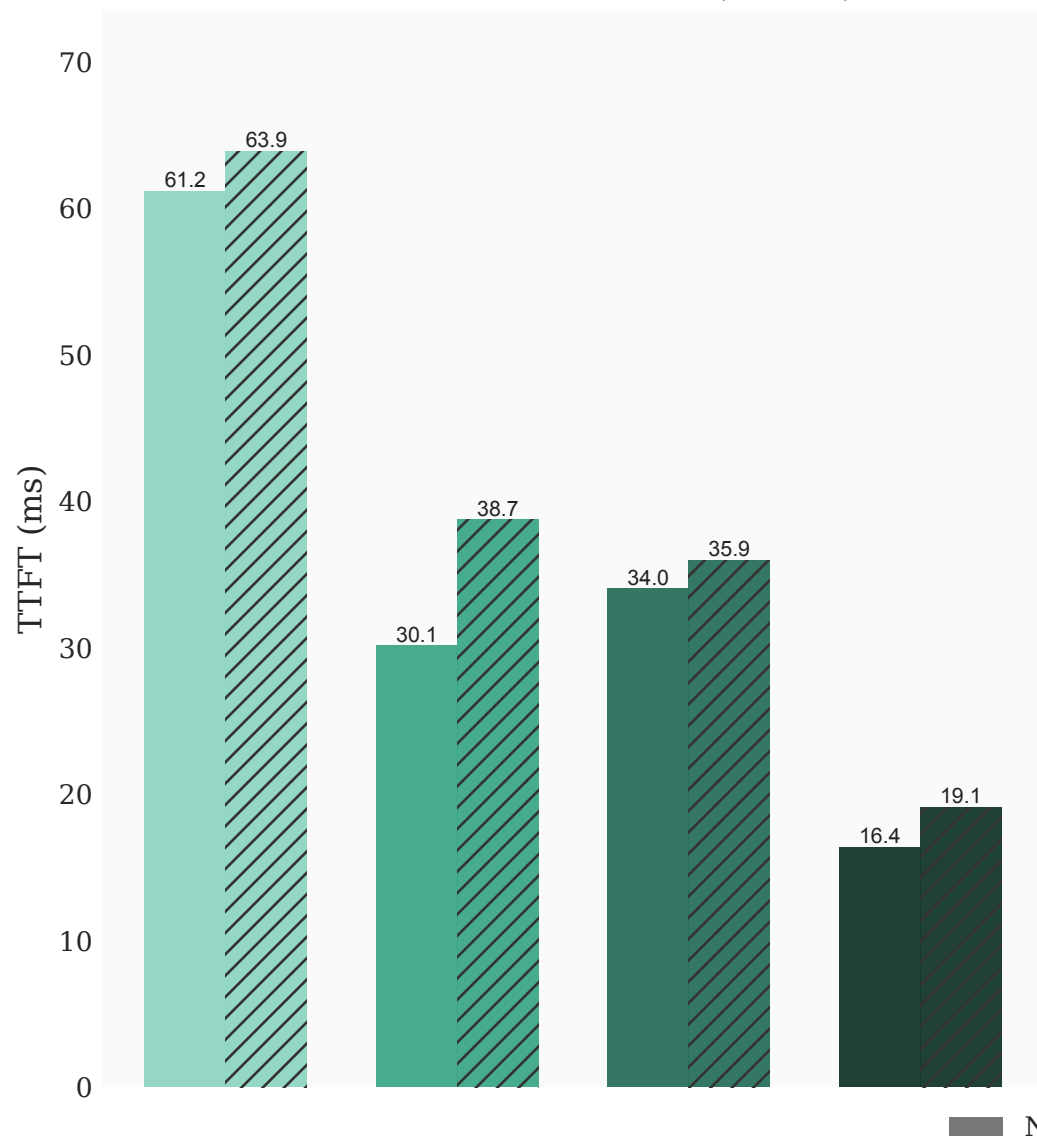
Time to First Token (P99)



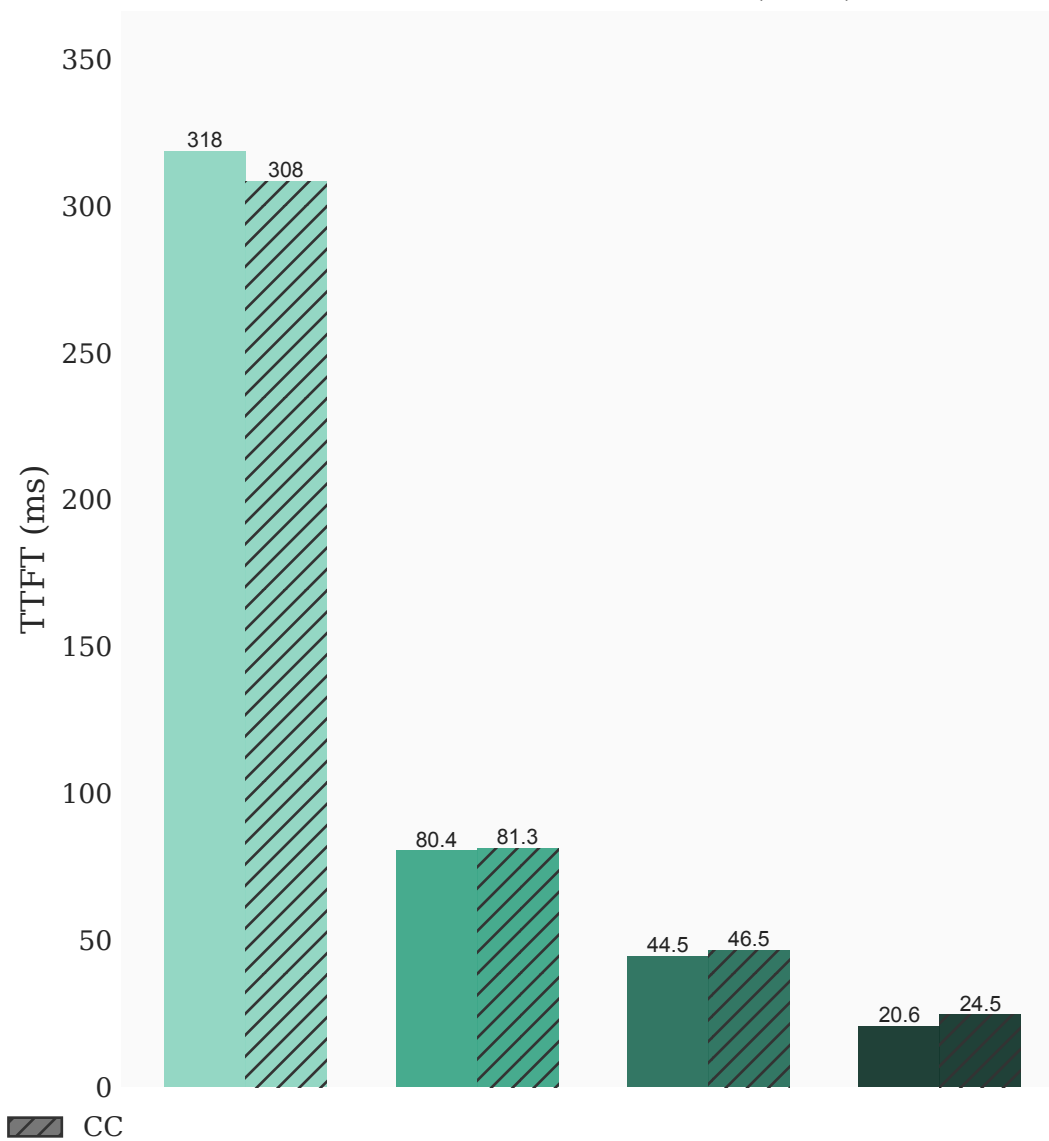
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

ShareGPT (Single Request)

Time to First Token (Mean)



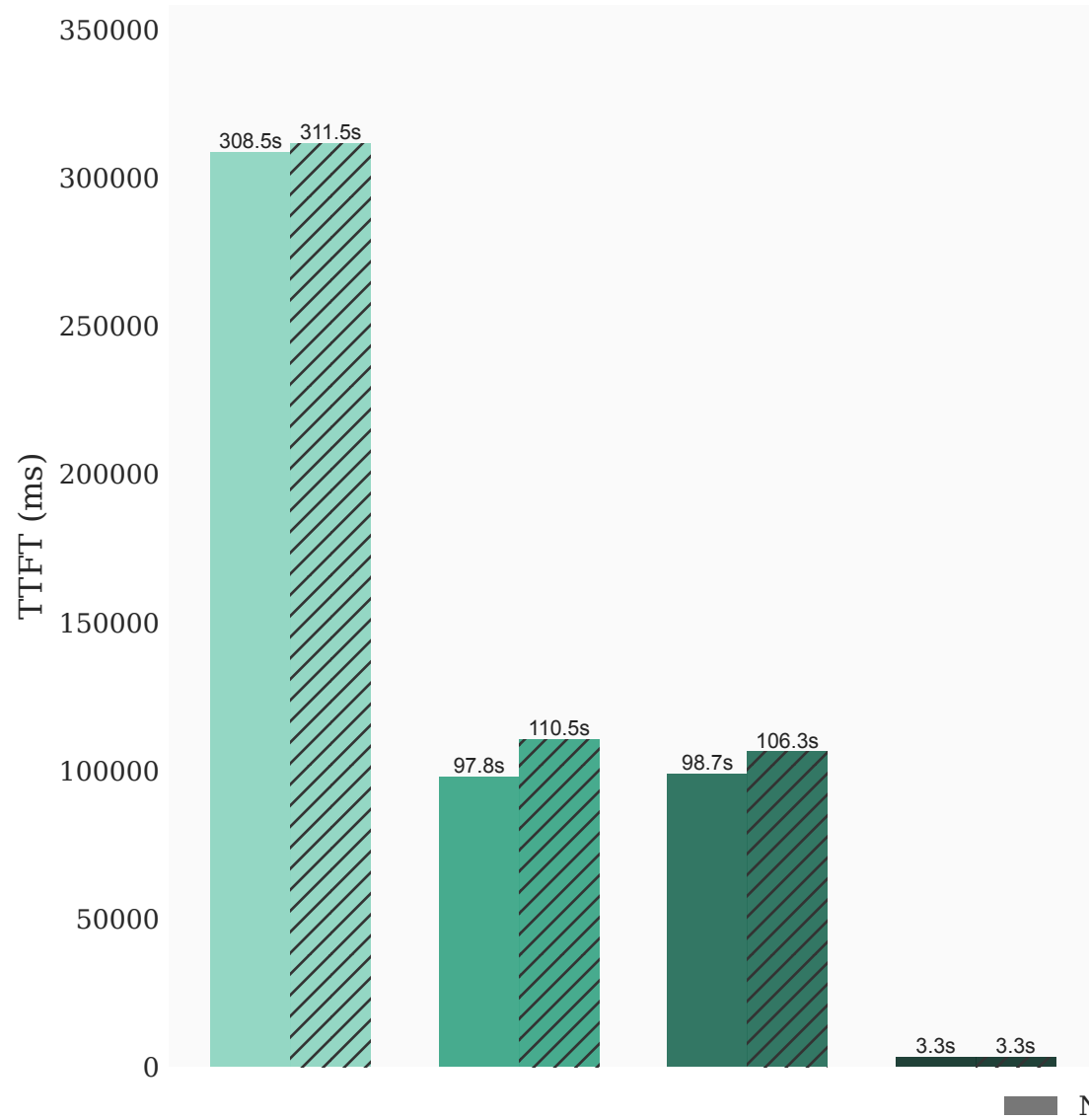
Time to First Token (P99)



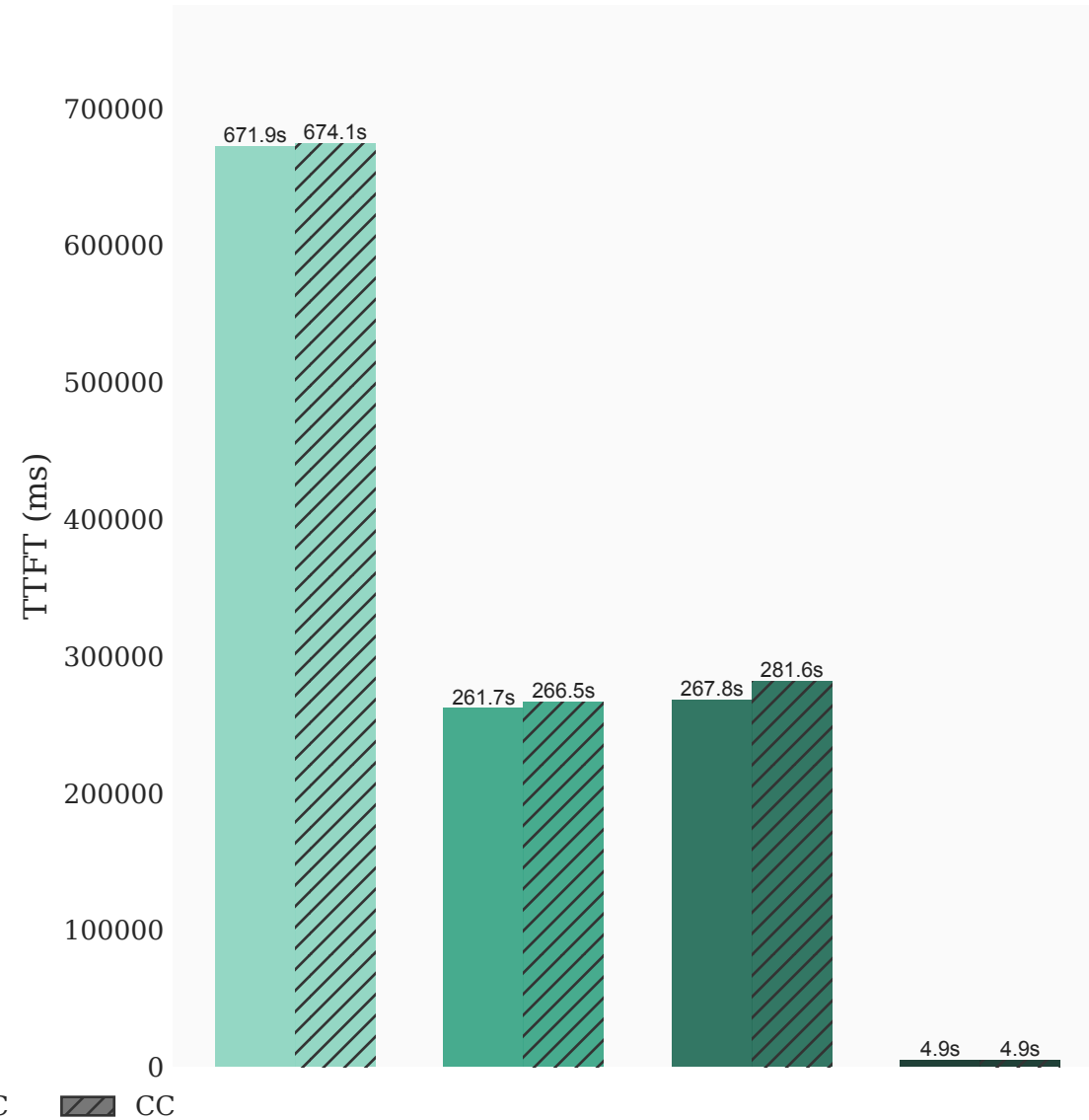
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

Edit 10K Characters (100 Request Rate)

Time to First Token (Mean)



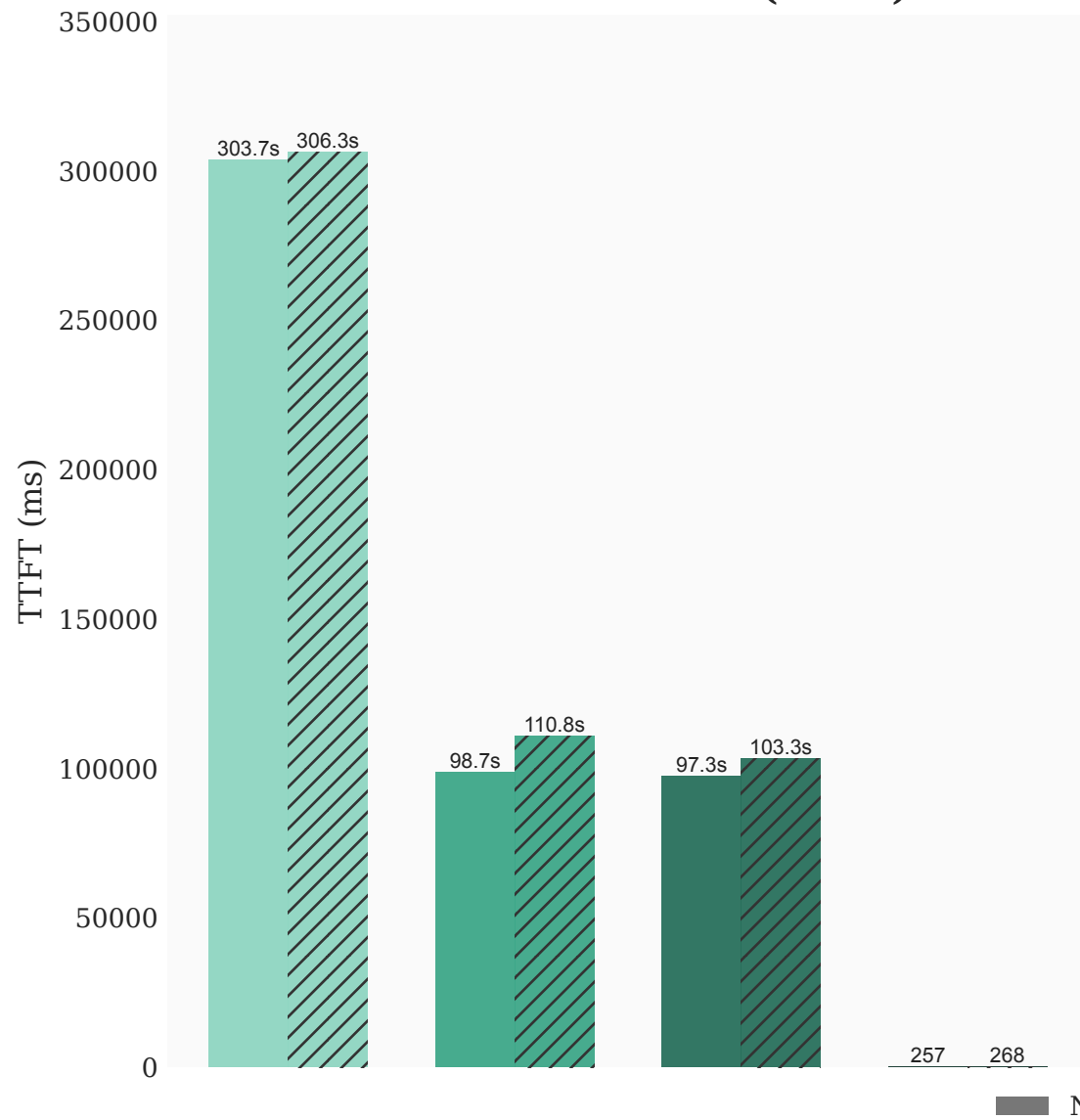
Time to First Token (P99)



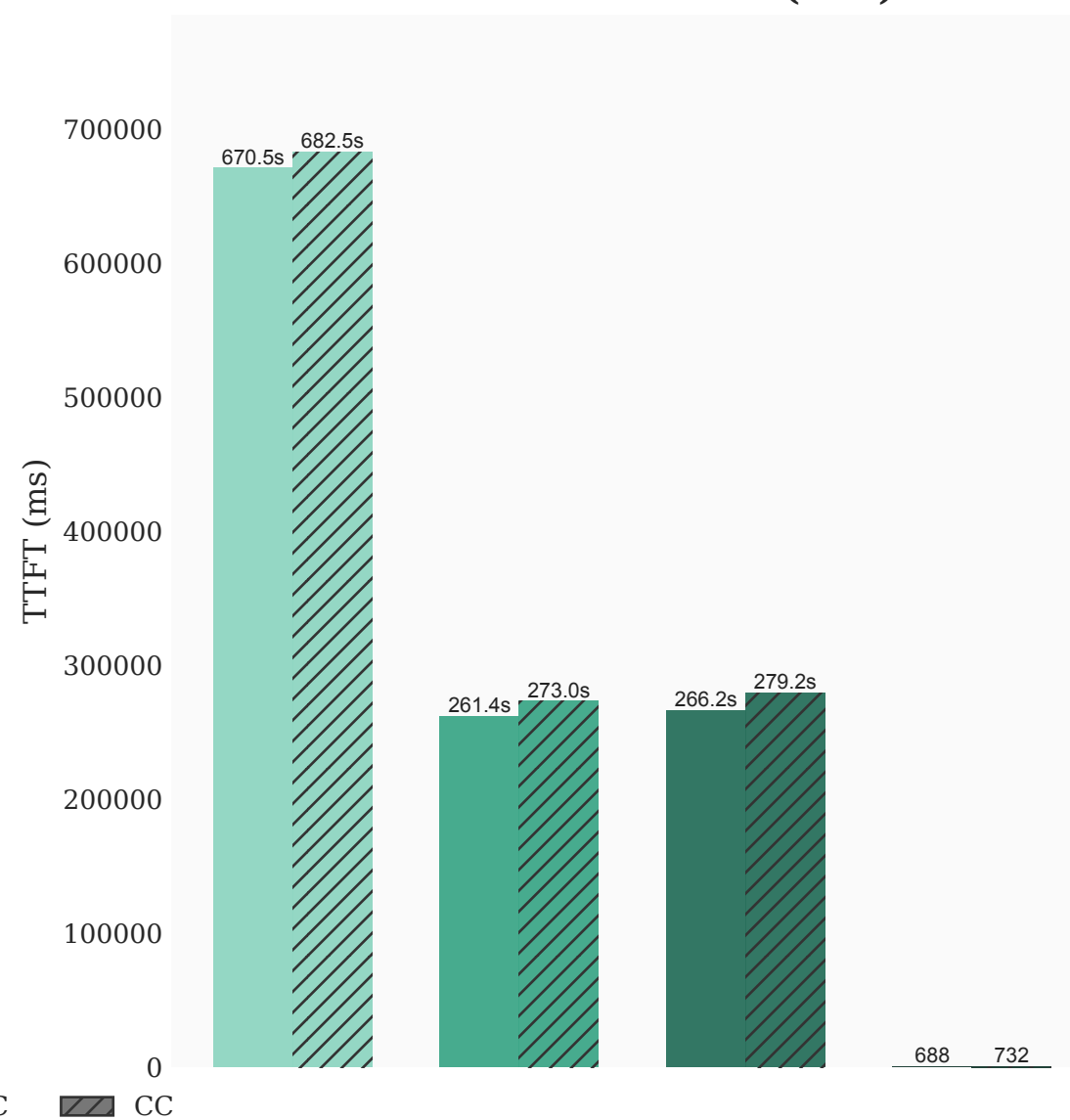
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

Edit 10K Characters (50 Request Rate)

Time to First Token (Mean)



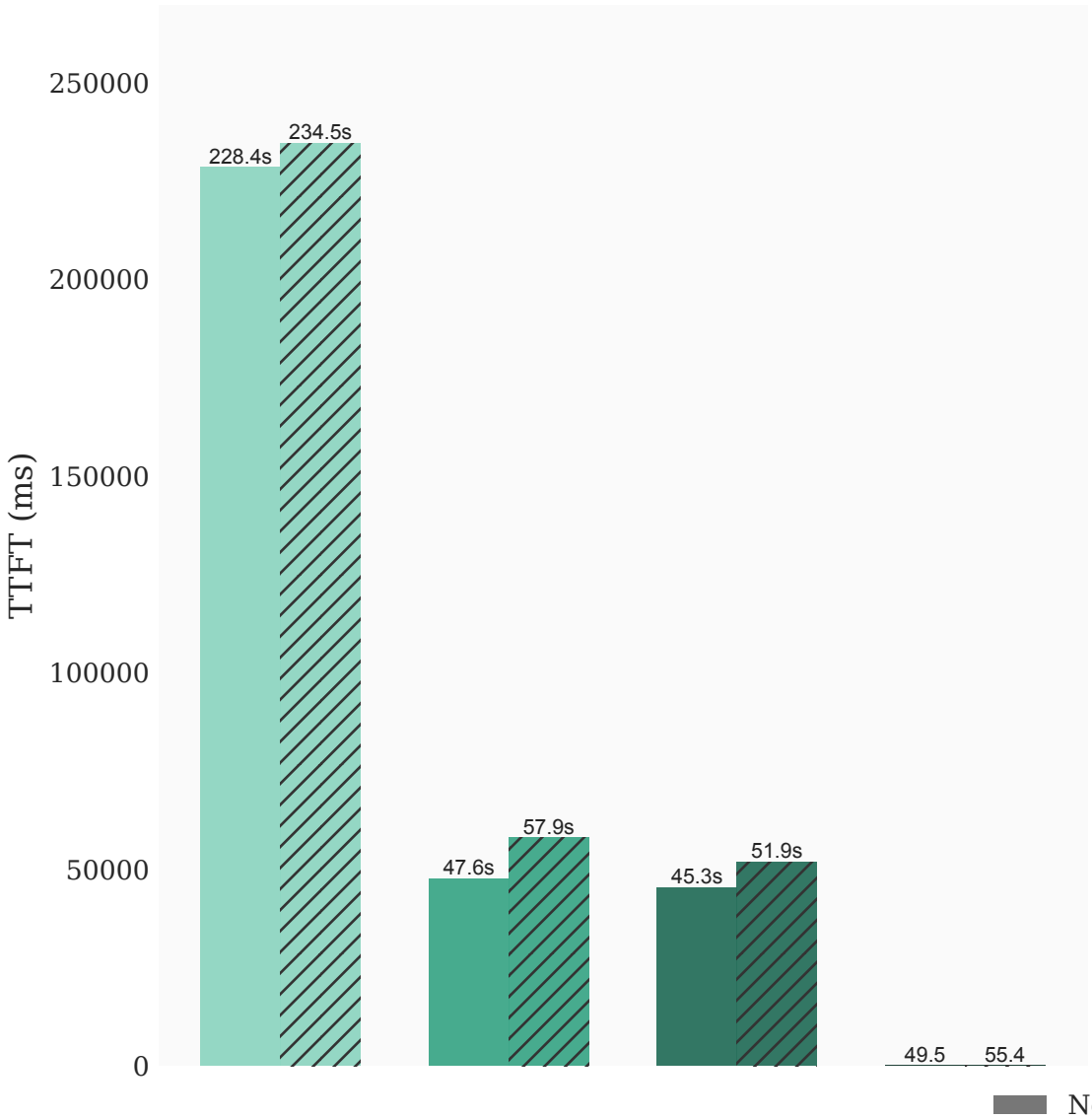
Time to First Token (P99)



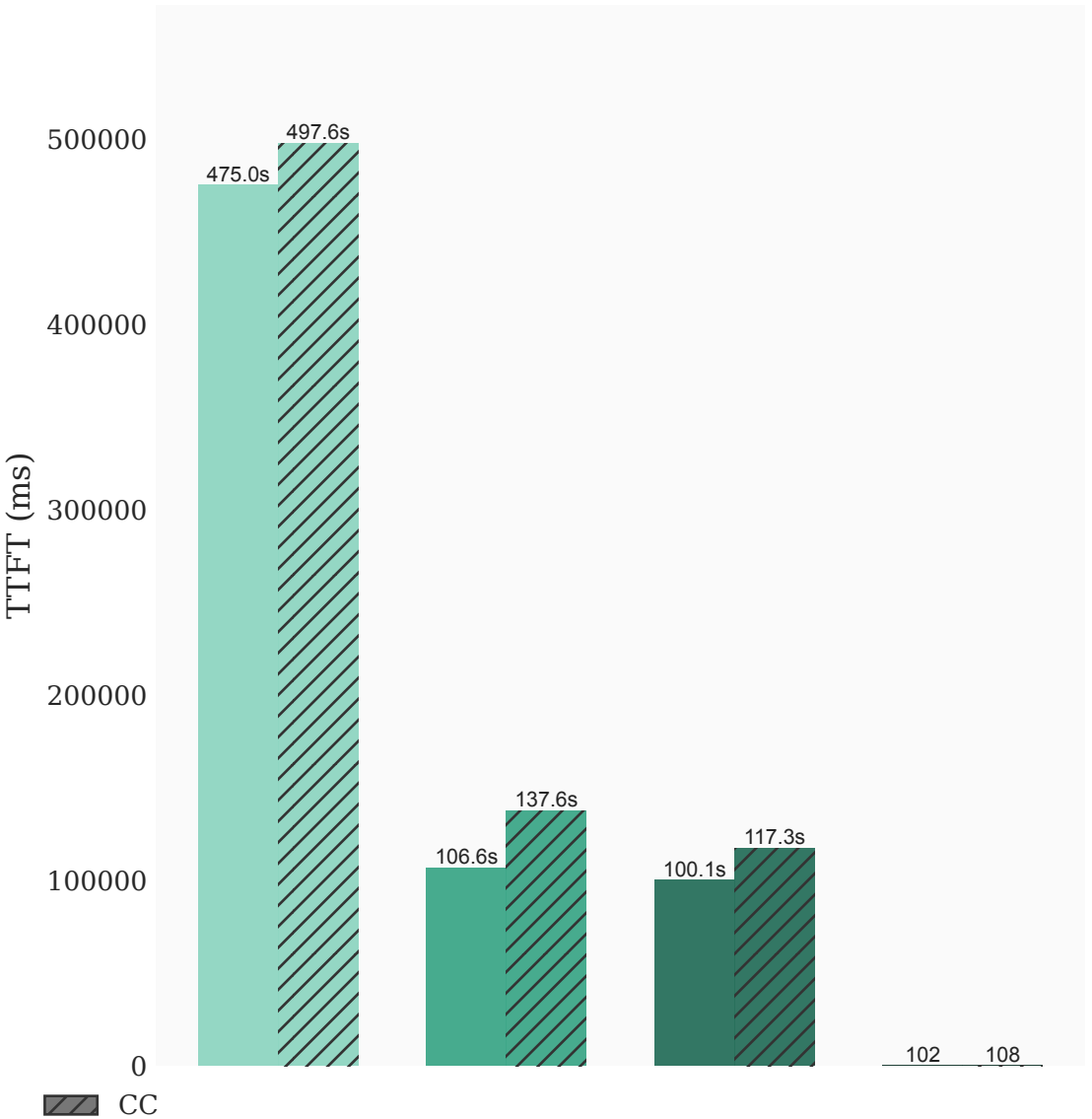
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

Edit 10K Characters (Single Request)

Time to First Token (Mean)



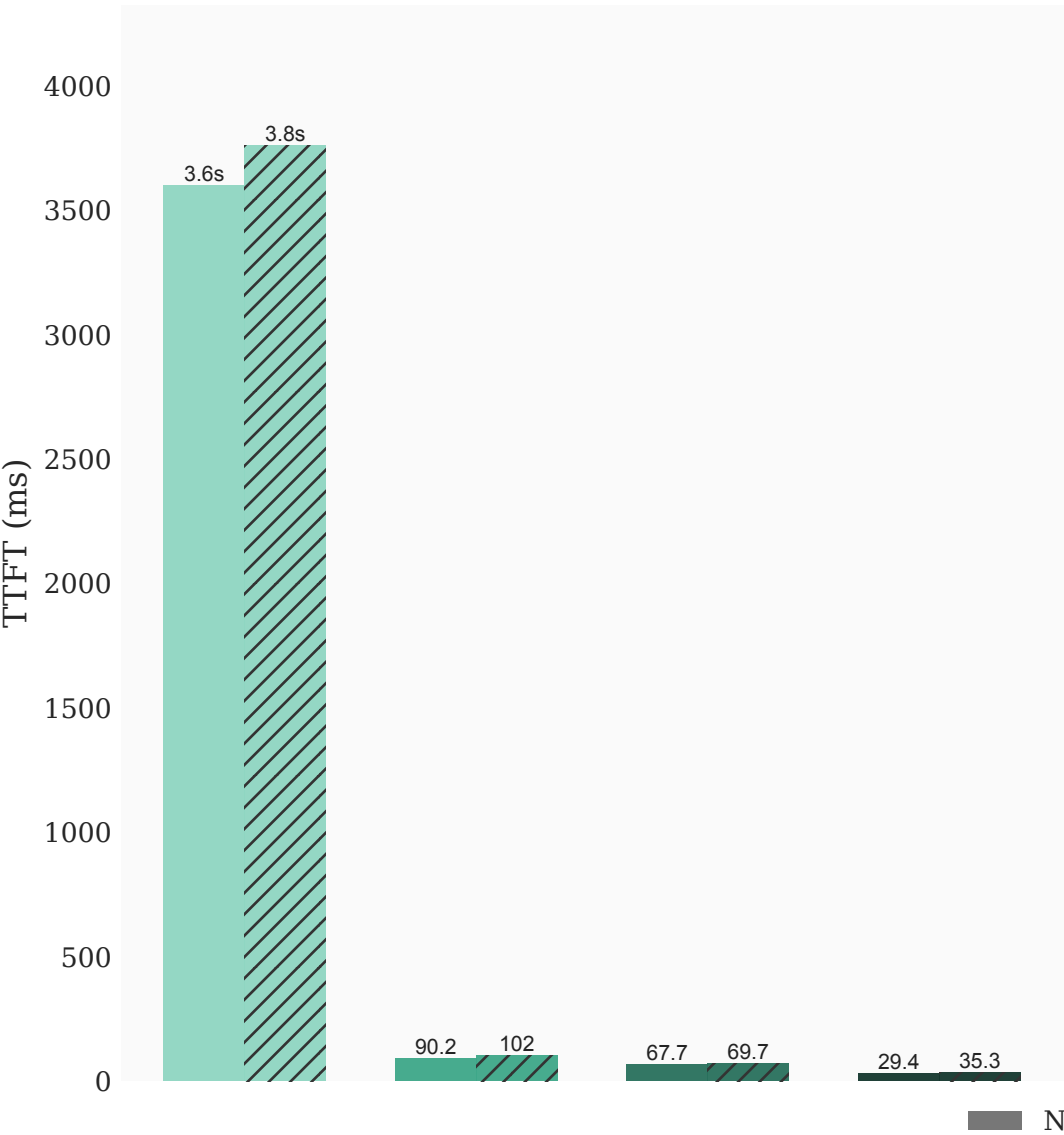
Time to First Token (P99)



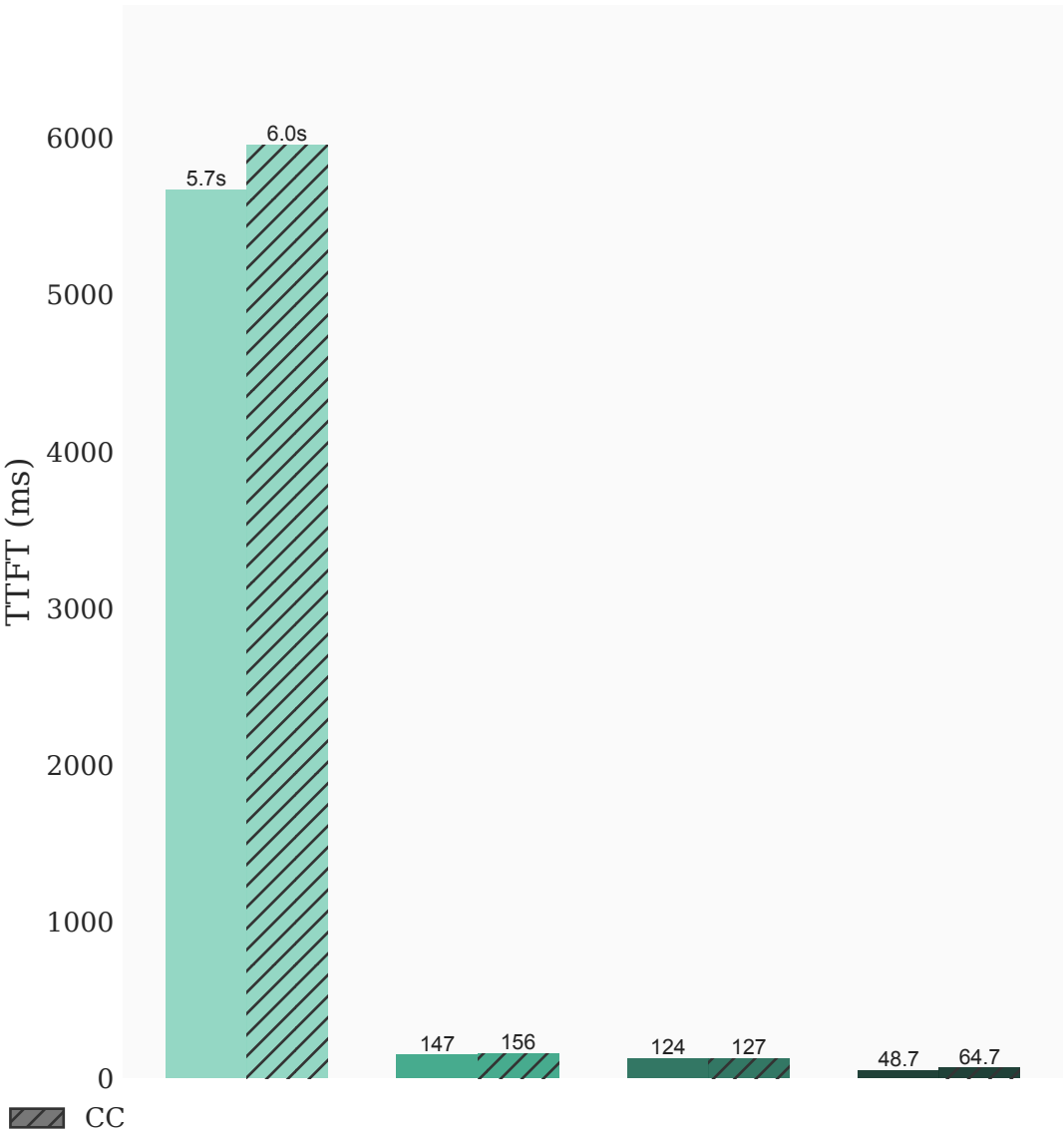
LLama 3.3 70B Int4 GPT OSS 120B Mistral 3.1 24B LLama 3.1 8B

Numina Math (100 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)

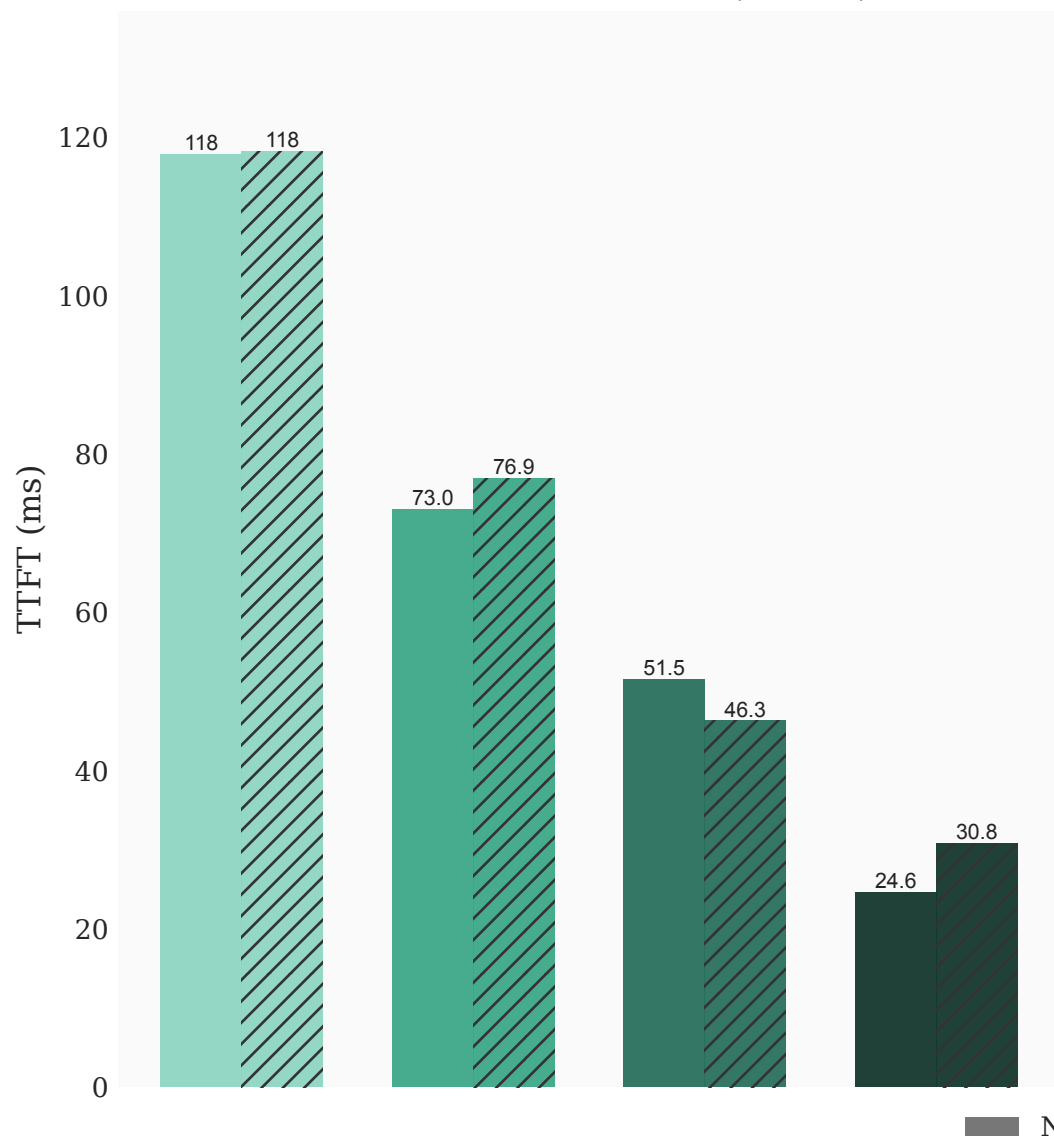


■ No CC ■ CC

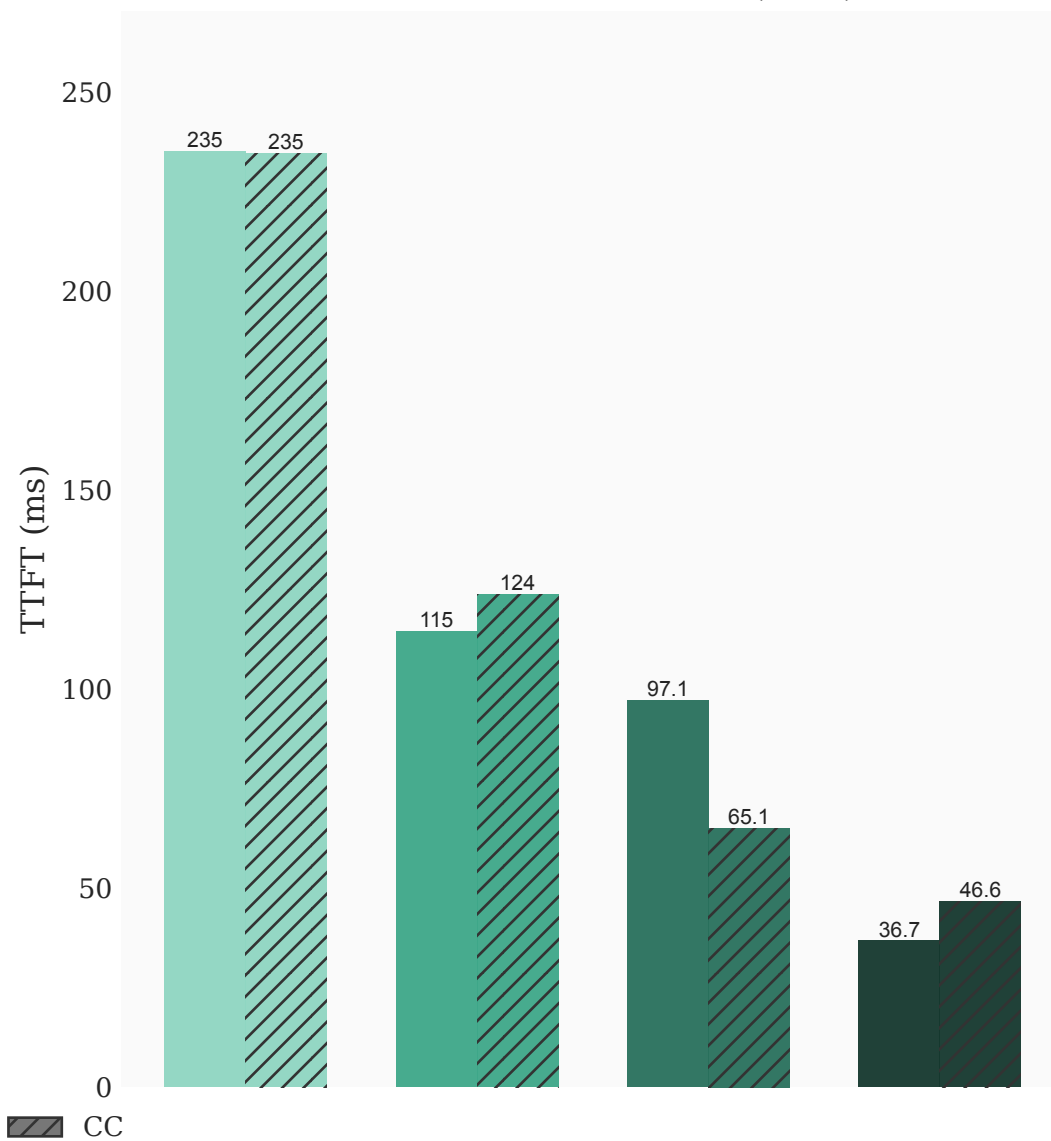
■ Llama 3.3 70B Int4 ■ GPT OSS 120B ■ Mistral 3.1 24B ■ Llama 3.1 8B

Numina Math (50 Request Rate)

Time to First Token (Mean)



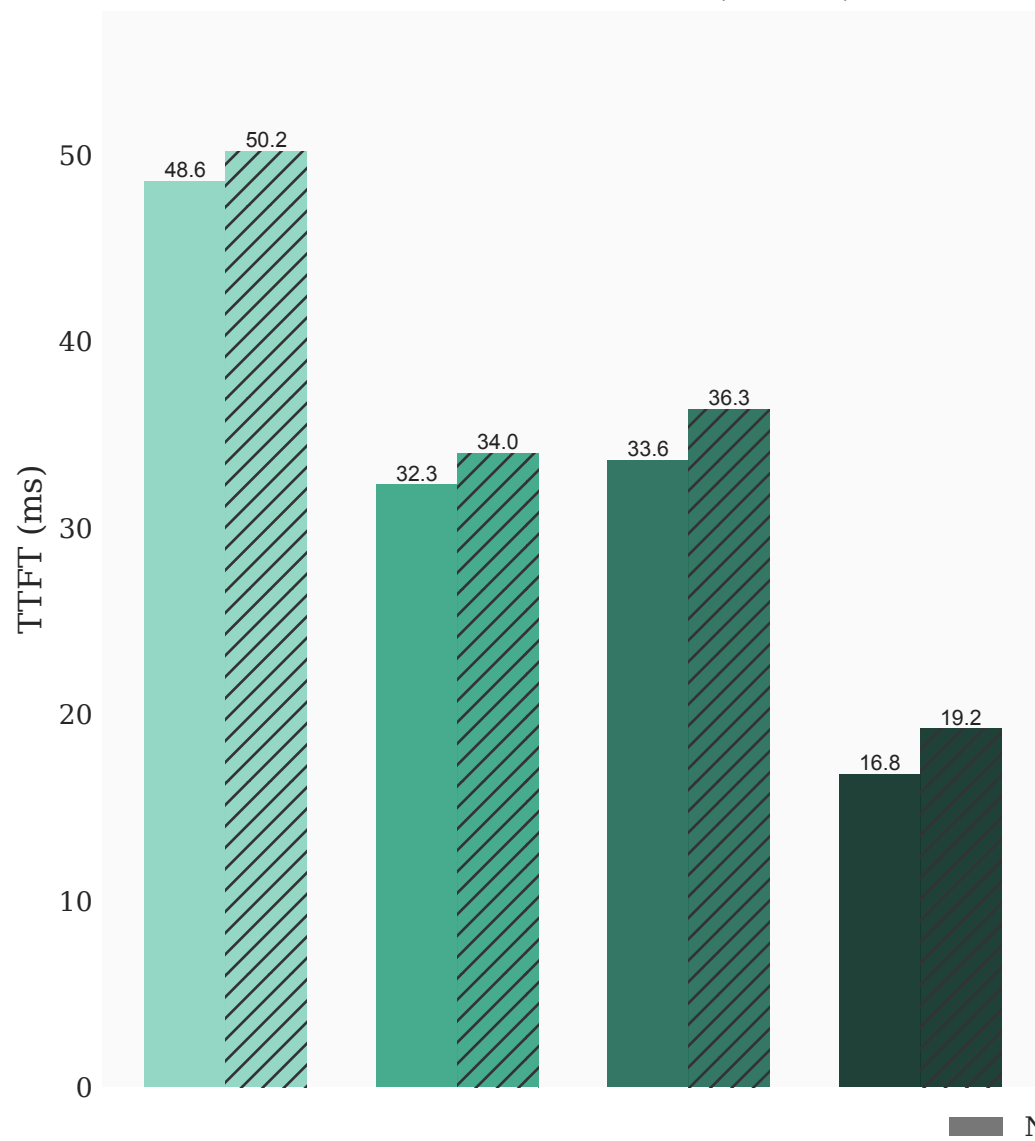
Time to First Token (P99)



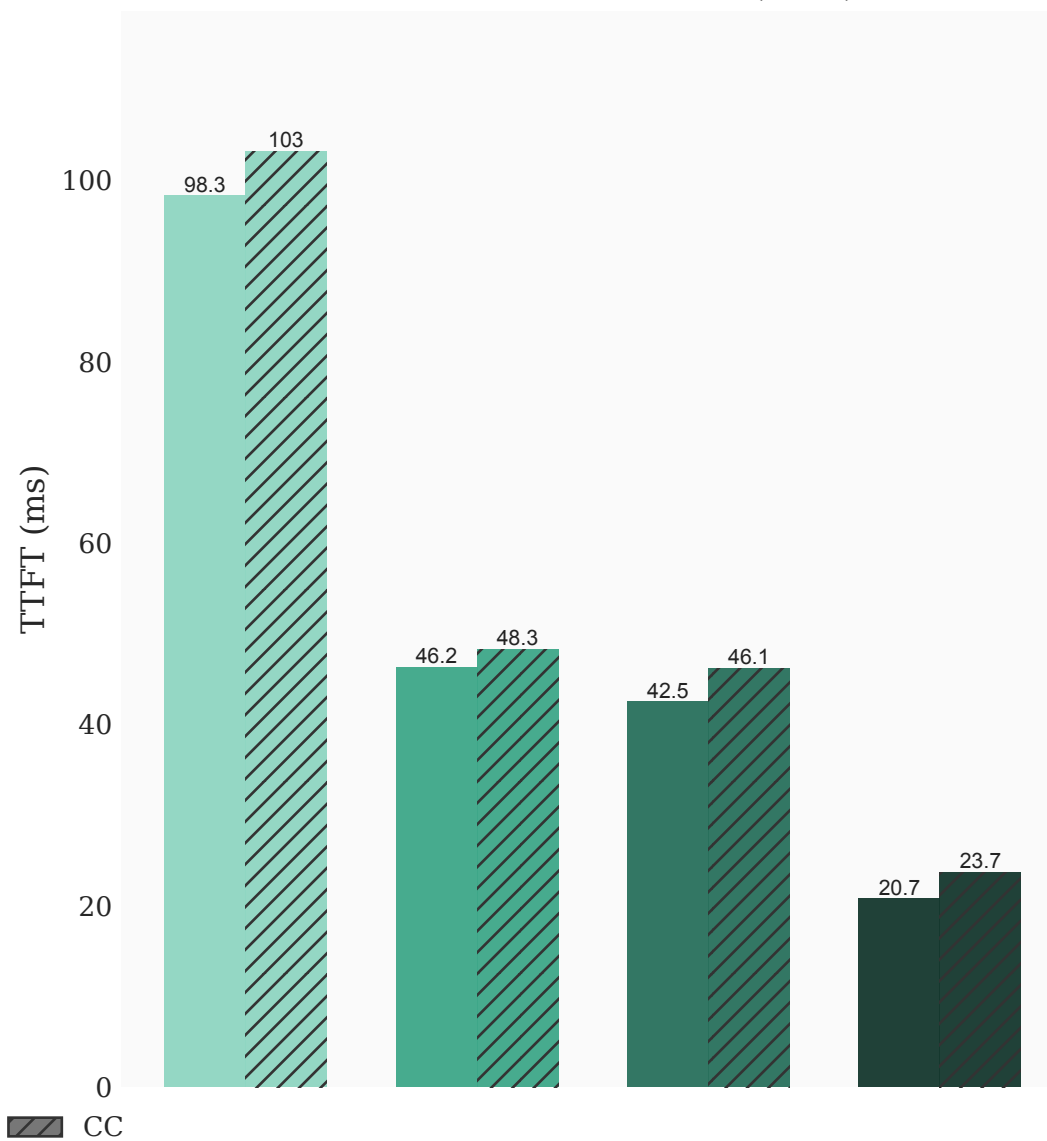
Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B

Numina Math (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



Legend: Llama 3.3 70B Int4, GPT OSS 120B, Mistral 3.1 24B, Llama 3.1 8B