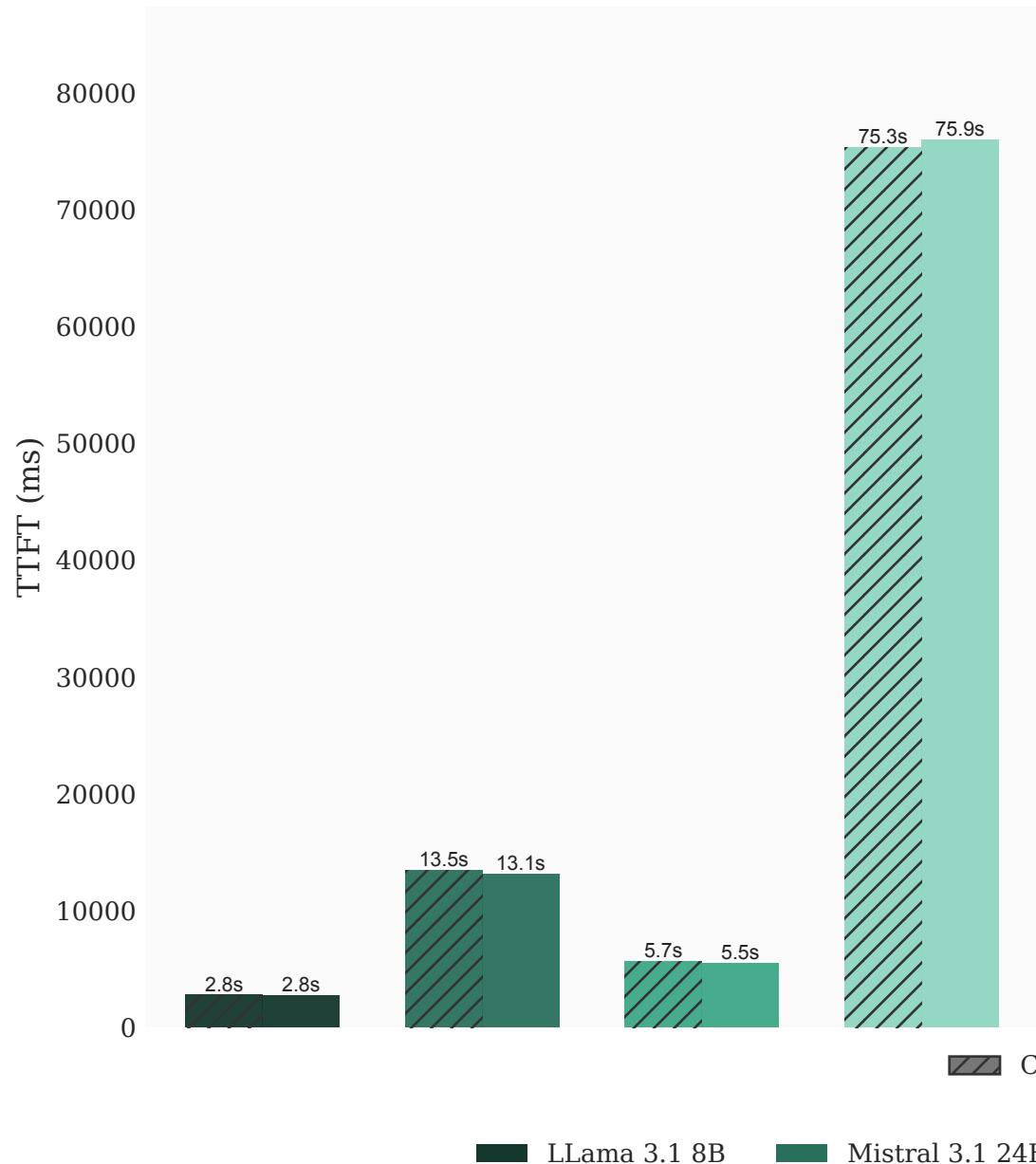
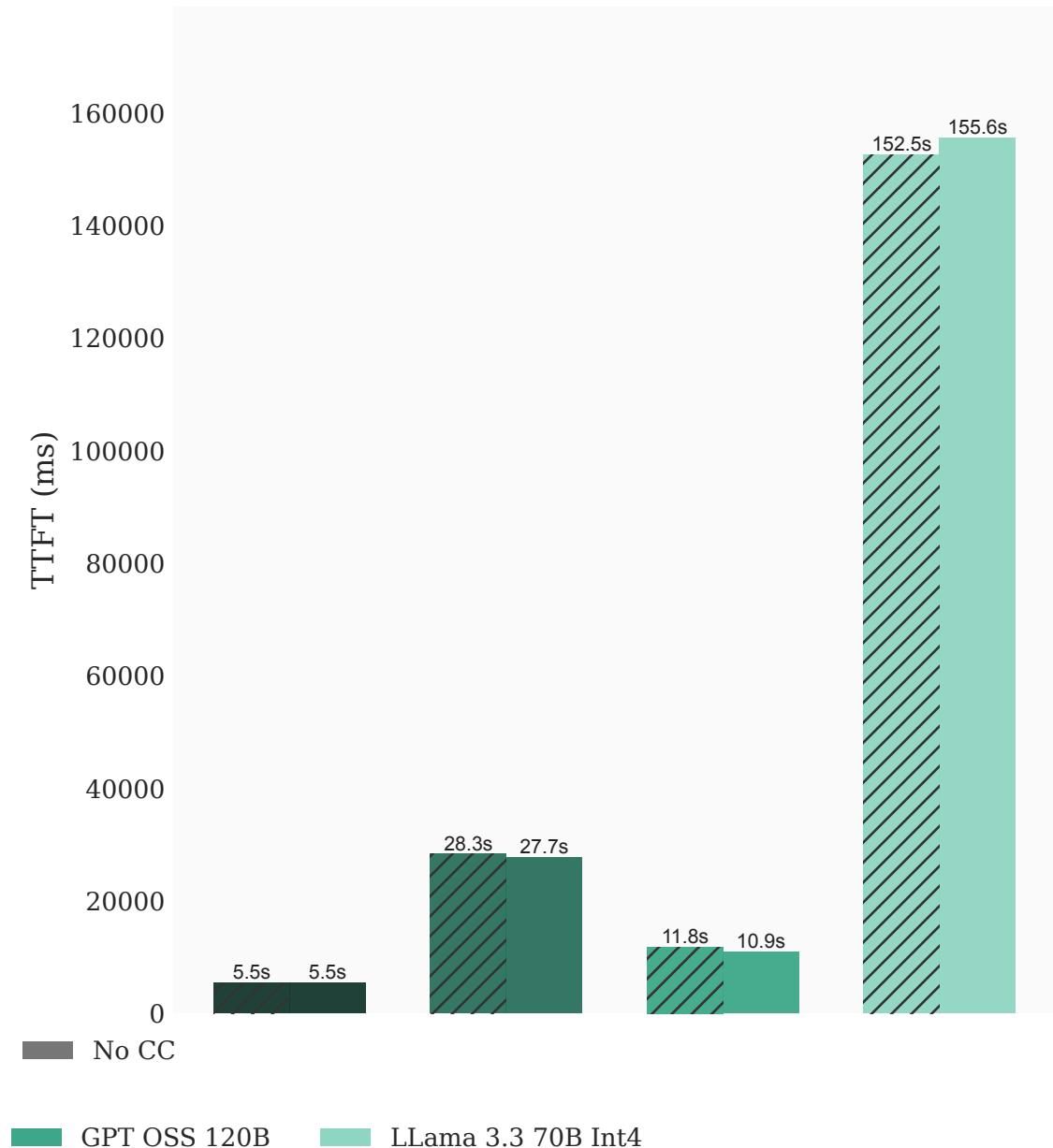


Random (1500 \Rightarrow 250) (100 Request Rate)

Time to First Token (Mean)

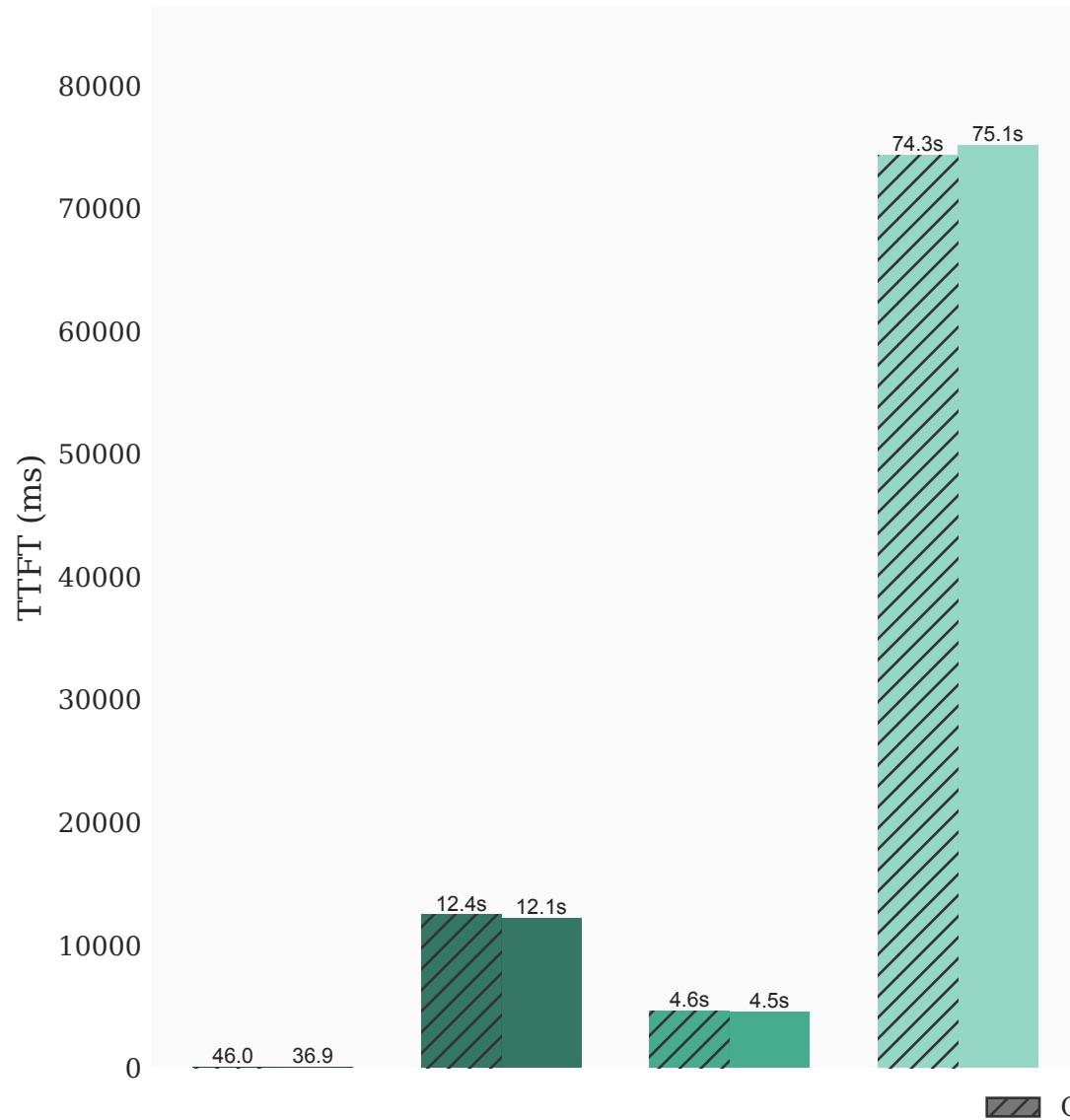


Time to First Token (P99)

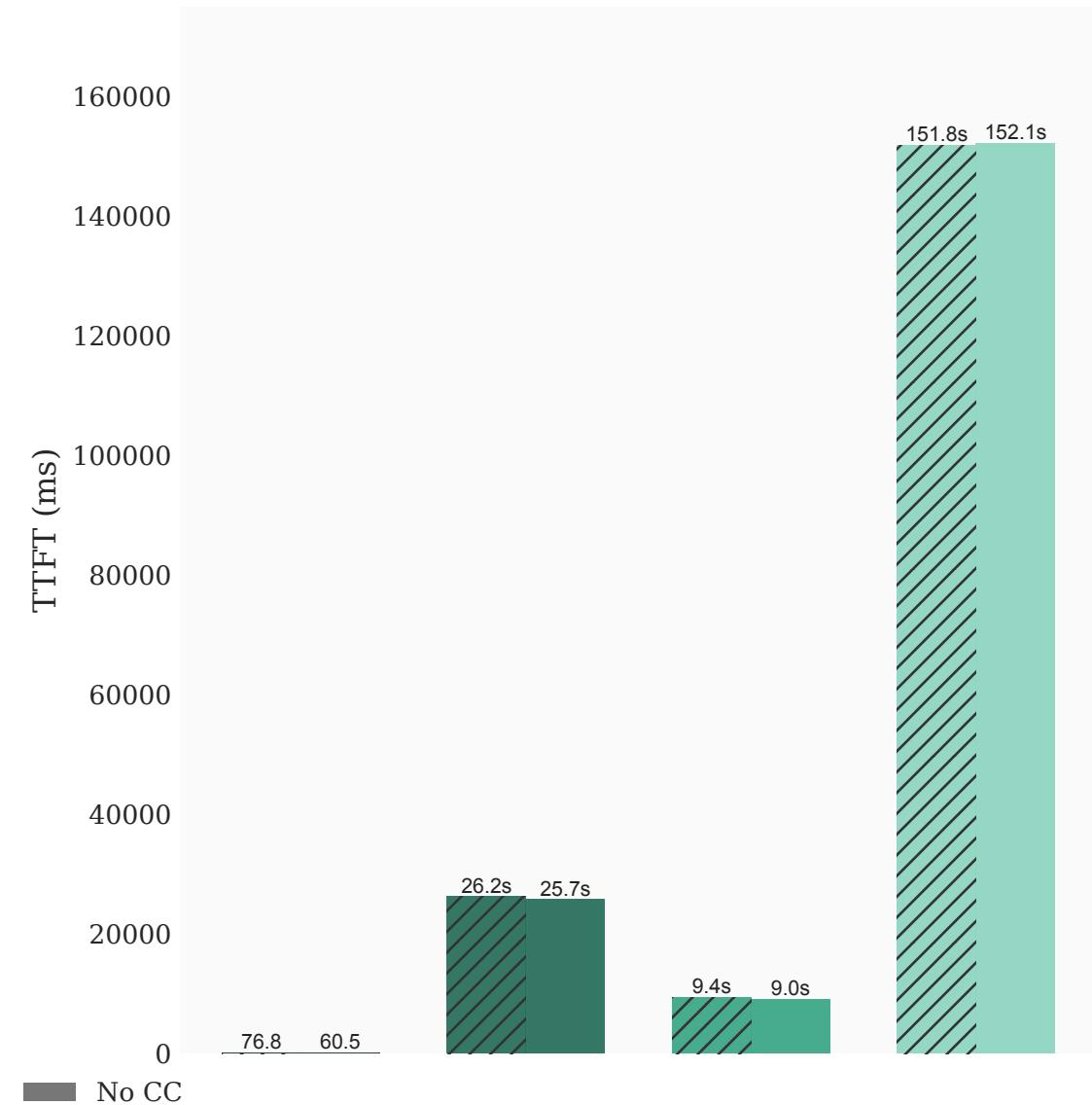


Random (1500 \Rightarrow 250) (50 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)



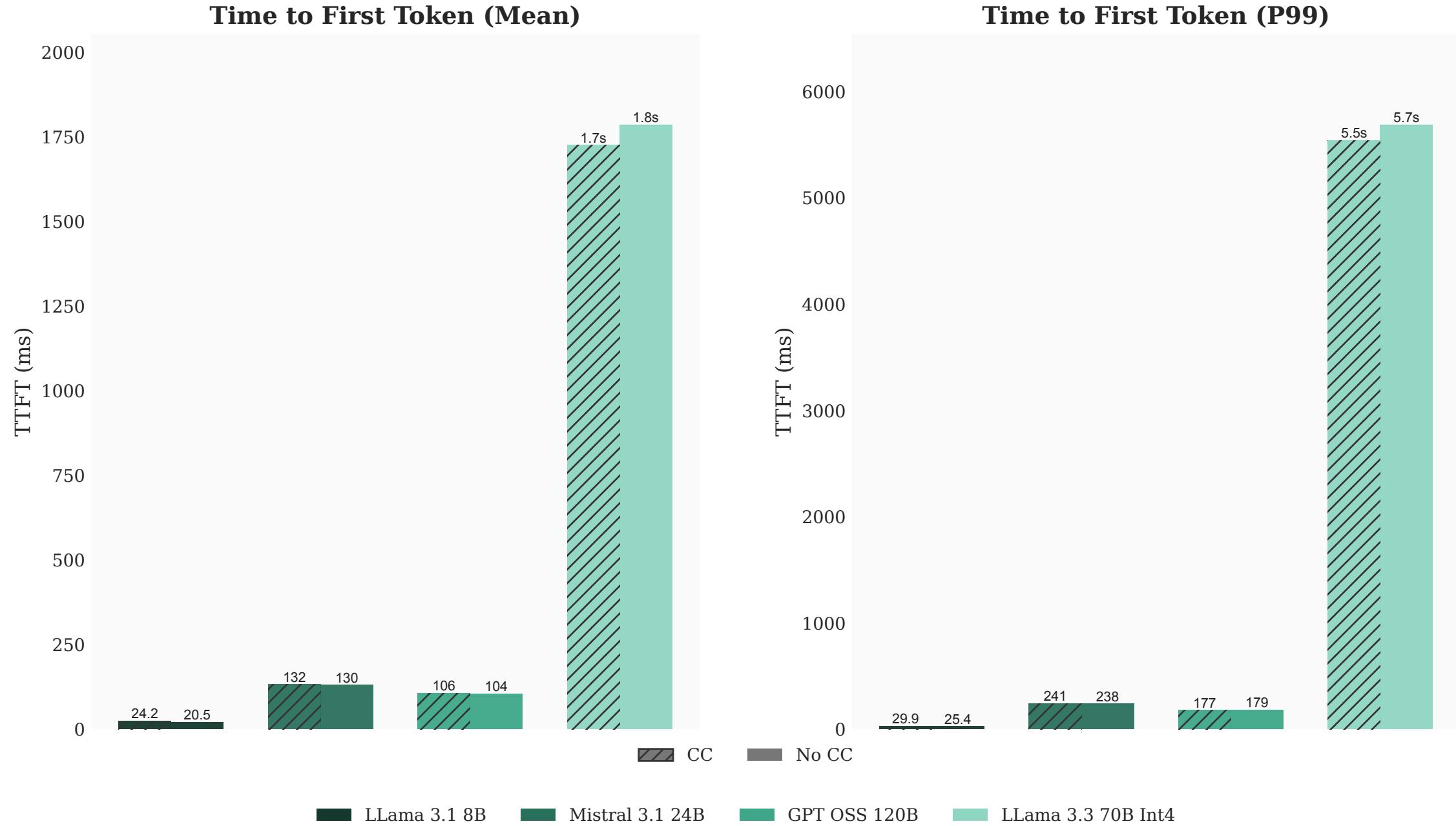
■ LLama 3.1 8B

■ Mistral 3.1 24B

■ GPT OSS 120B

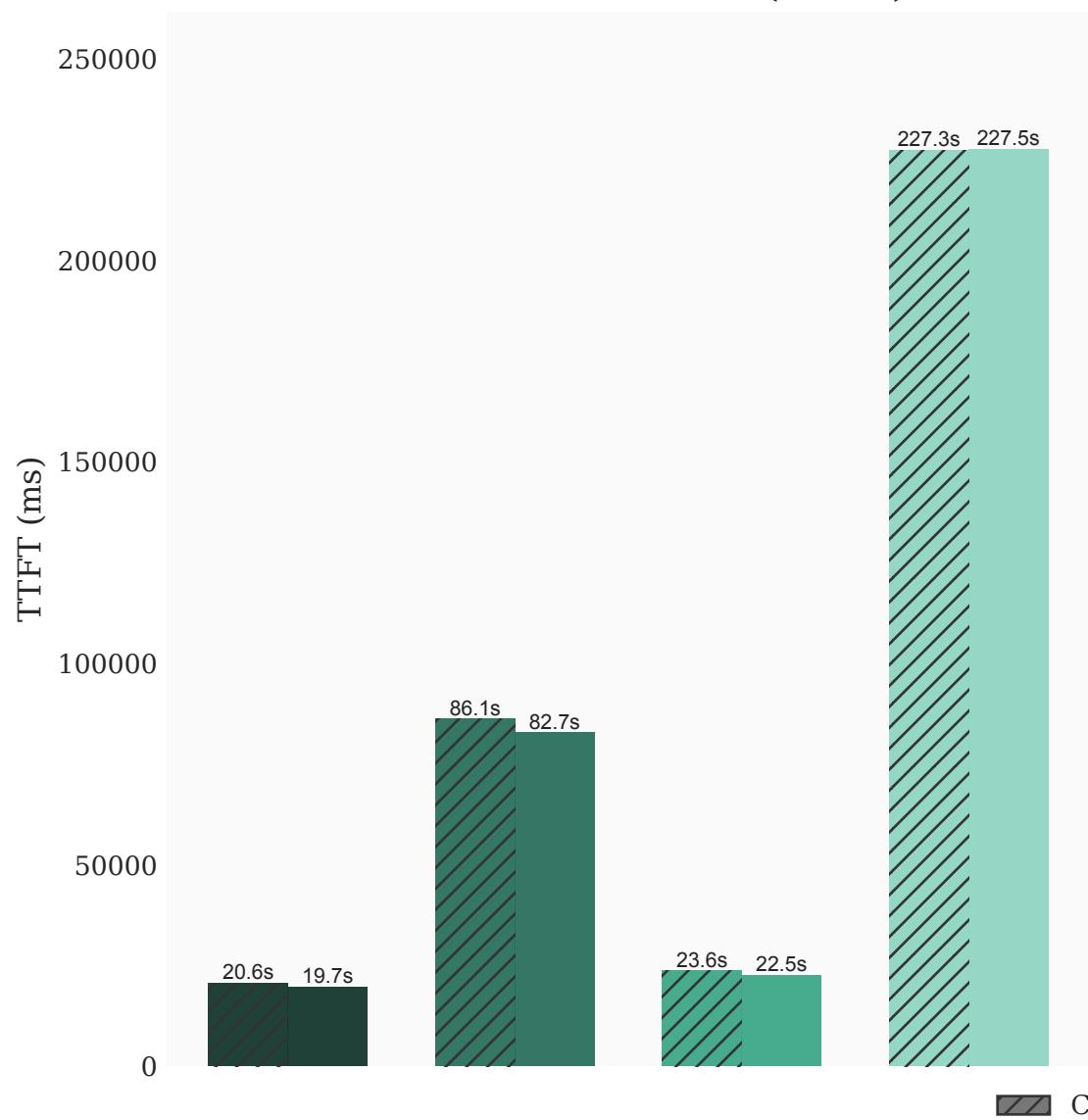
■ LLama 3.3 70B Int4

Random (1500 ⇒ 250) (Single Request)

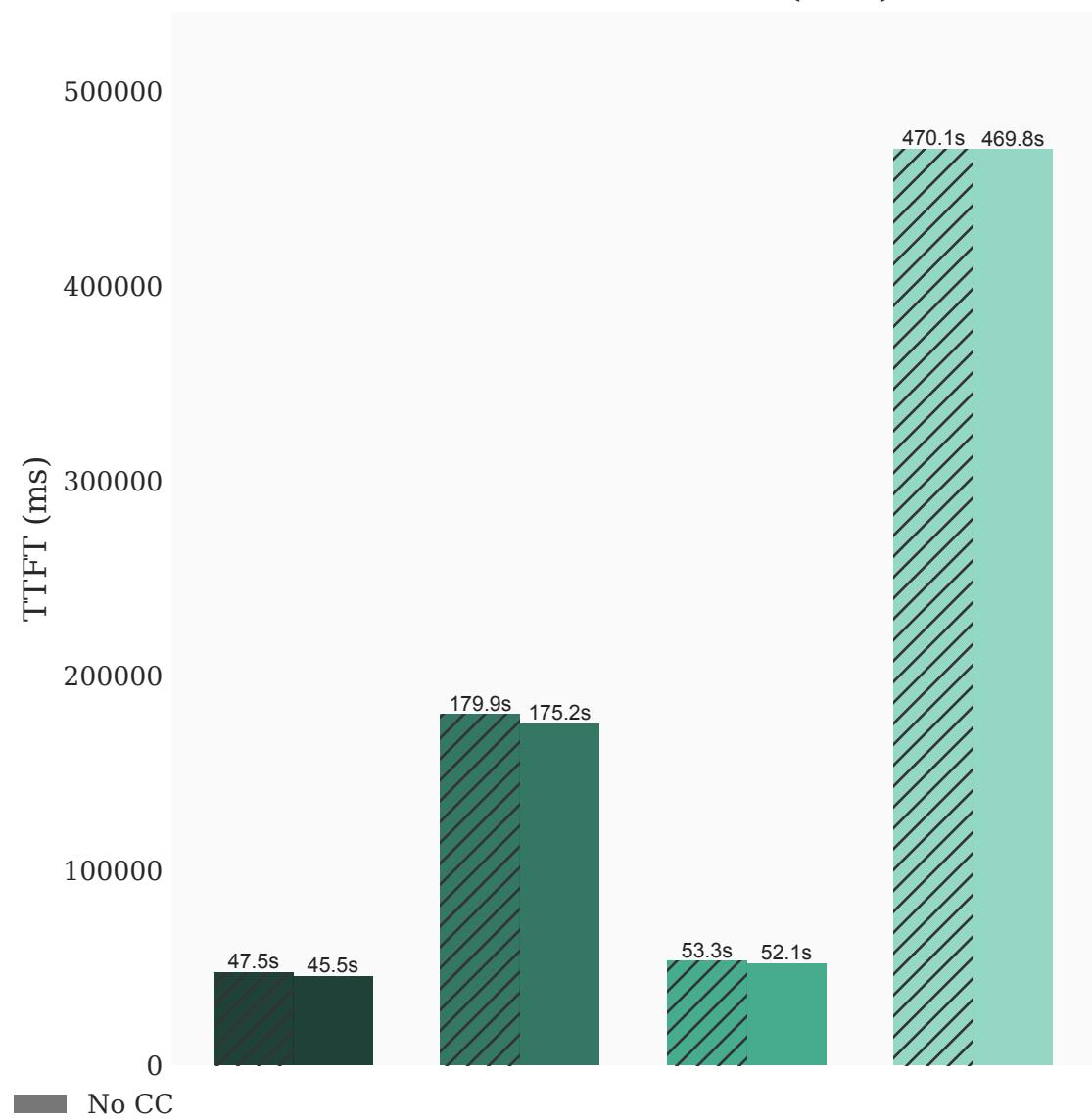


Random (4000 \Rightarrow 1000) (100 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

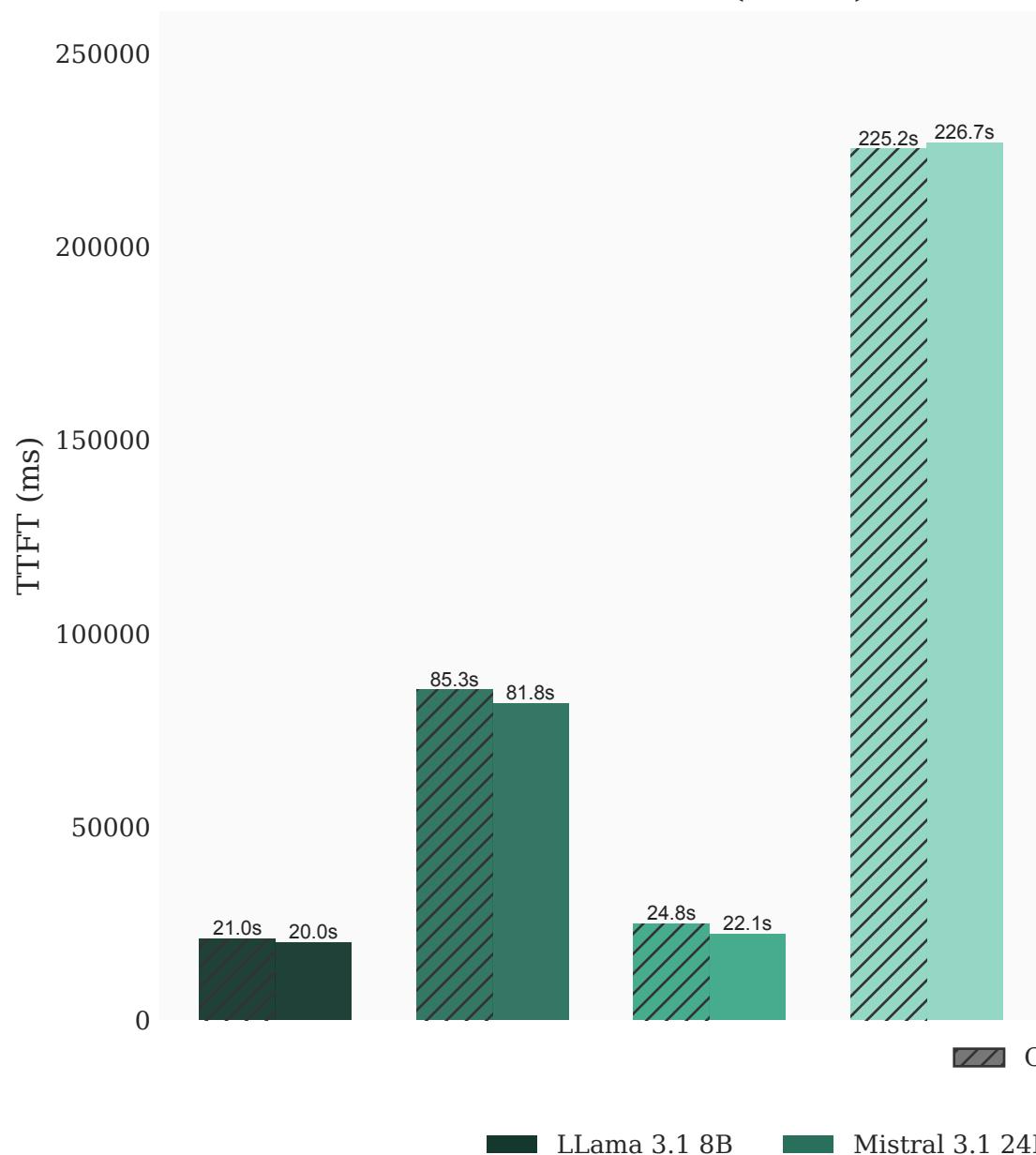
■ Mistral 3.1 24B

■ GPT OSS 120B

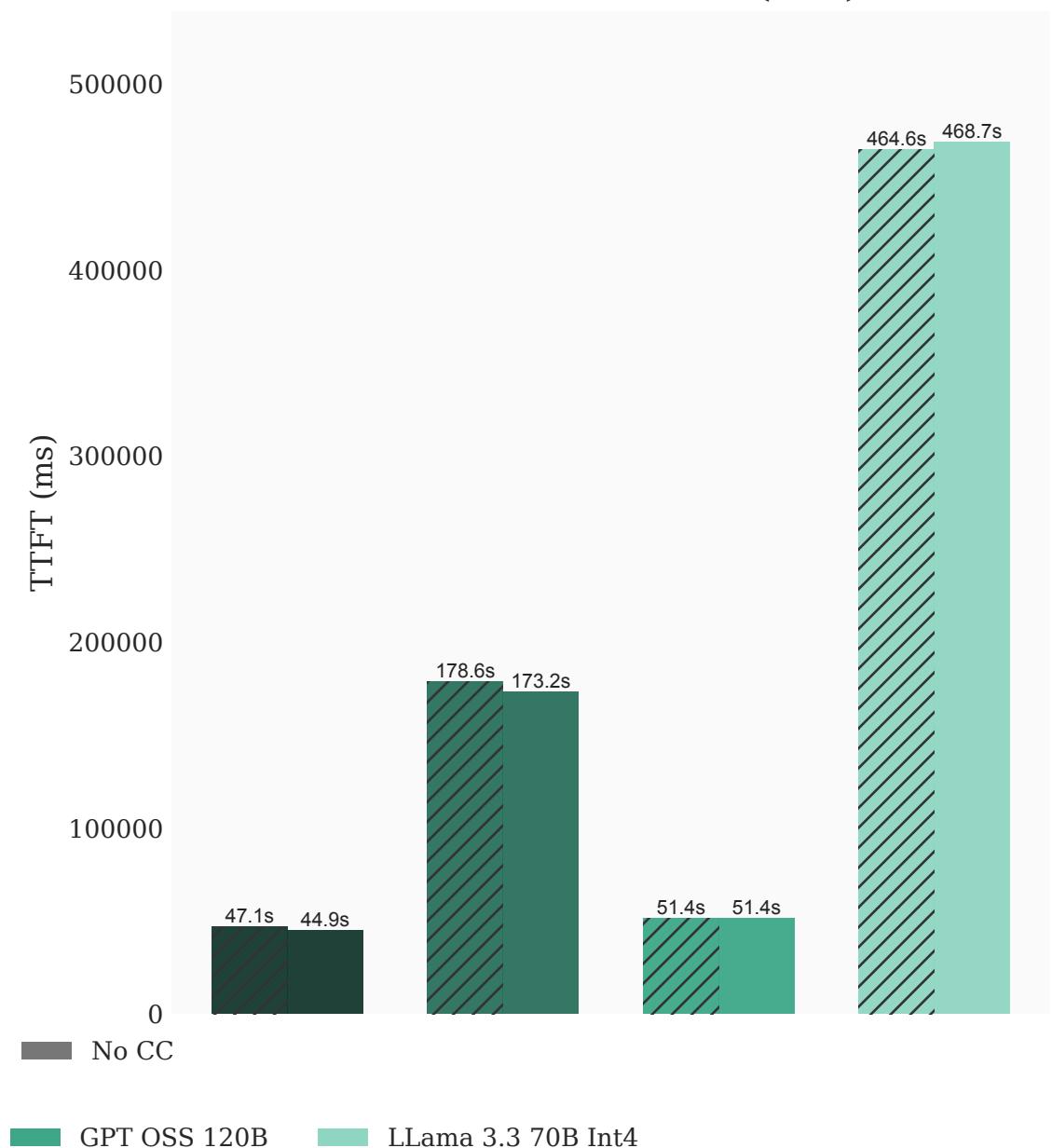
■ LLama 3.3 70B Int4

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

Time to First Token (Mean)

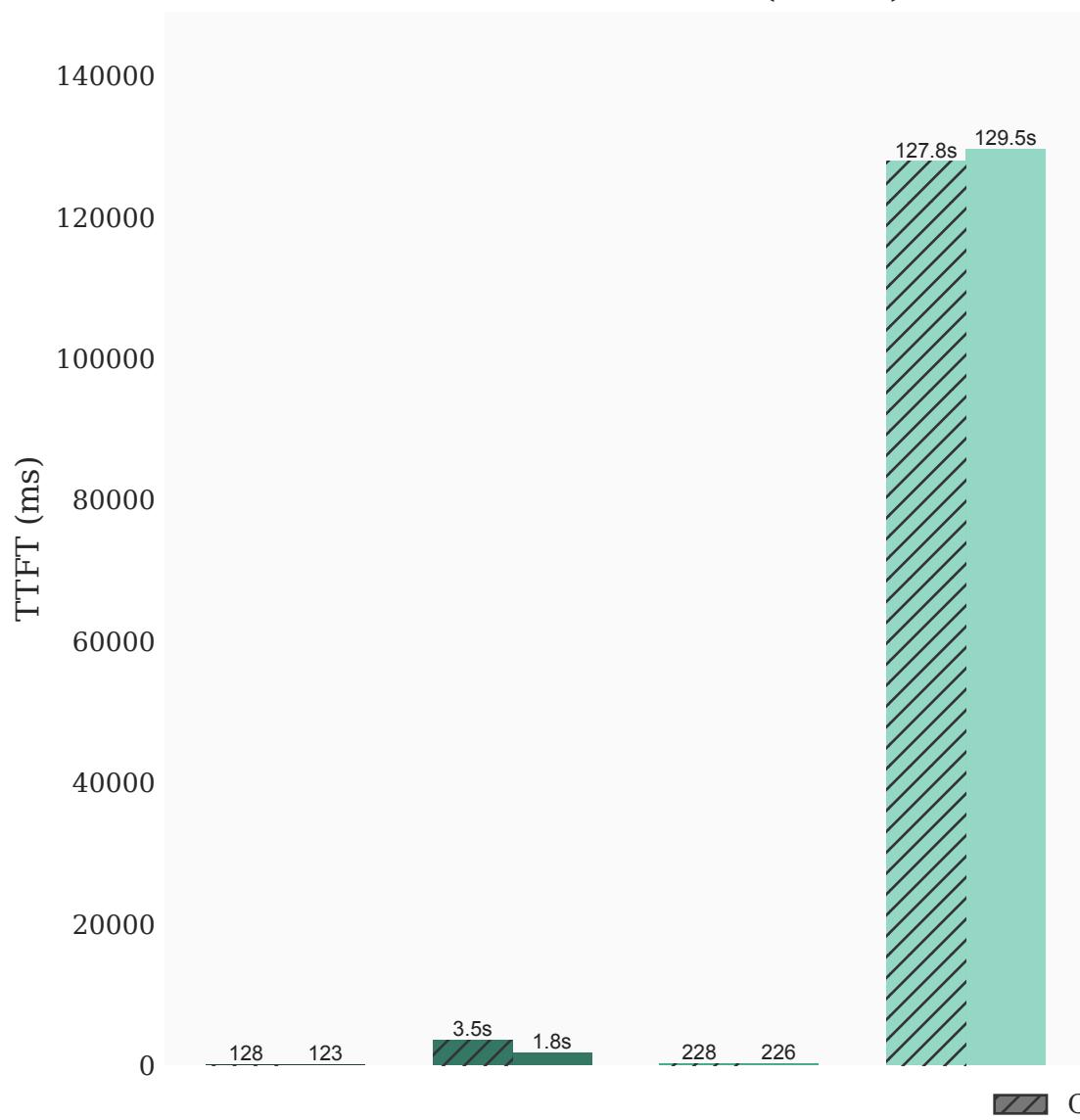


Time to First Token (P99)

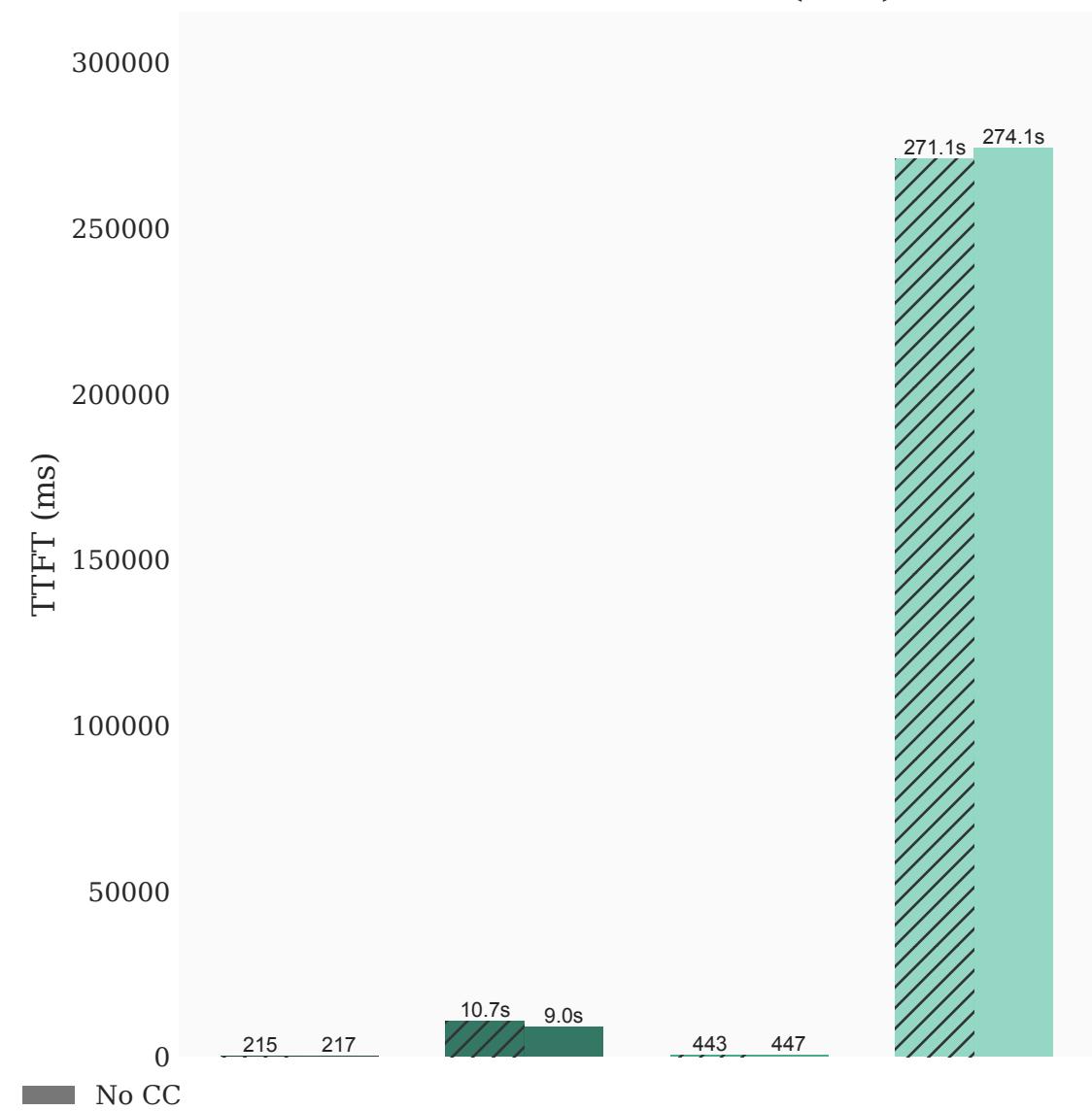


Random (4000 \Rightarrow 1000) (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

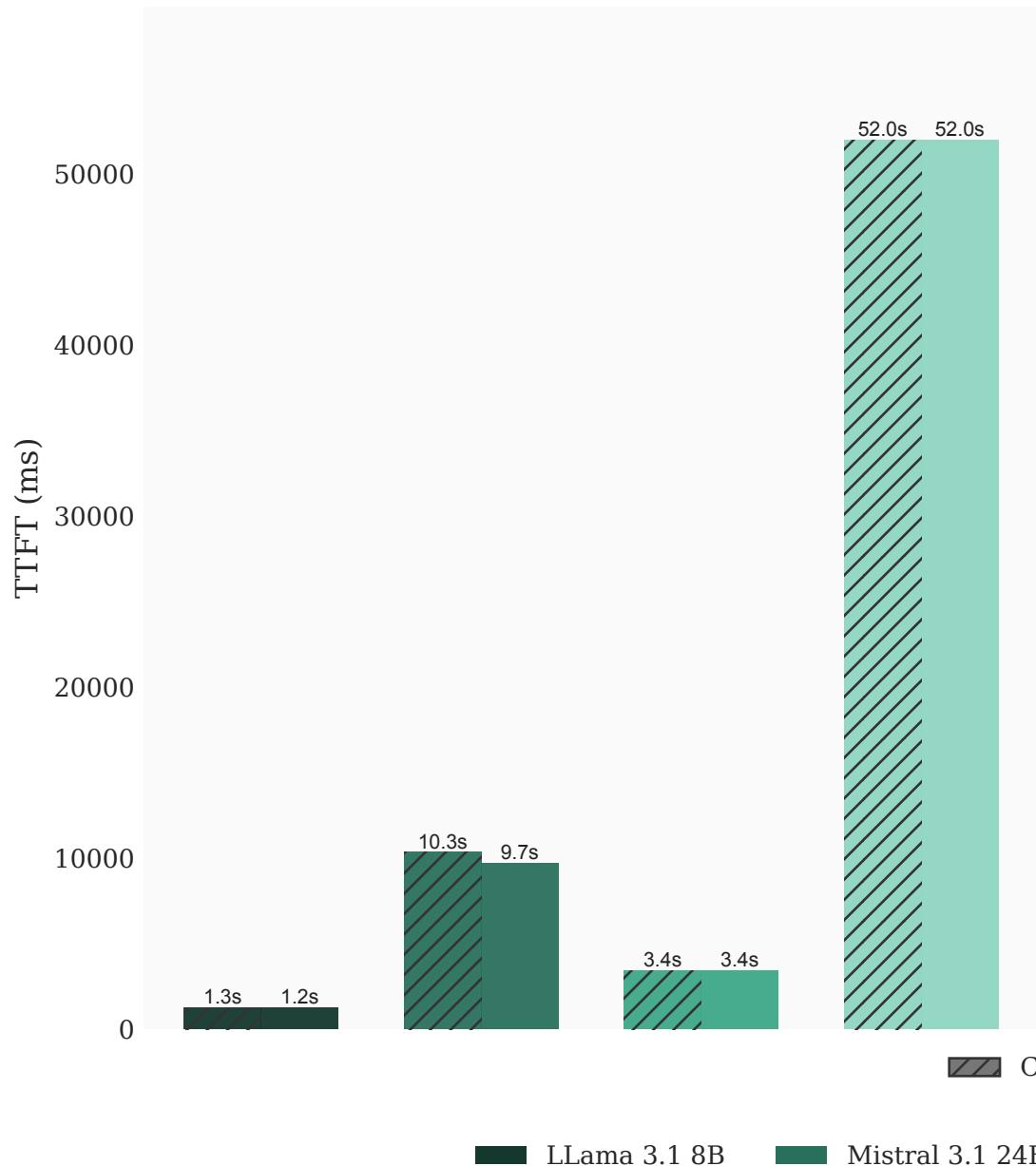
■ Mistral 3.1 24B

■ GPT OSS 120B

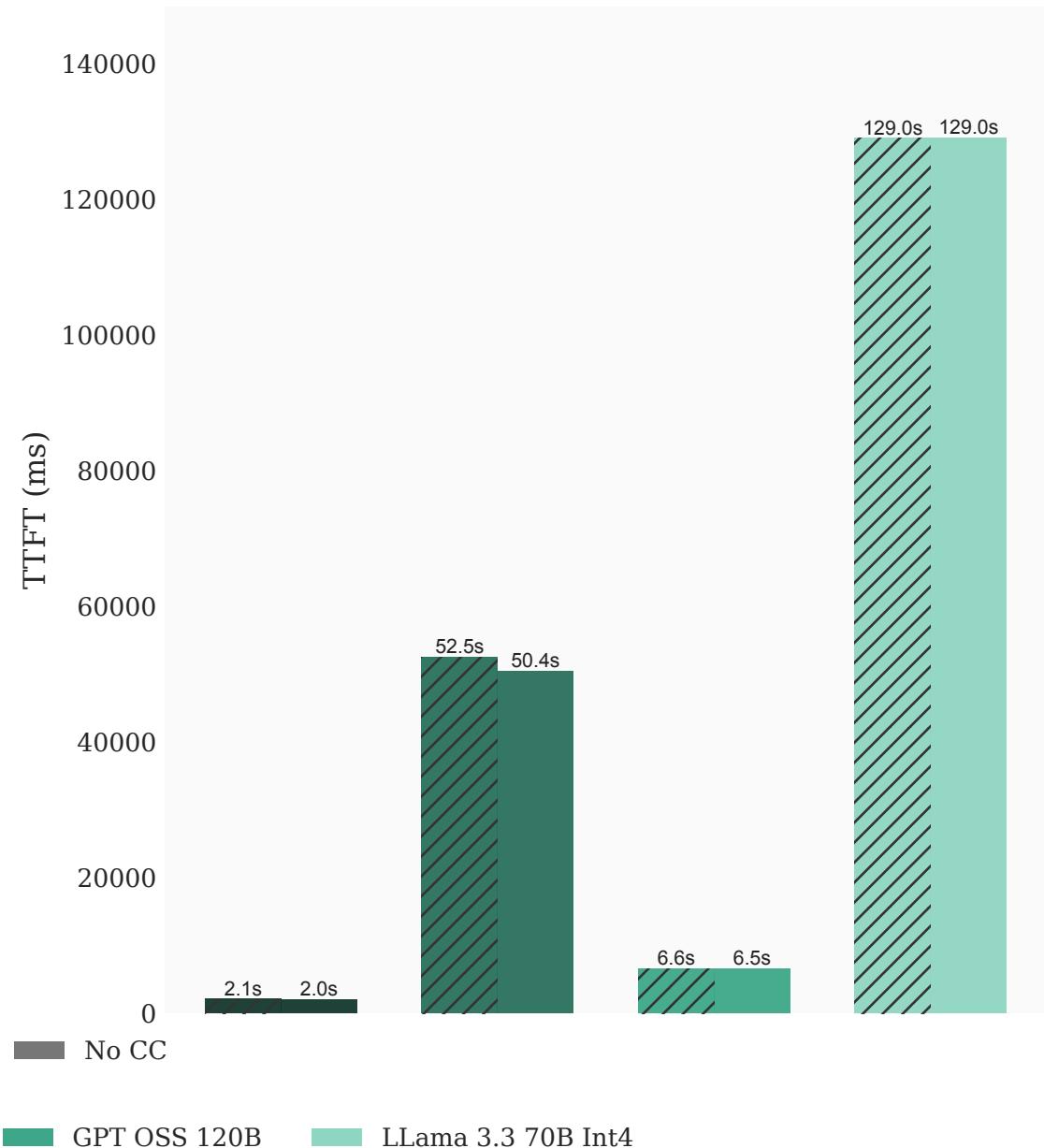
■ LLama 3.3 70B Int4

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

Time to First Token (Mean)

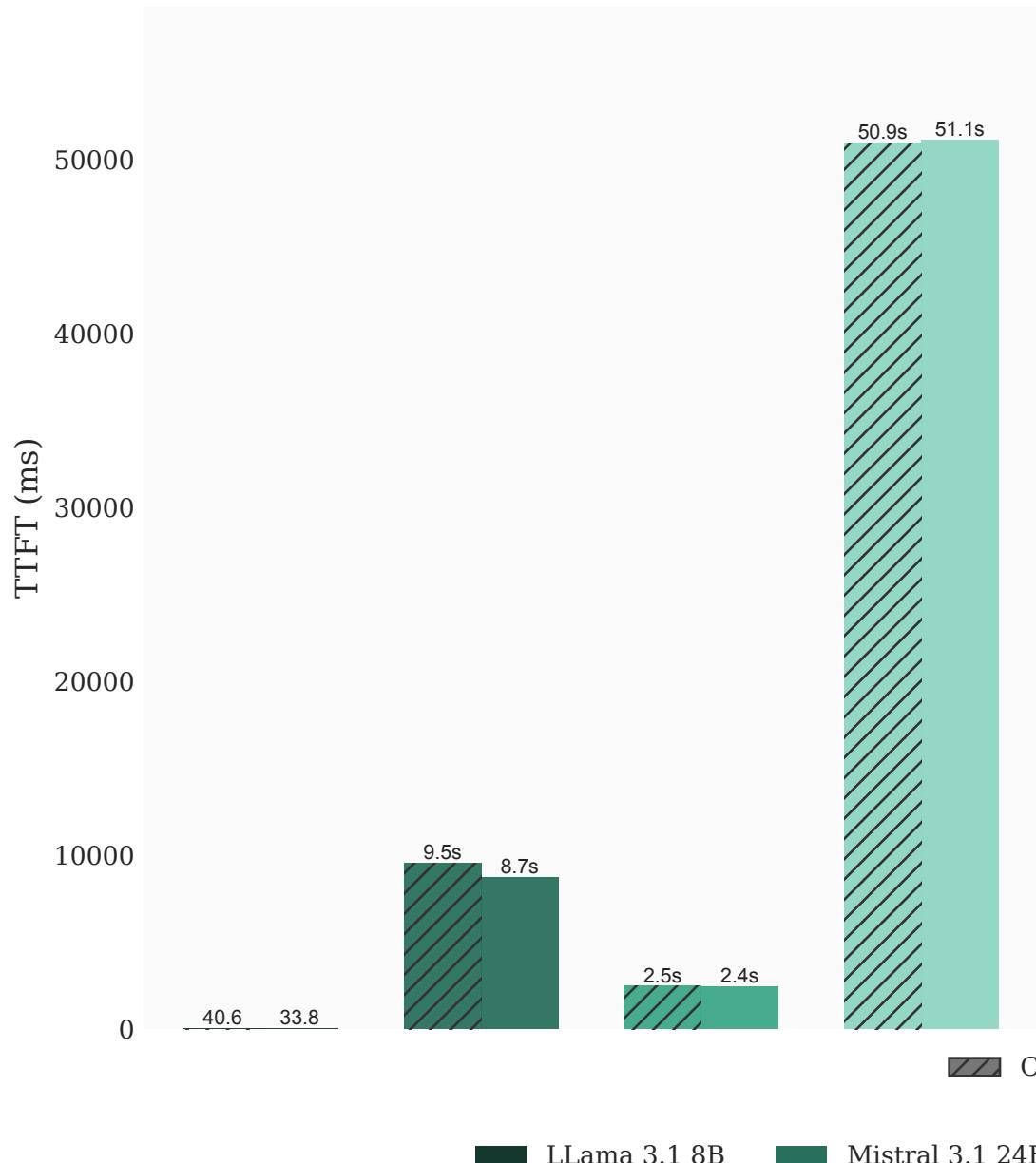


Time to First Token (P99)

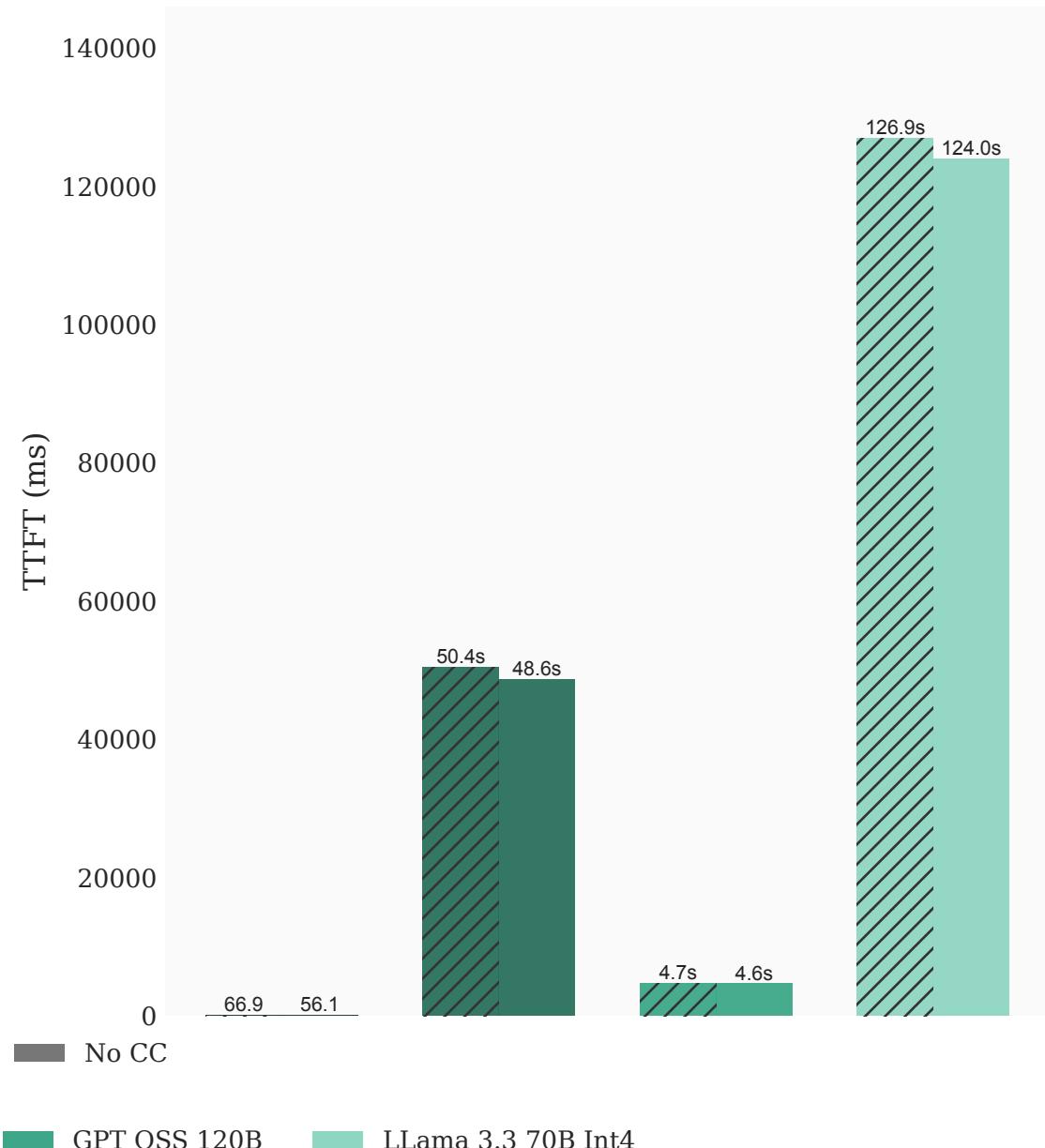


Random (1000 → 1000) (50 Request Rate)

Time to First Token (Mean)

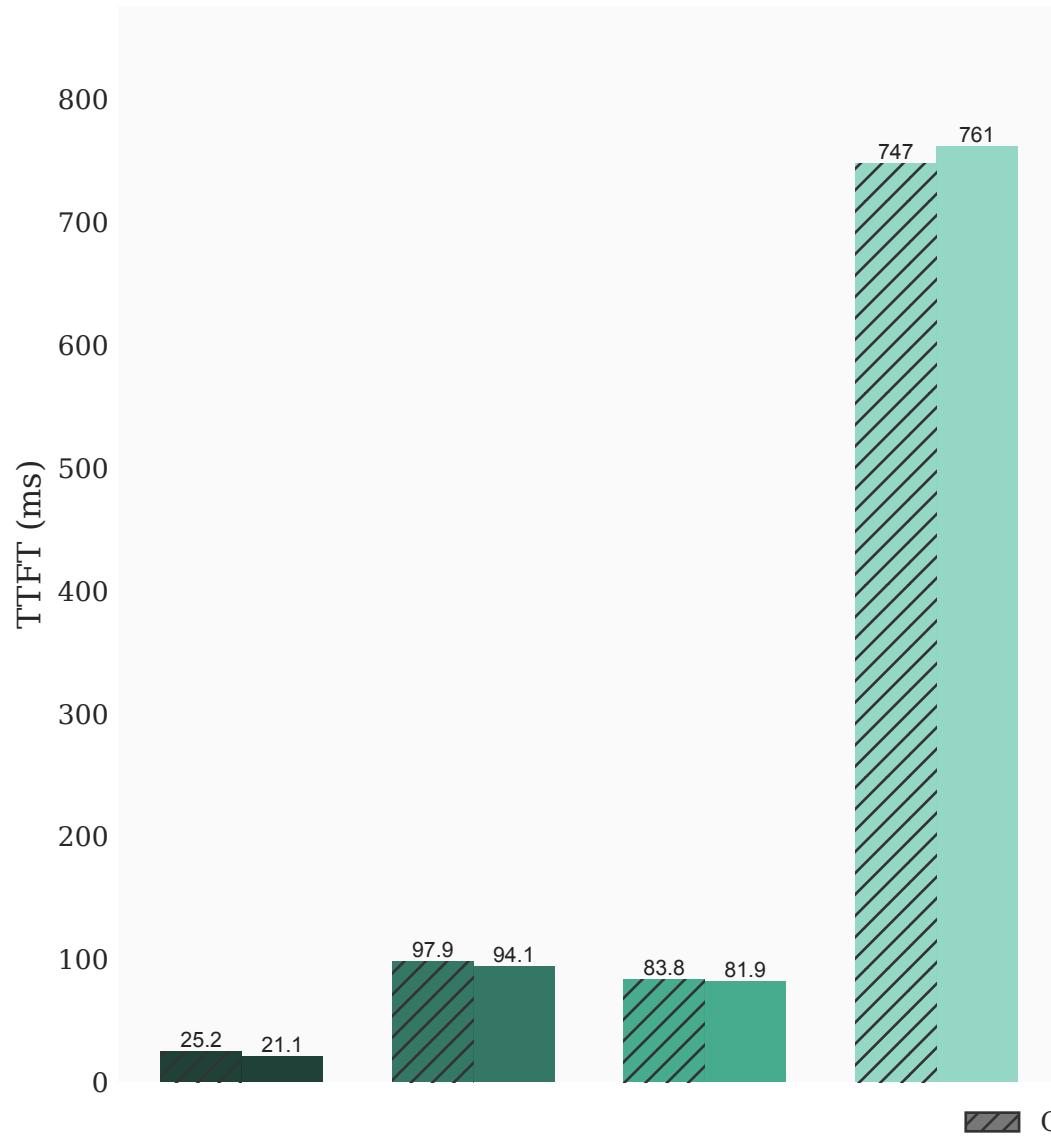


Time to First Token (P99)

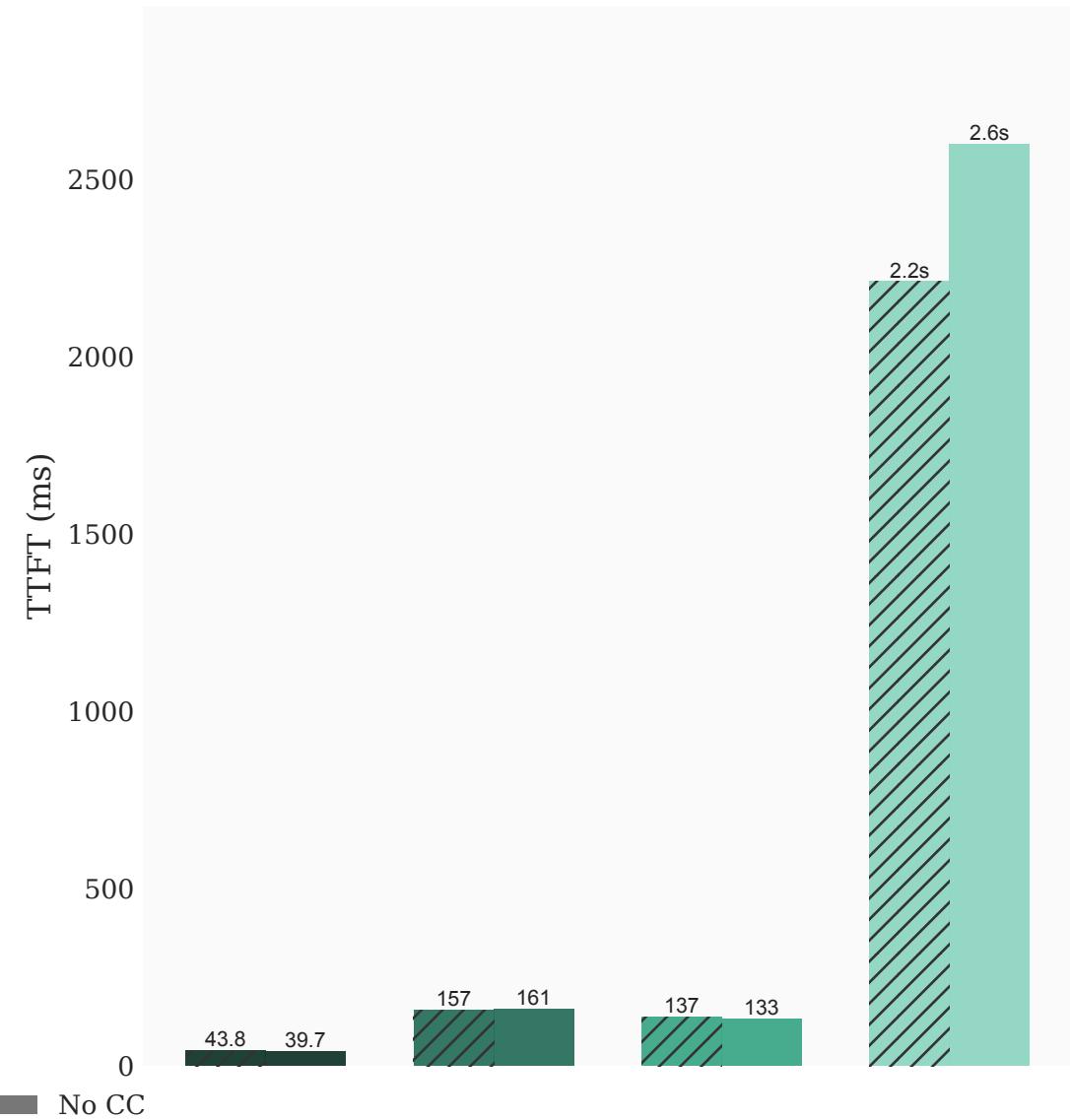


Random (1000 \Rightarrow 1000) (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

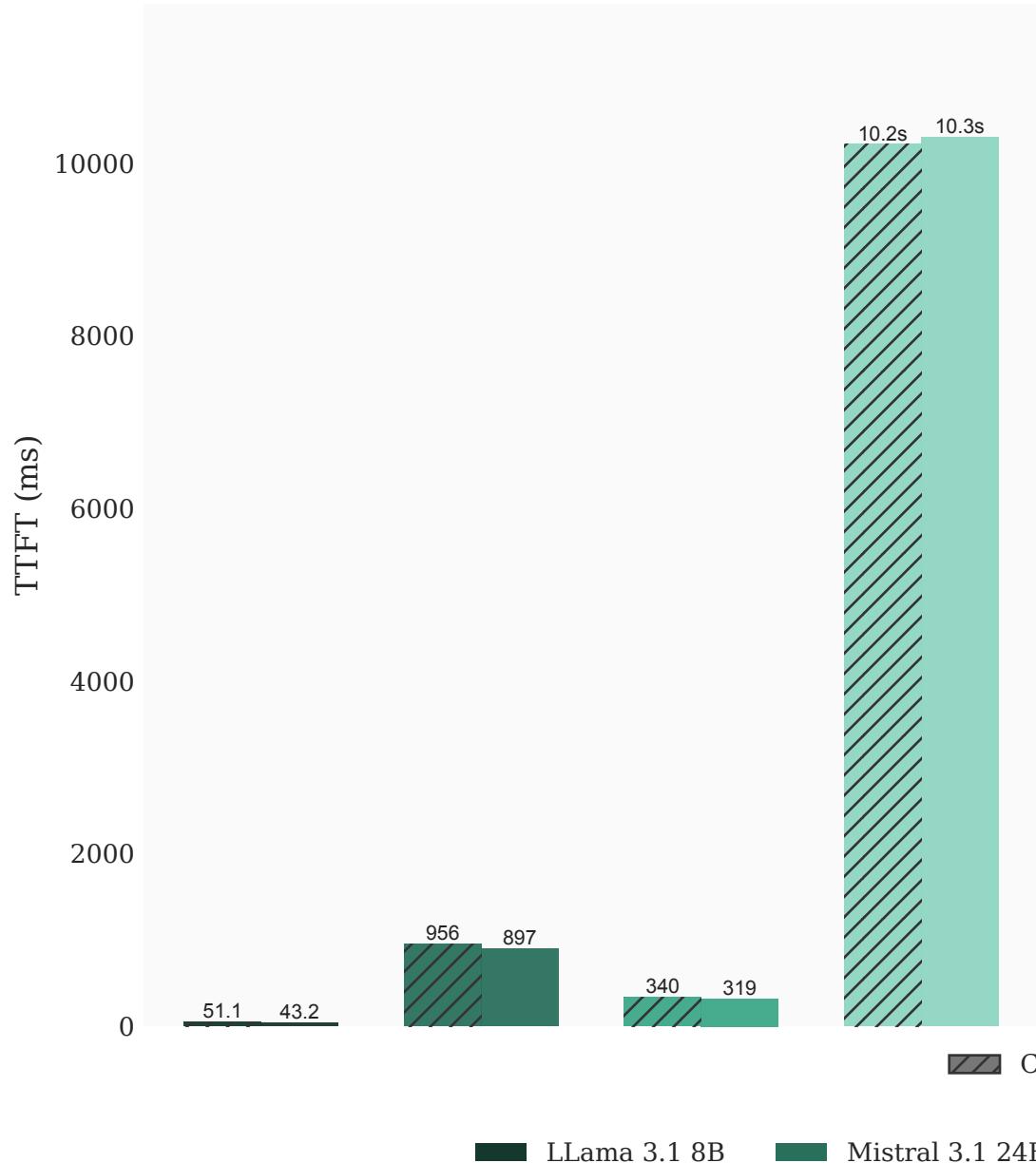
■ Mistral 3.1 24B

■ GPT OSS 120B

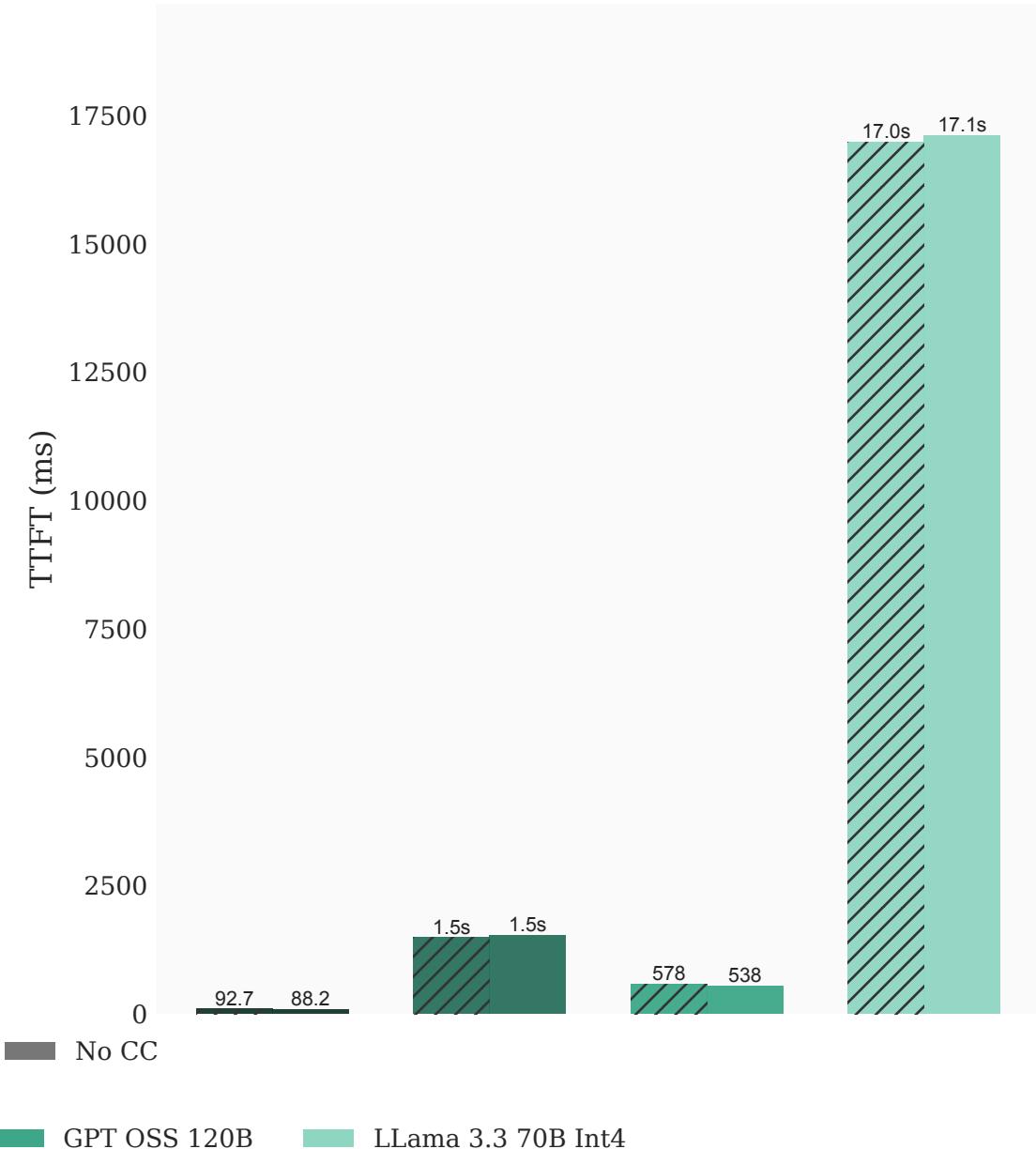
■ LLama 3.3 70B Int4

ShareGPT (100 Request Rate)

Time to First Token (Mean)

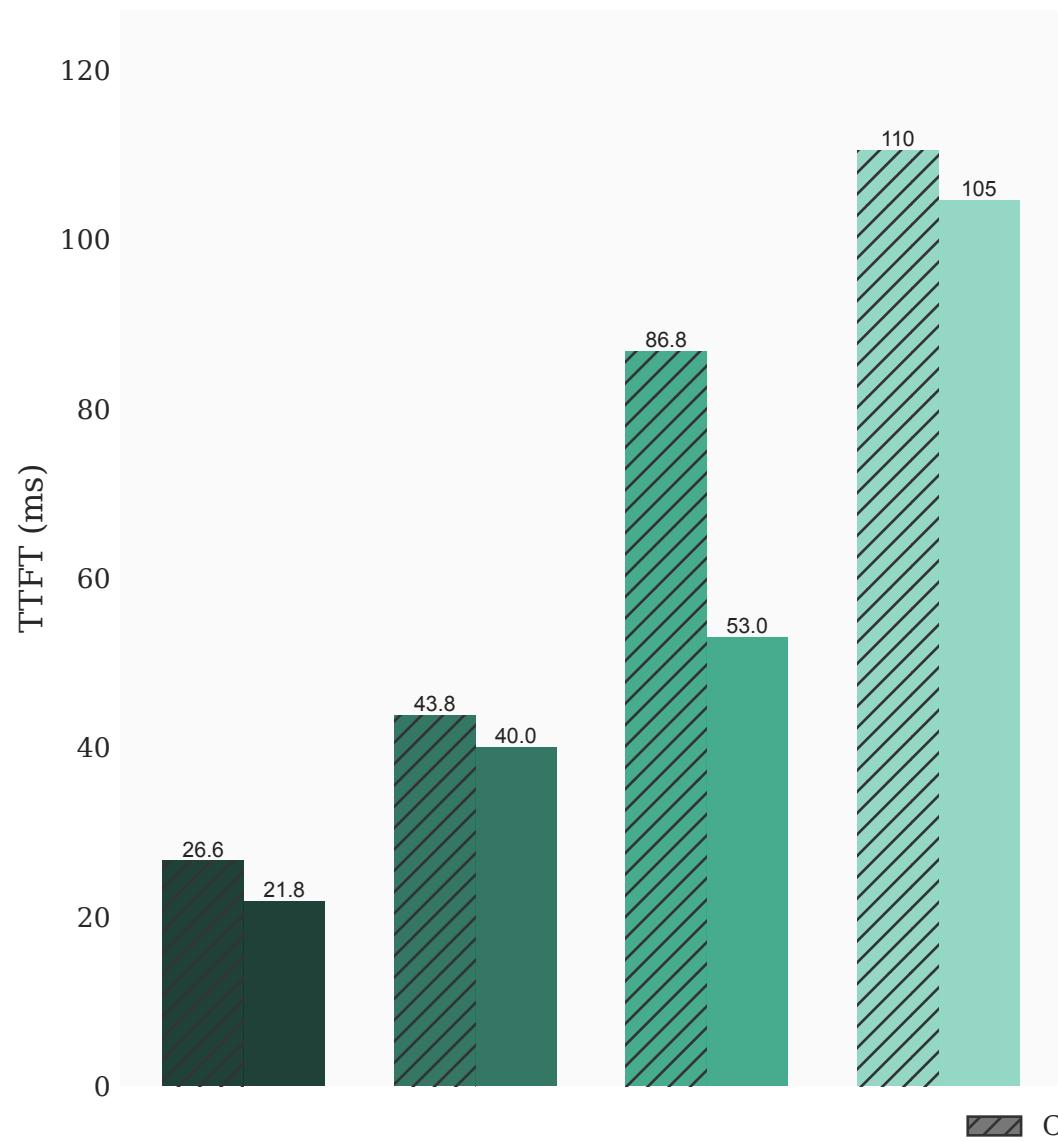


Time to First Token (P99)

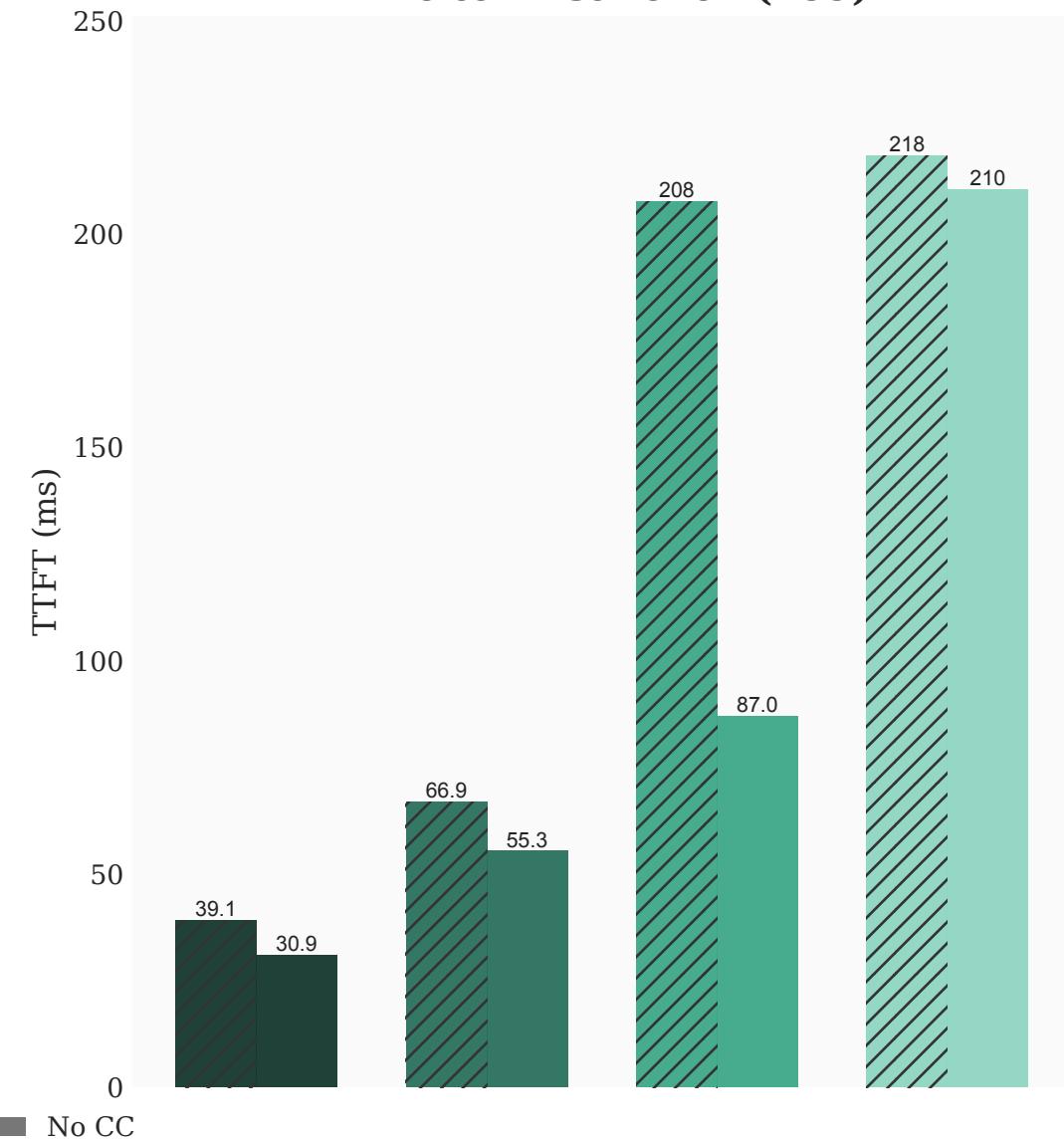


ShareGPT (50 Request Rate)

Time to First Token (Mean)

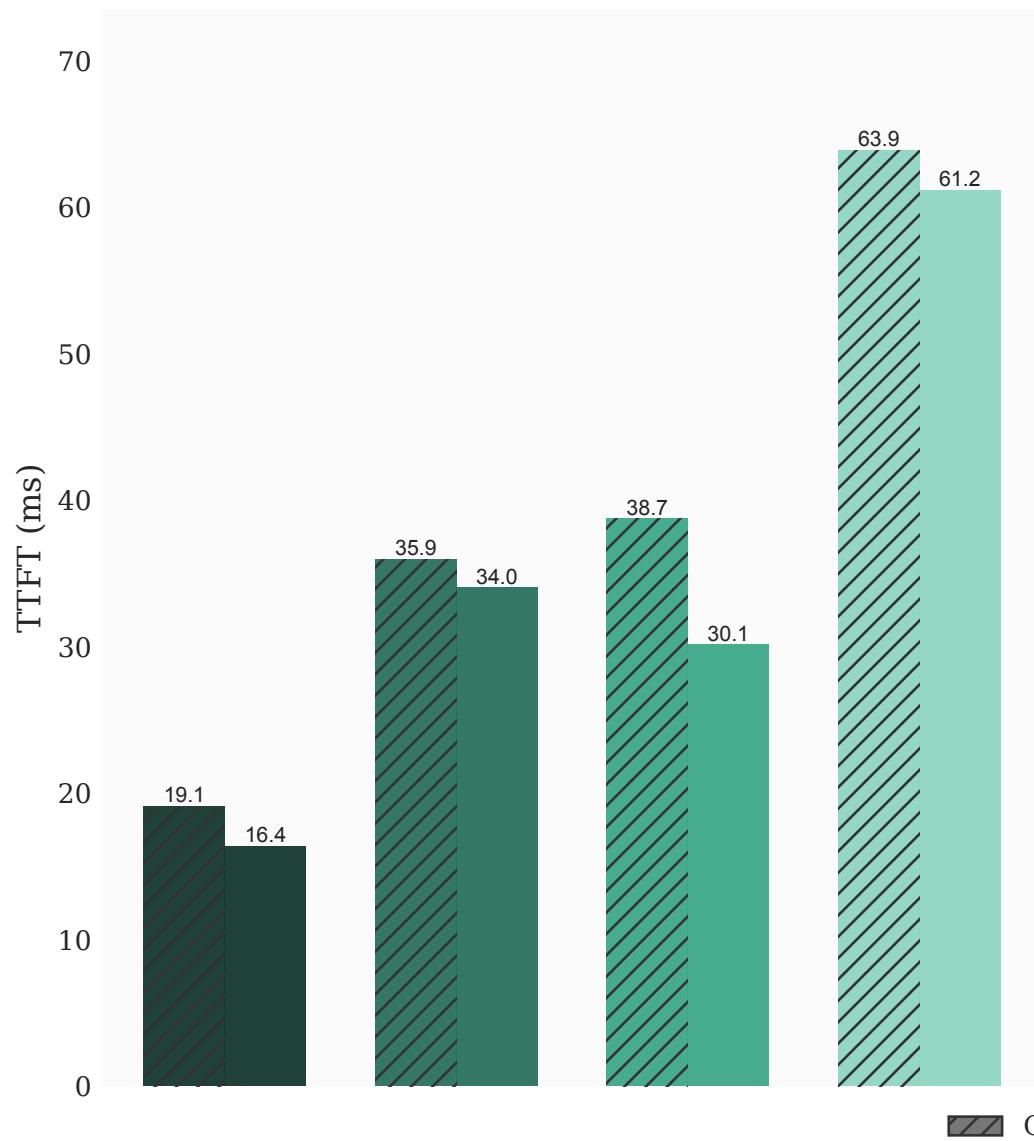


Time to First Token (P99)

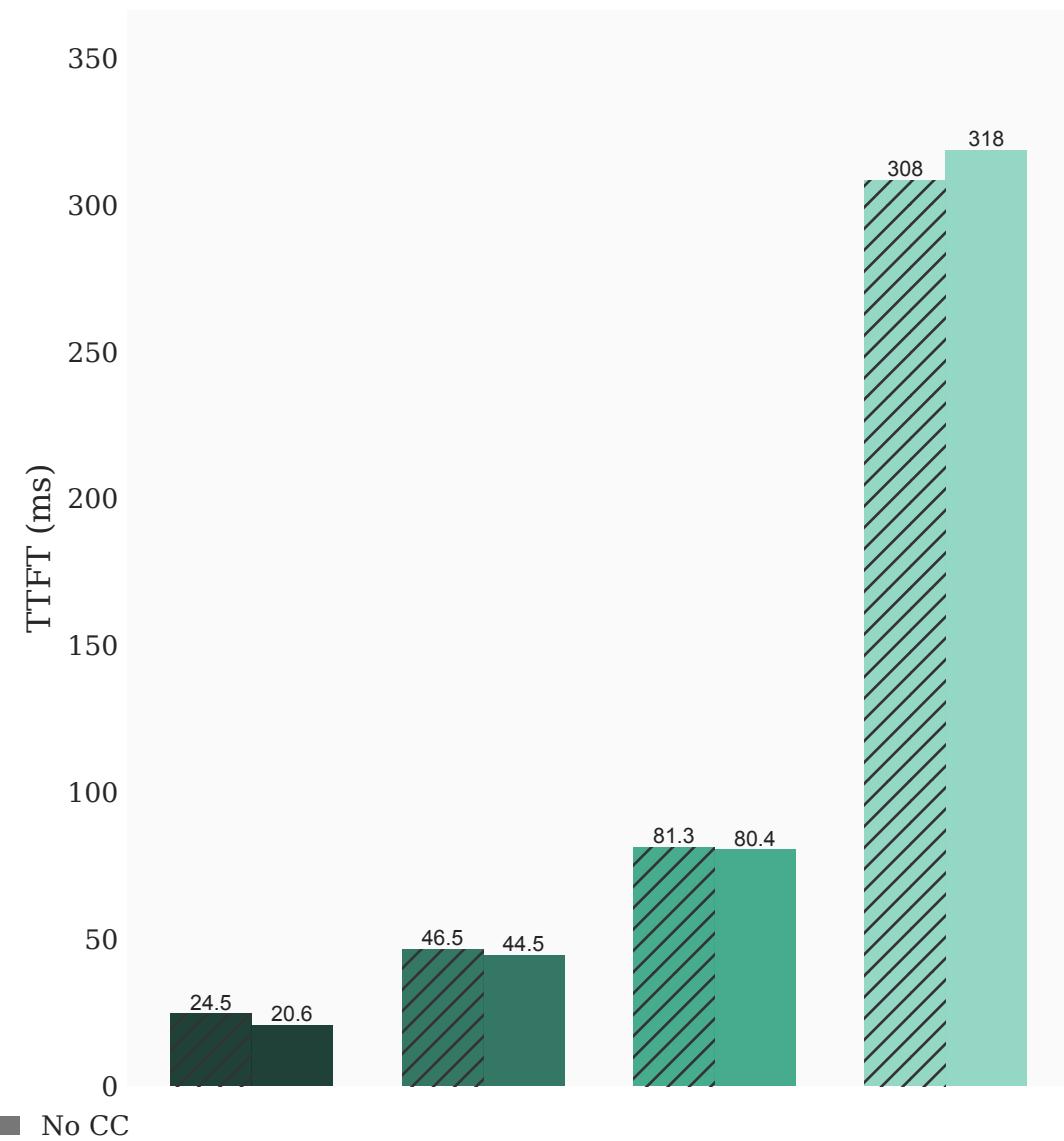


ShareGPT (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

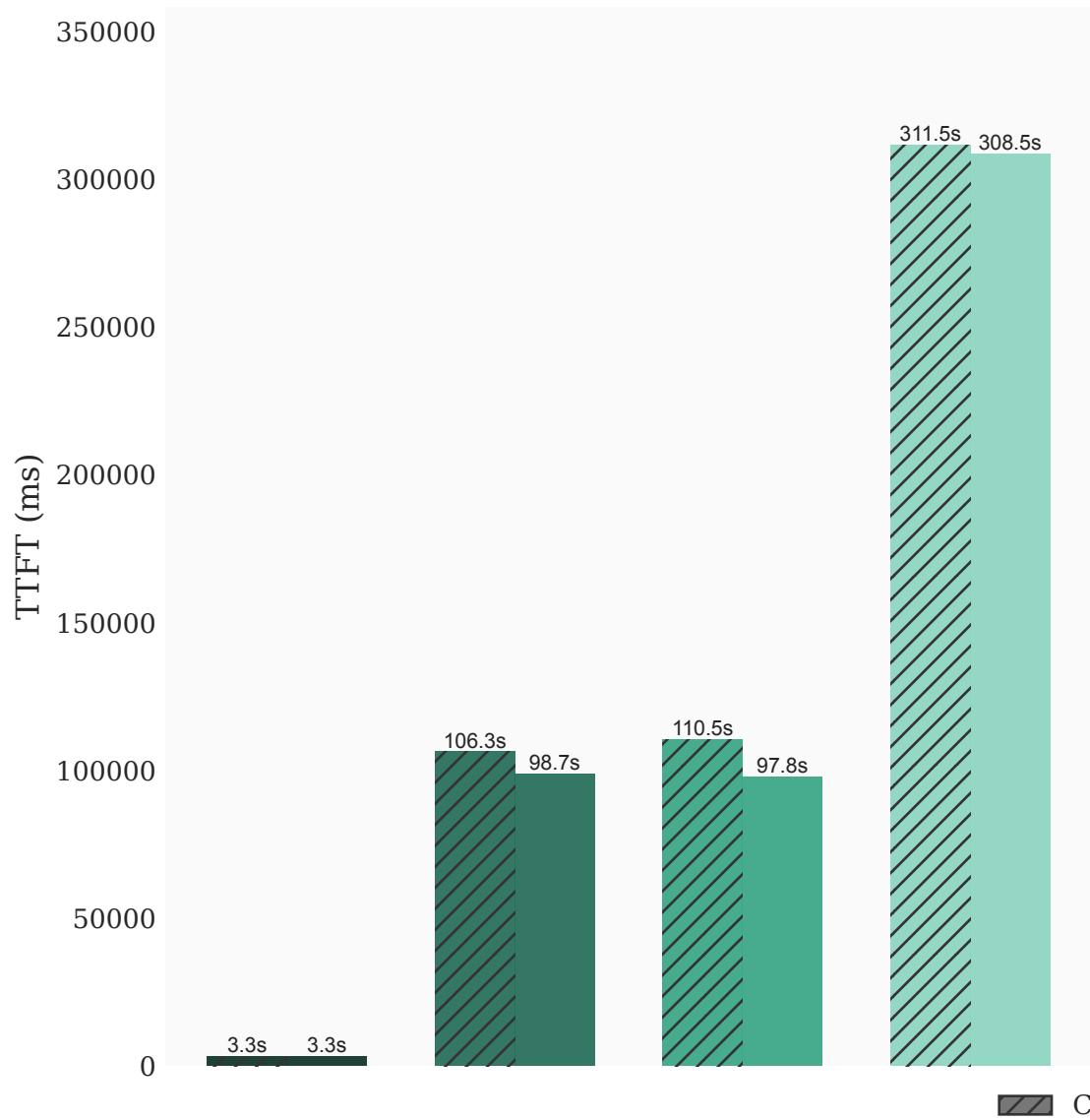
■ Mistral 3.1 24B

■ GPT OSS 120B

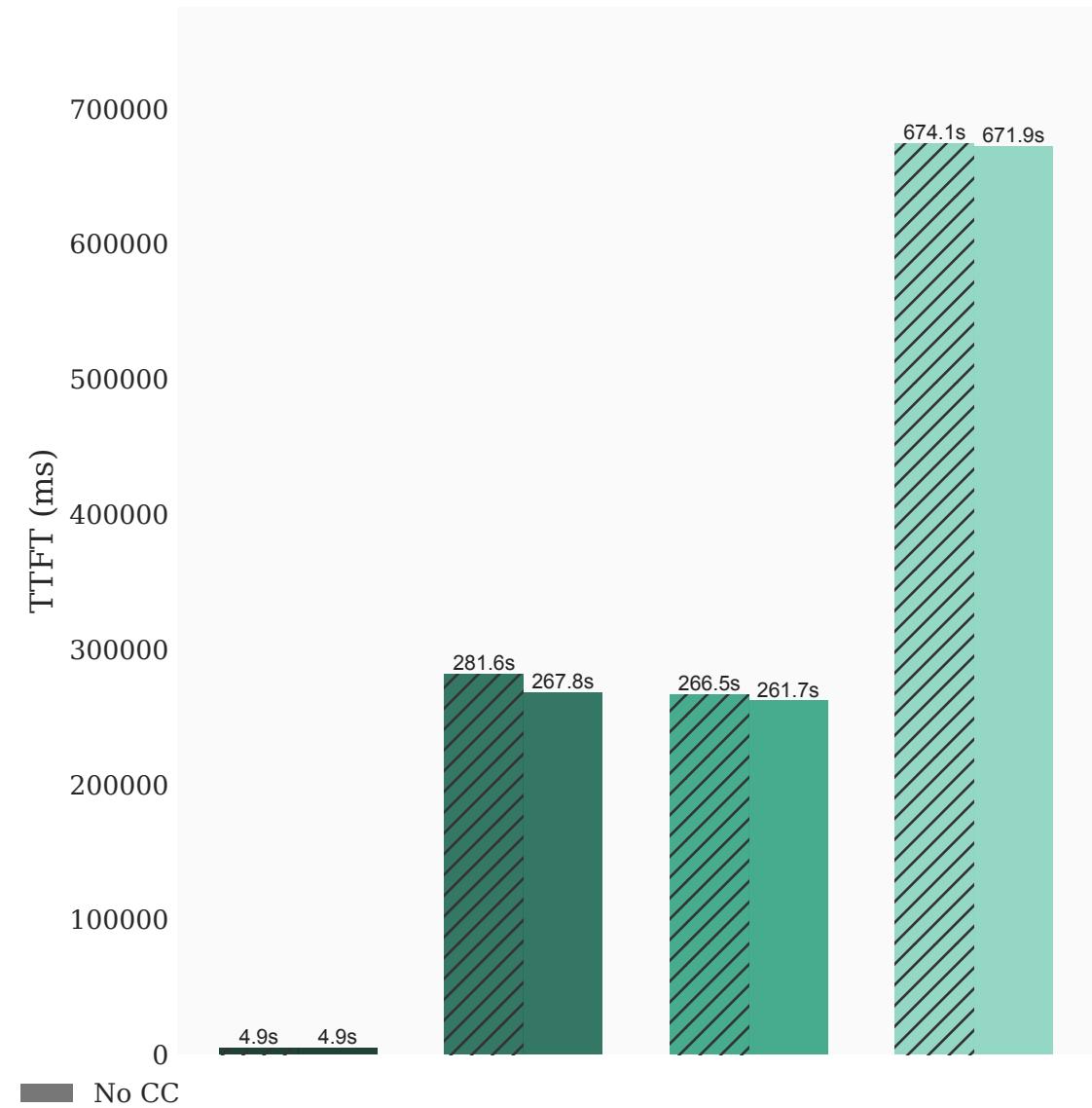
■ LLama 3.3 70B Int4

Edit 10K Characters (100 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

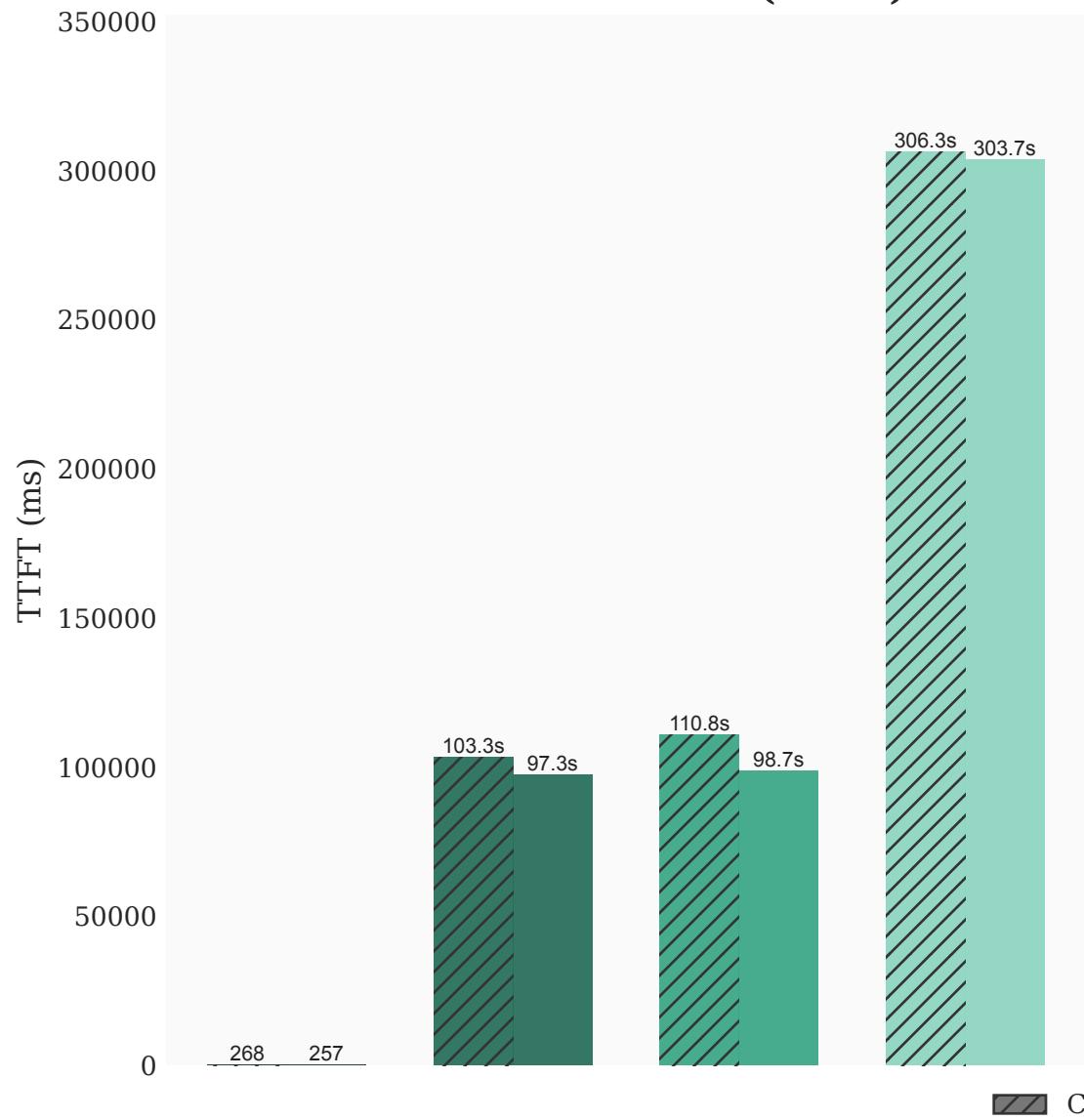
■ Mistral 3.1 24B

■ GPT OSS 120B

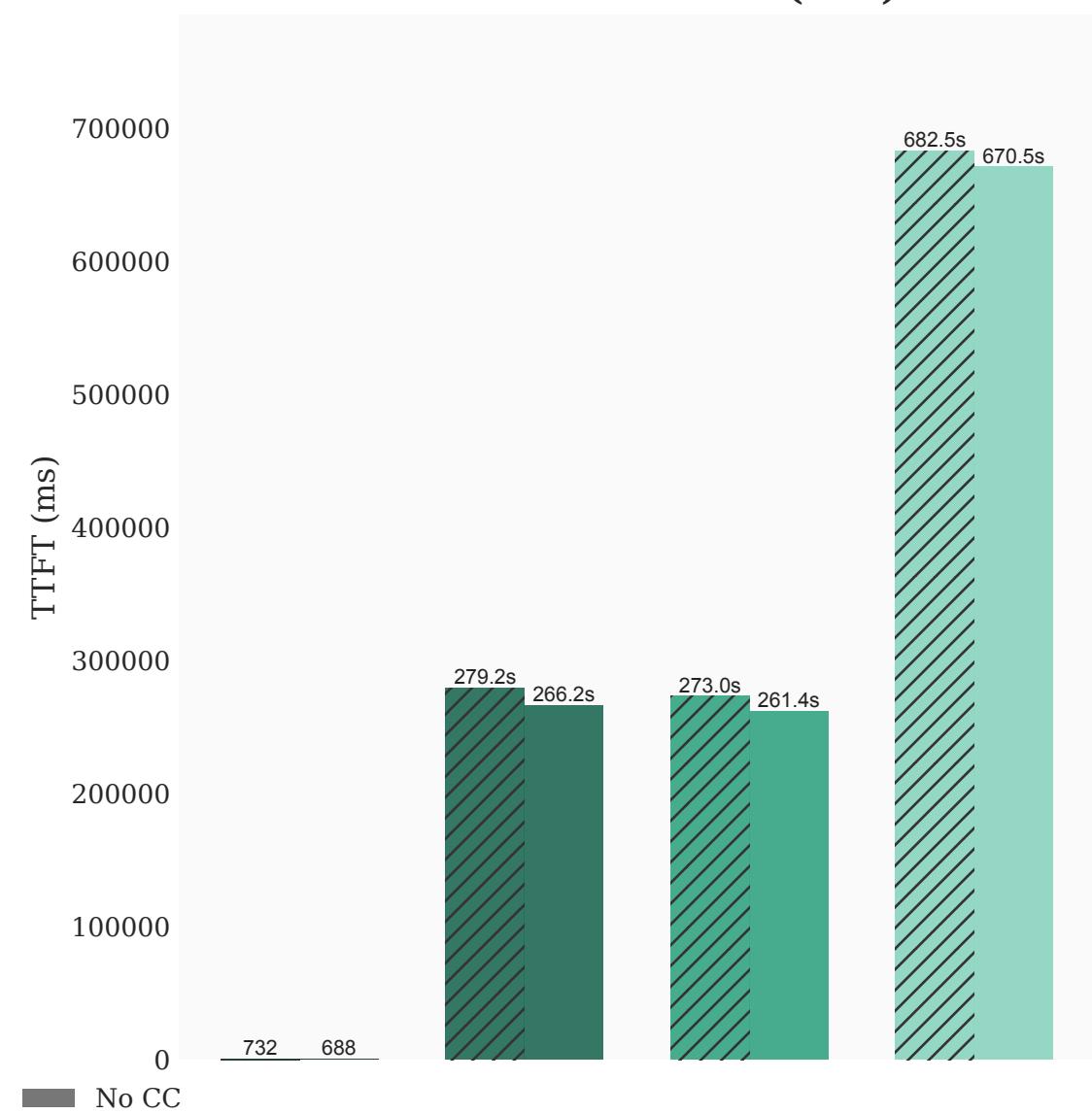
■ LLama 3.3 70B Int4

Edit 10K Characters (50 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

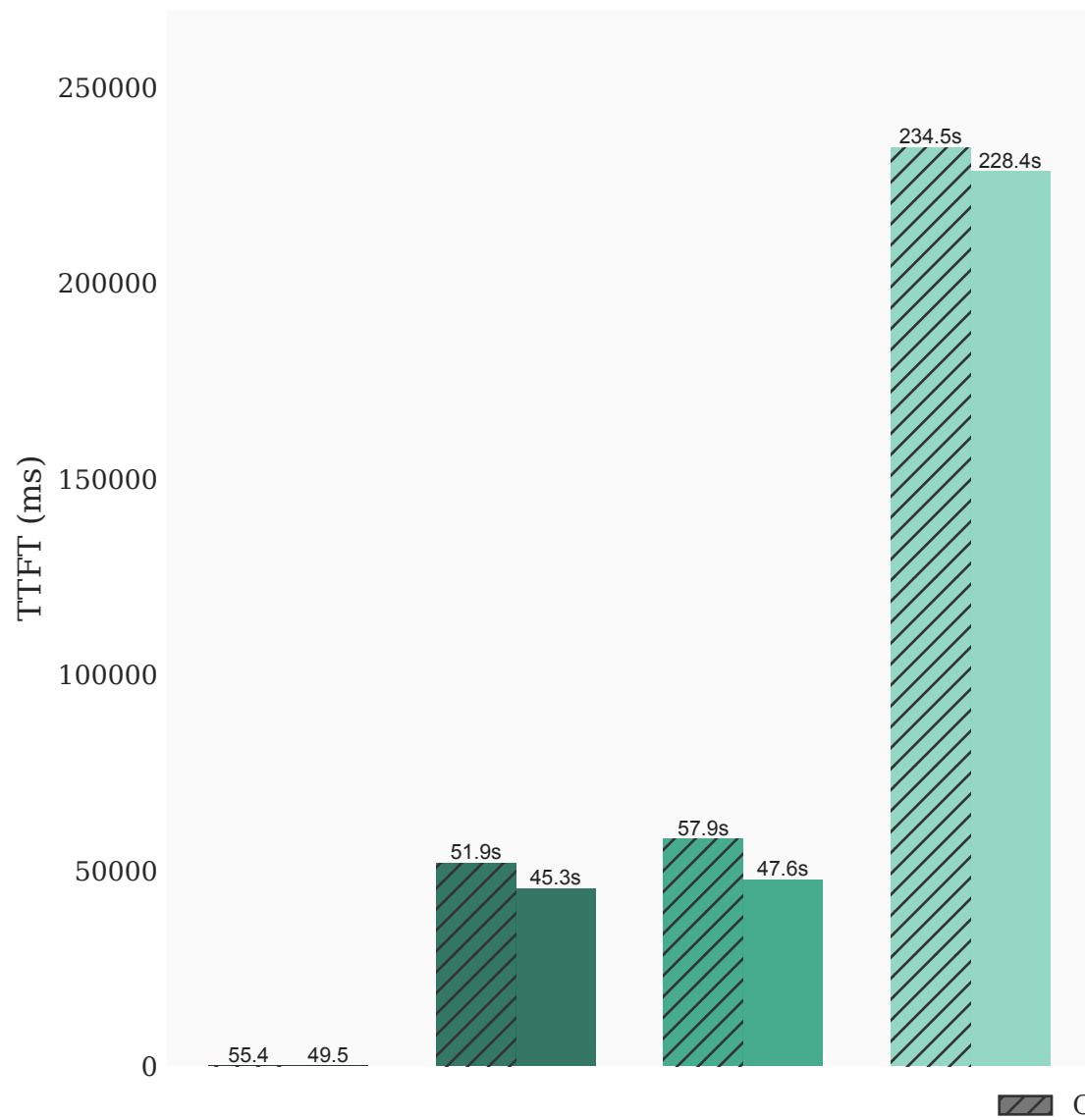
■ Mistral 3.1 24B

■ GPT OSS 120B

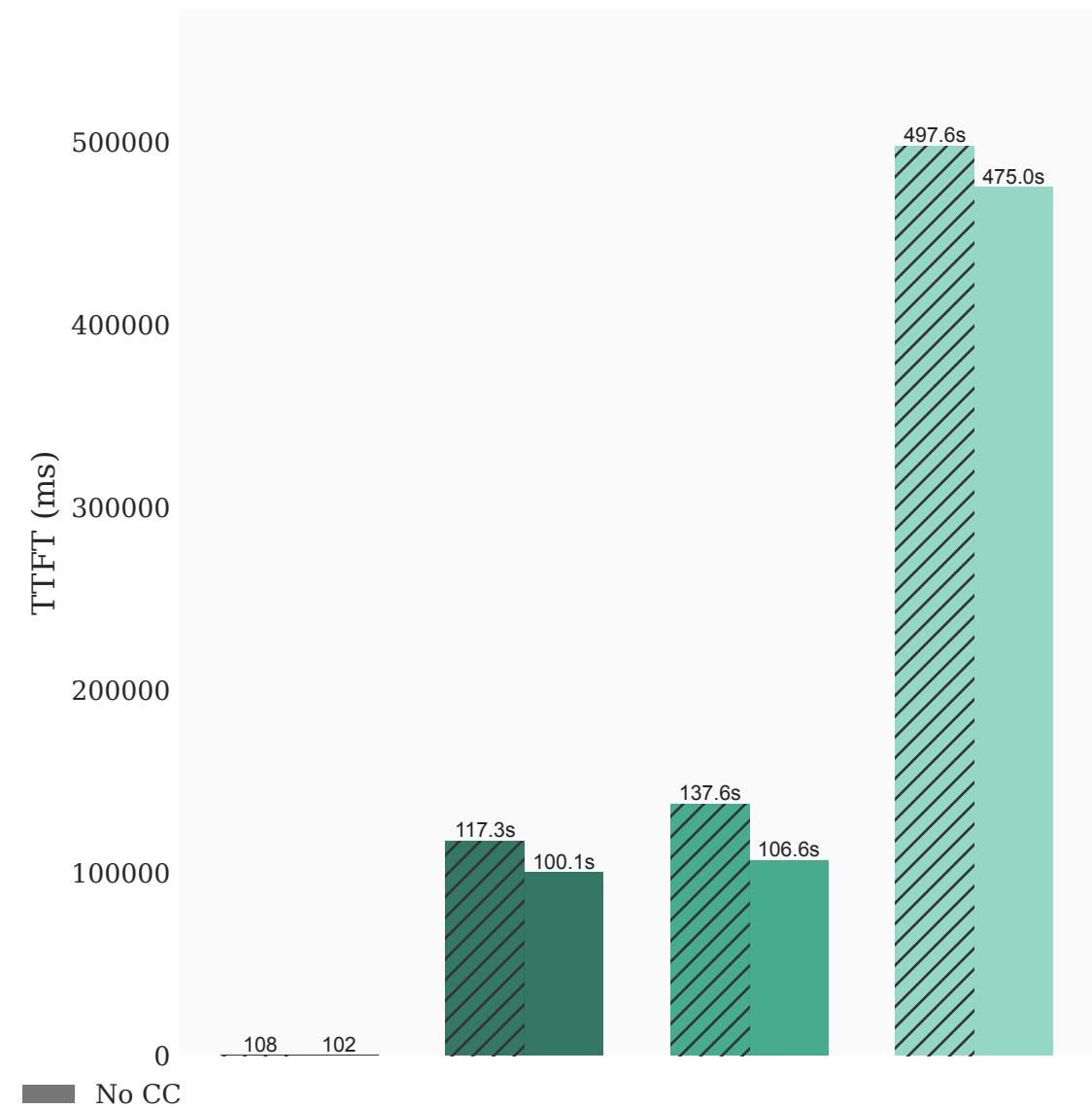
■ LLama 3.3 70B Int4

Edit 10K Characters (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



LLama 3.1 8B

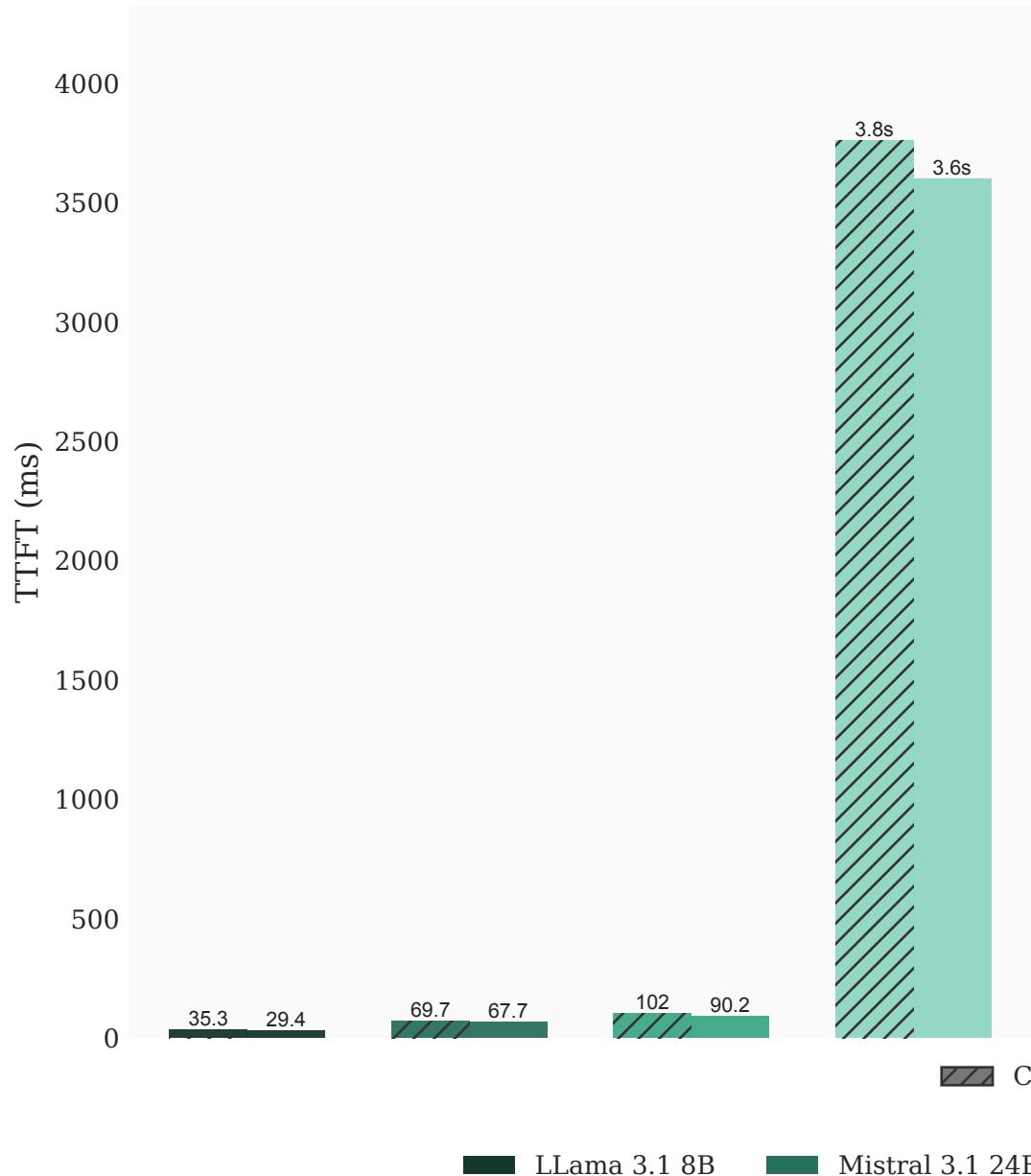
Mistral 3.1 24B

GPT OSS 120B

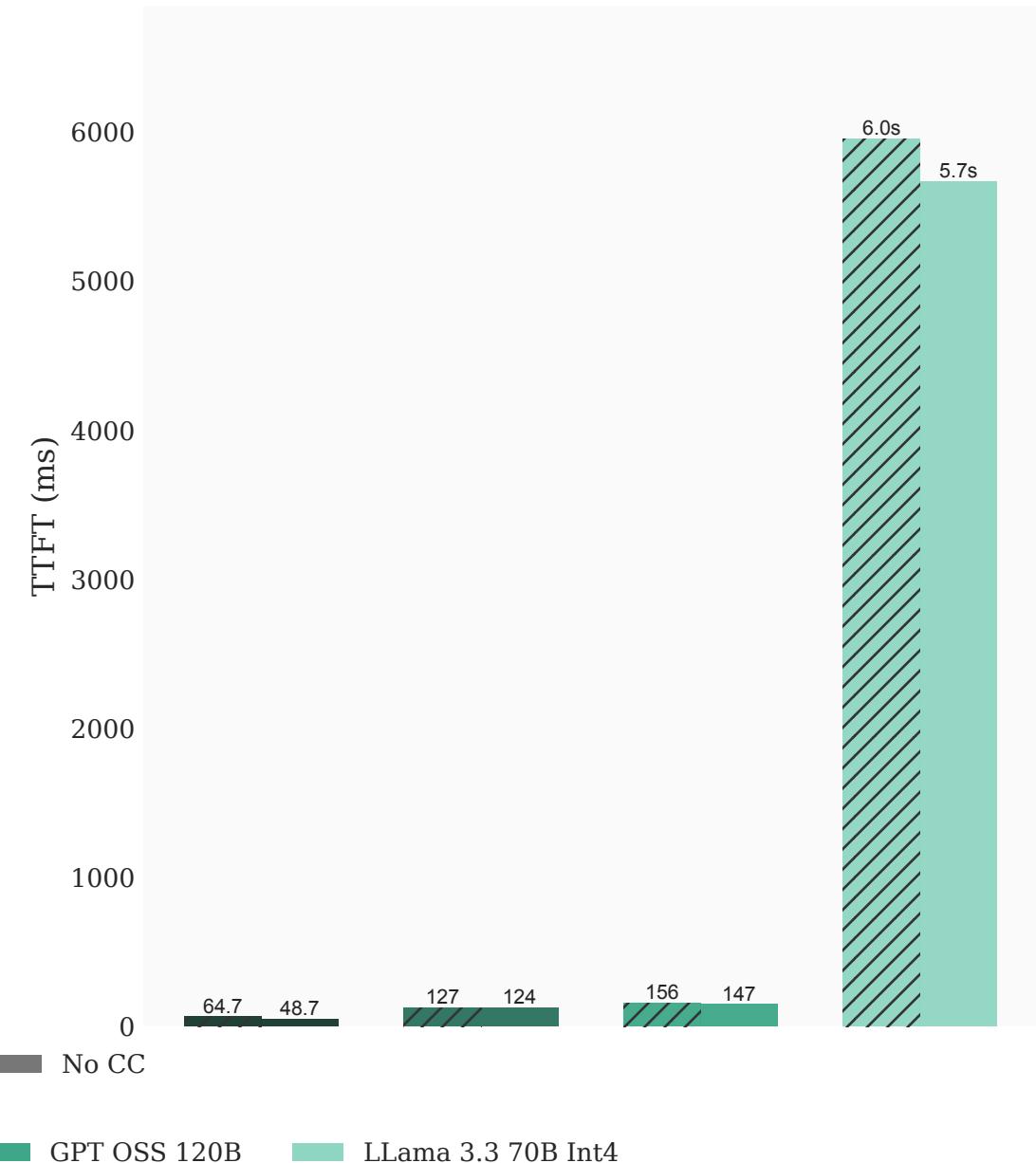
LLama 3.3 70B Int4

Numina Math (100 Request Rate)

Time to First Token (Mean)

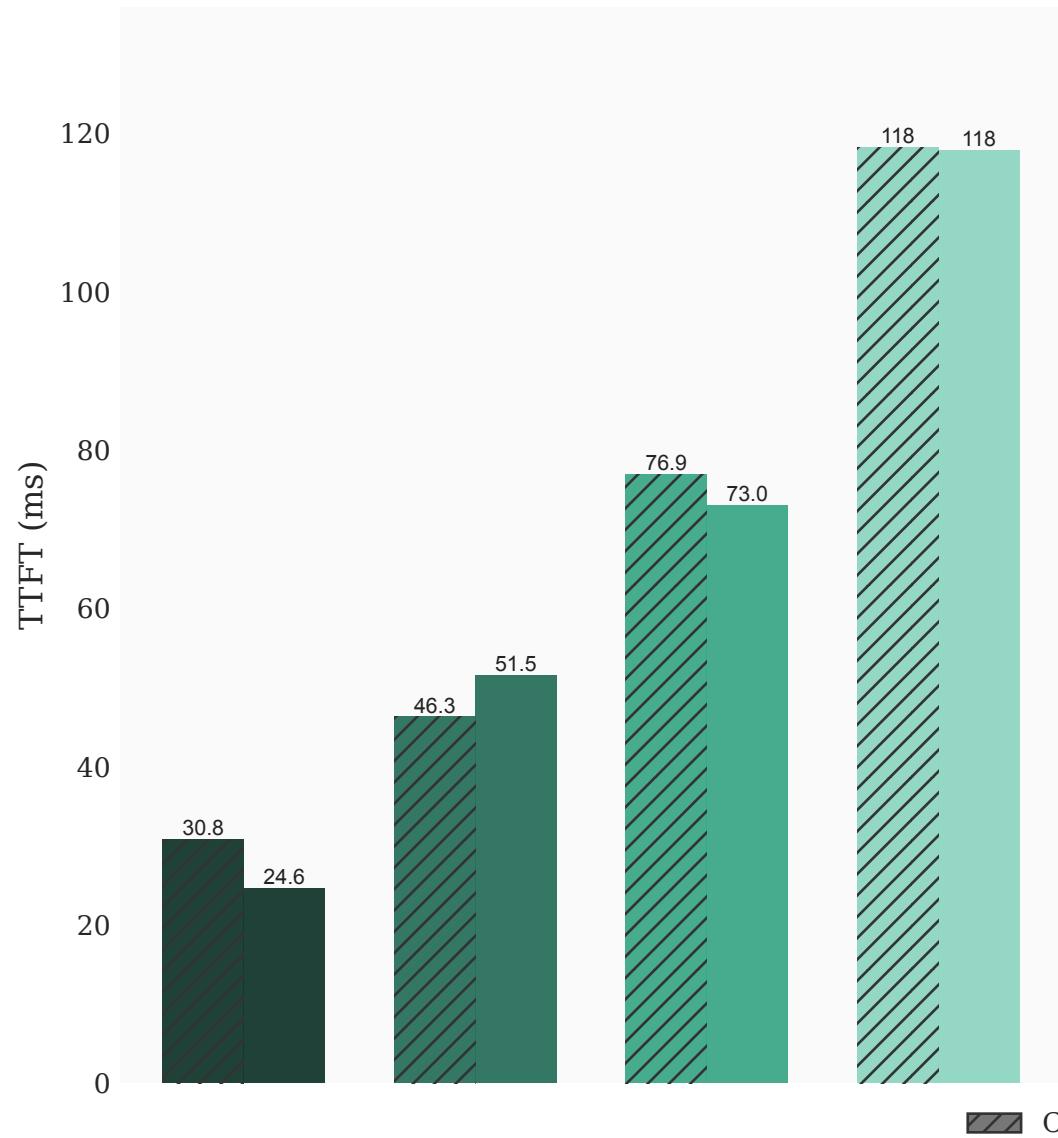


Time to First Token (P99)

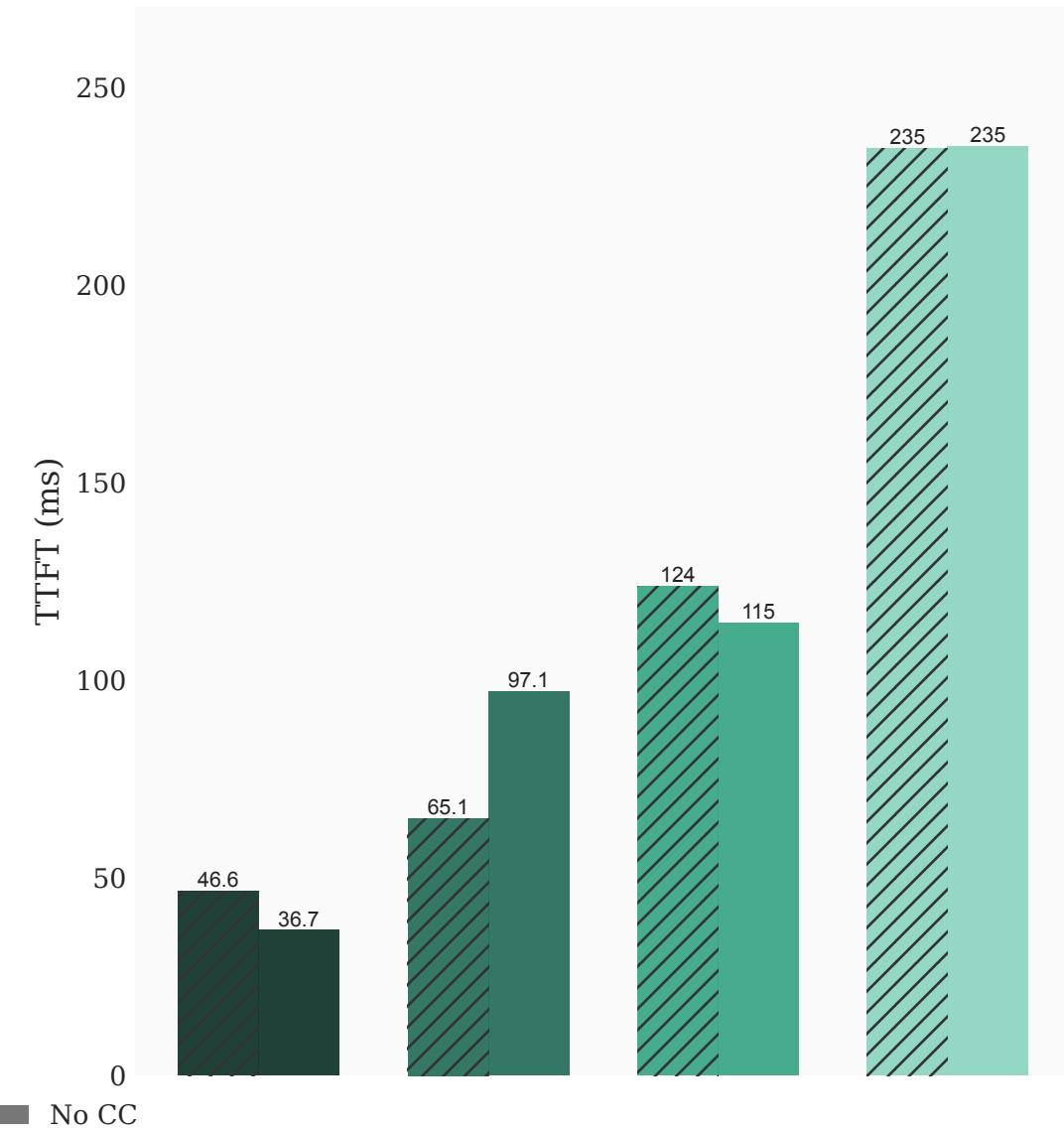


Numina Math (50 Request Rate)

Time to First Token (Mean)



Time to First Token (P99)



LLama 3.1 8B

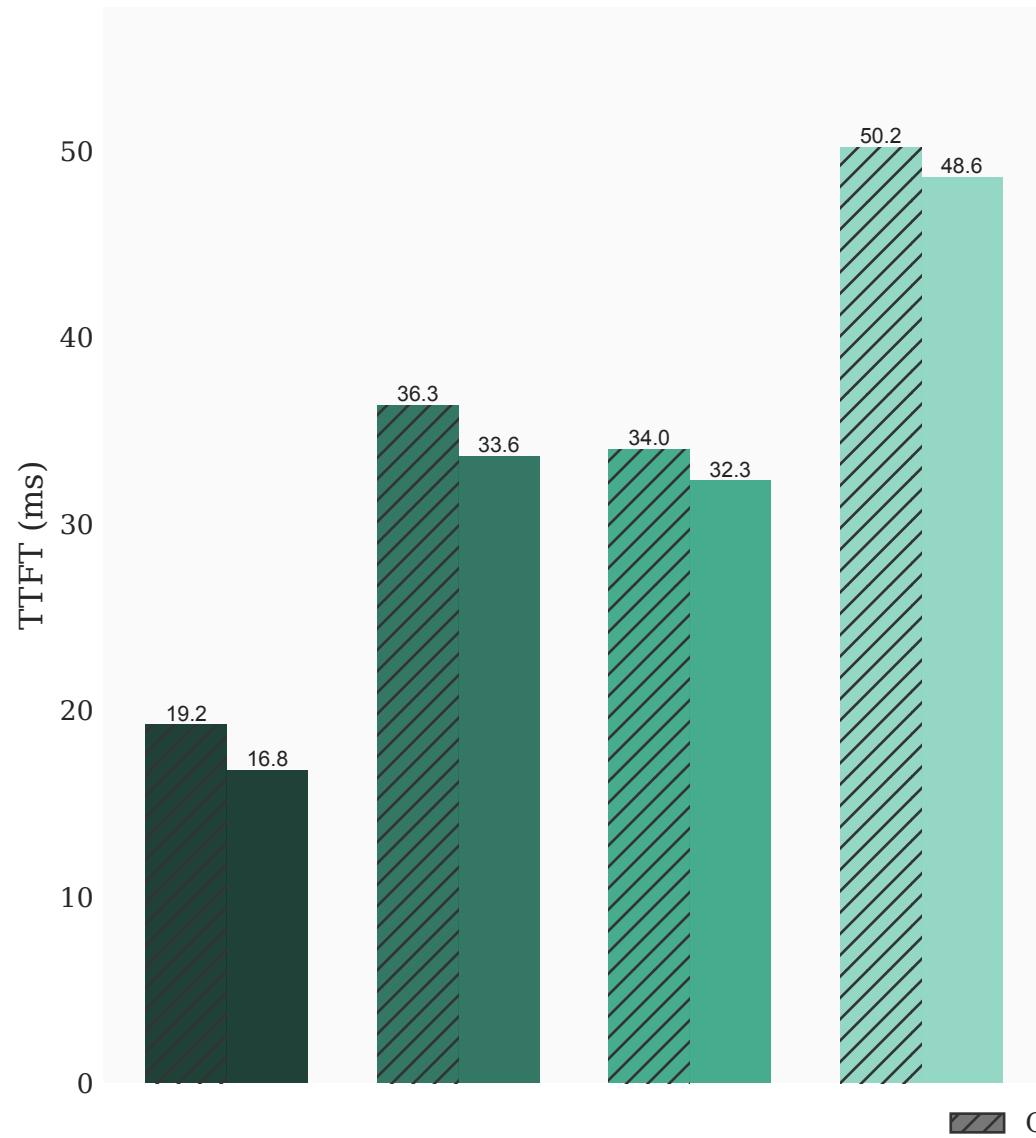
Mistral 3.1 24B

GPT OSS 120B

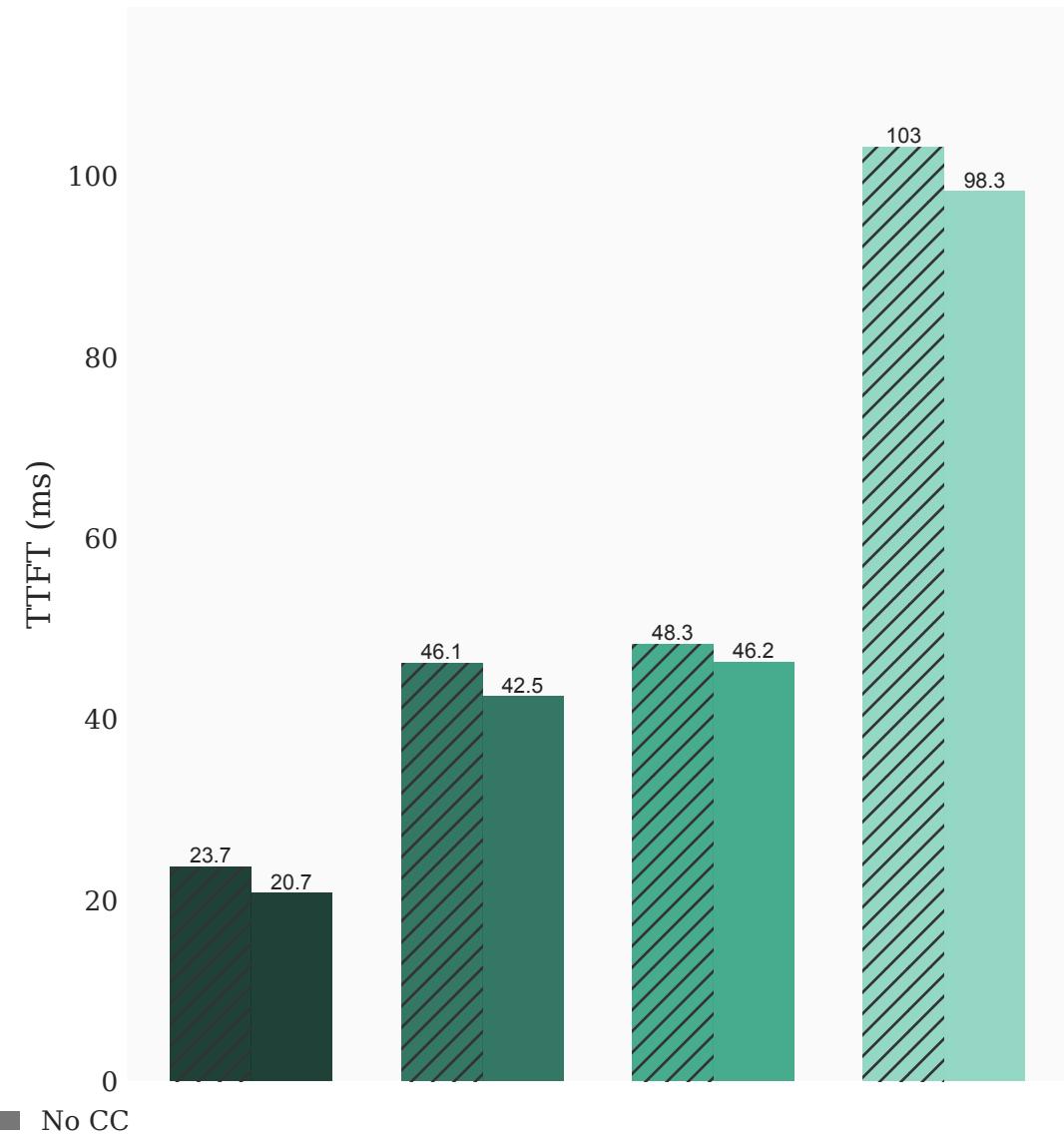
LLama 3.3 70B Int4

Numina Math (Single Request)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.1 8B

■ Mistral 3.1 24B

■ GPT OSS 120B

■ LLama 3.3 70B Int4