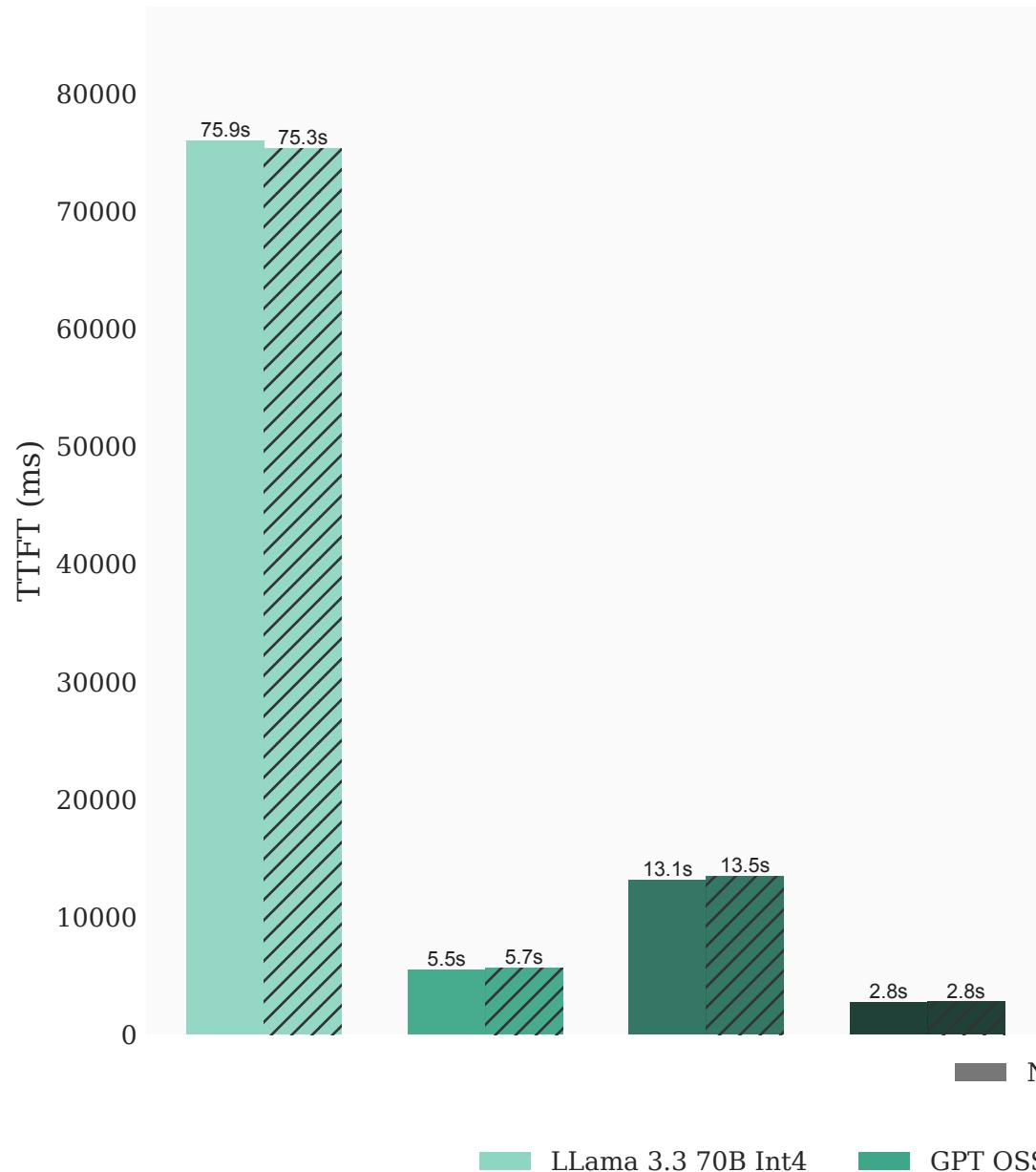
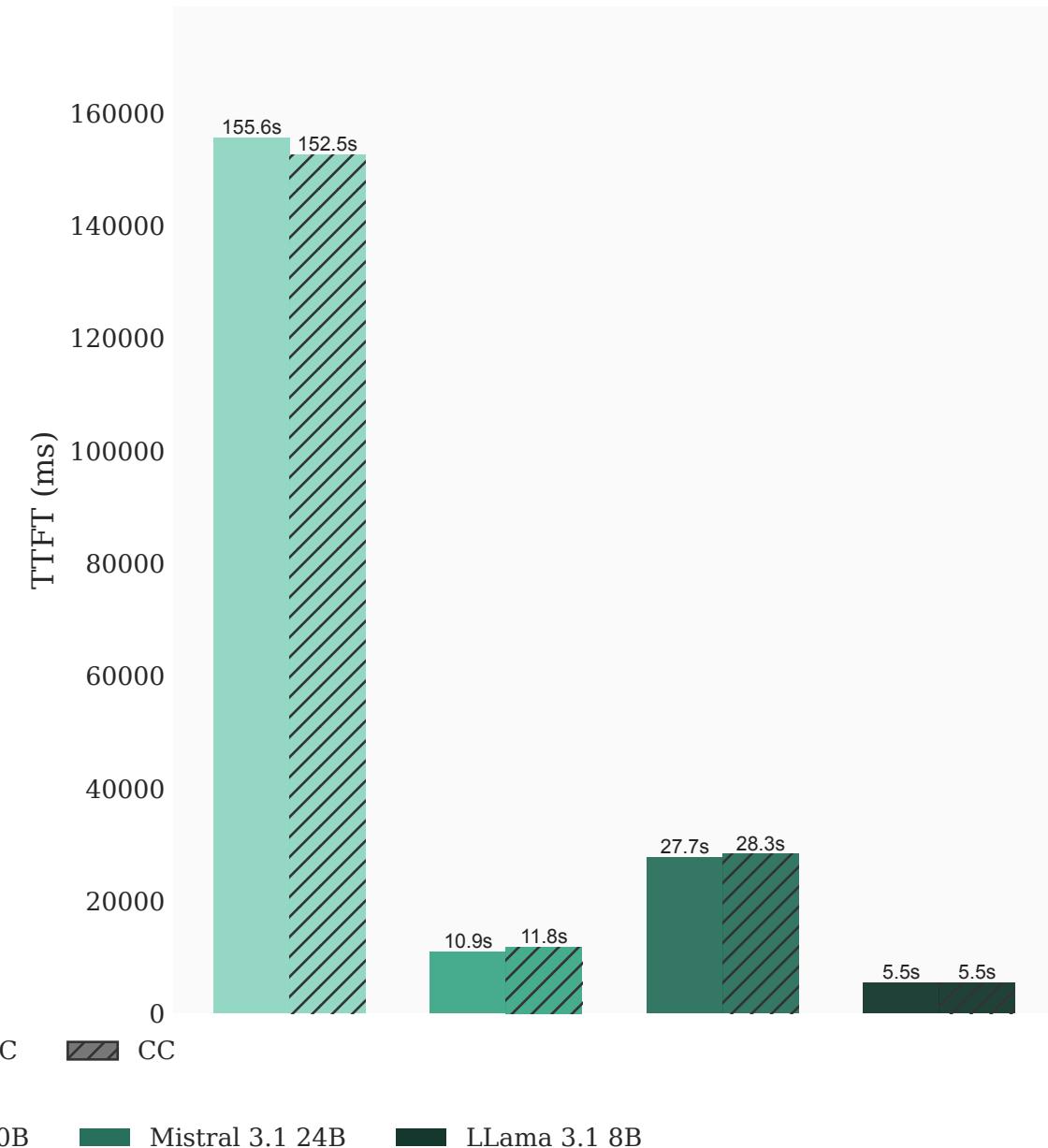


Random (1500 \Rightarrow 250) (Request Rate 100)

Time to First Token (Mean)

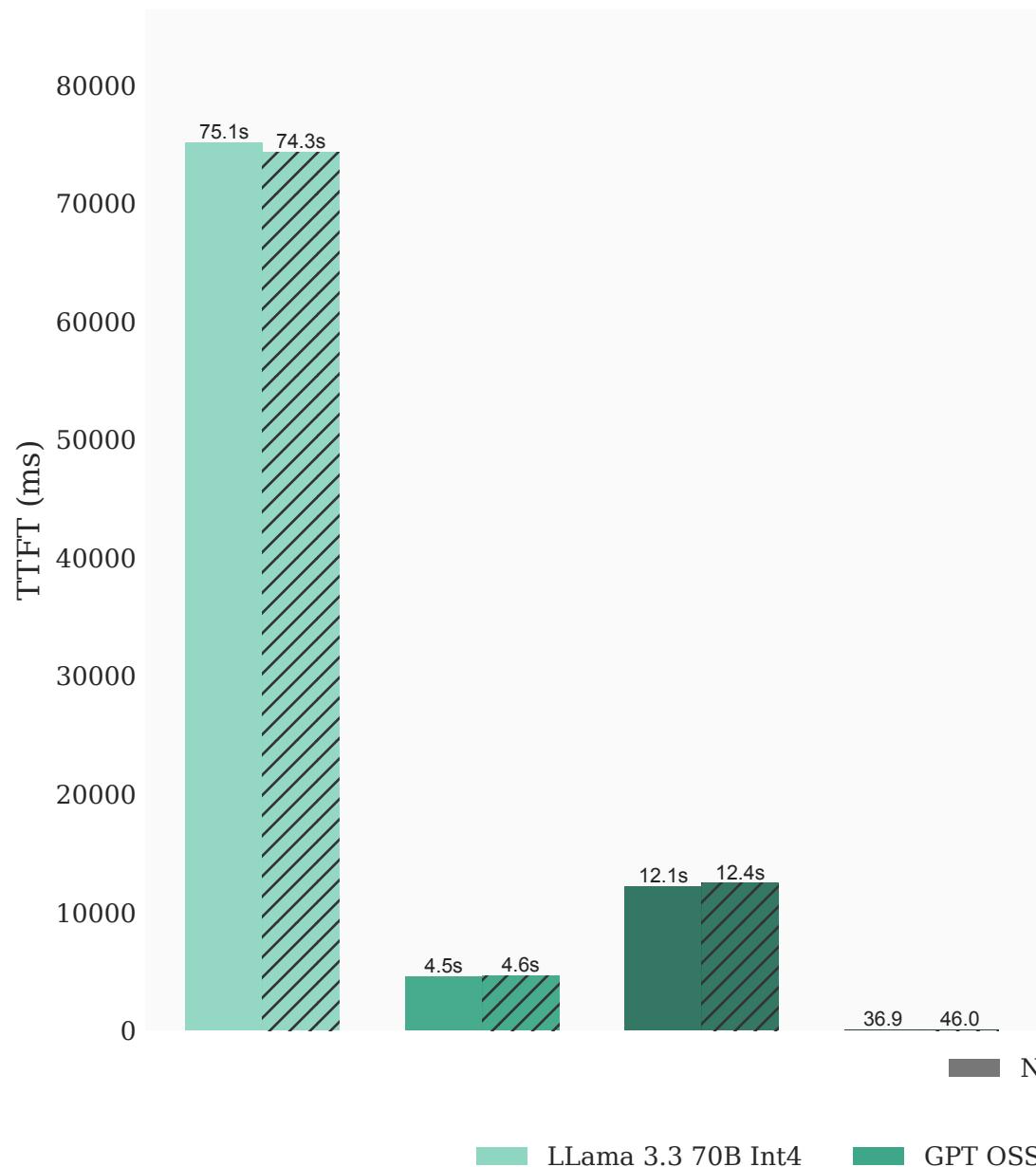


Time to First Token (P99)

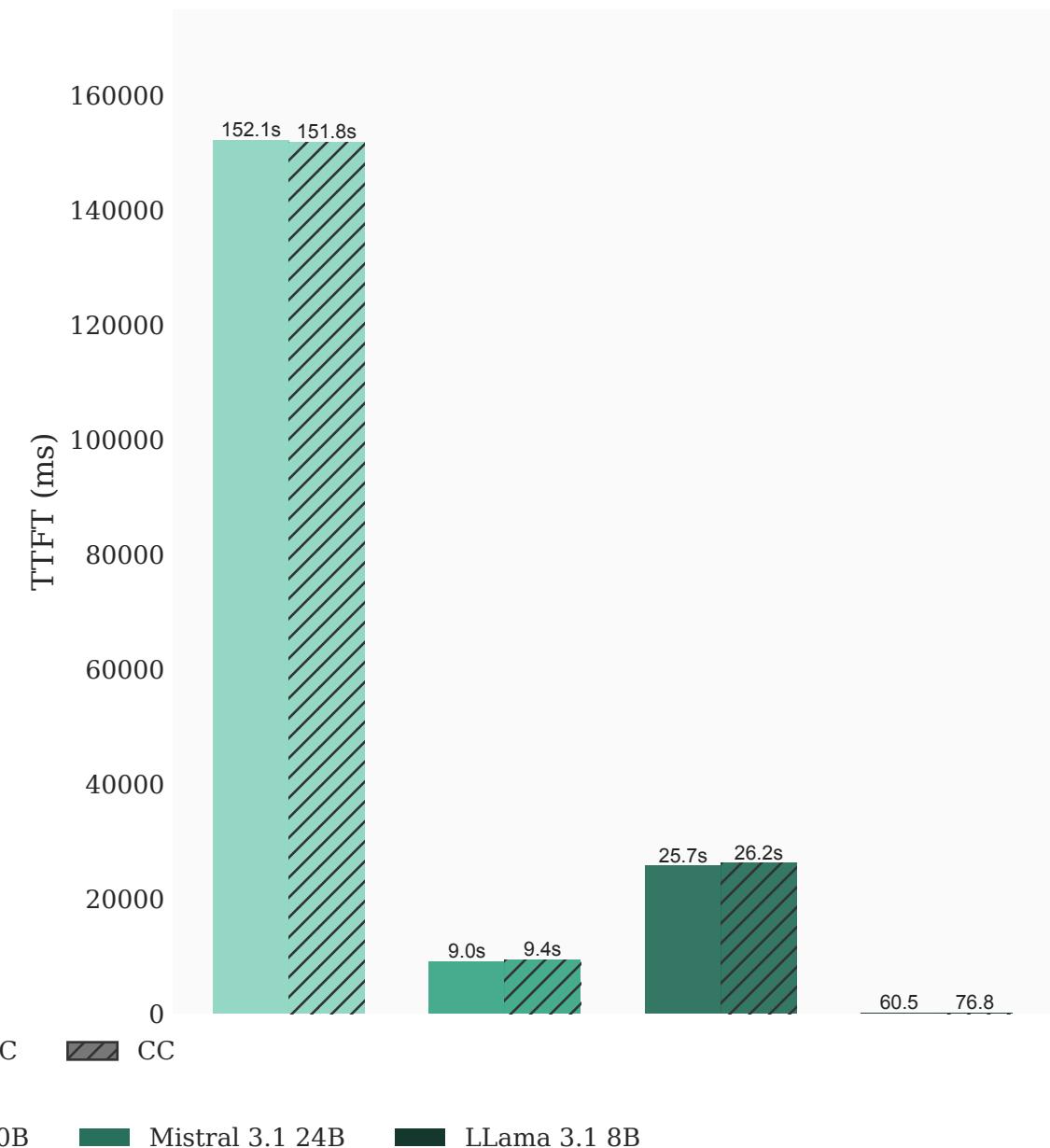


Random (1500 \Rightarrow 250) (Request Rate 50)

Time to First Token (Mean)

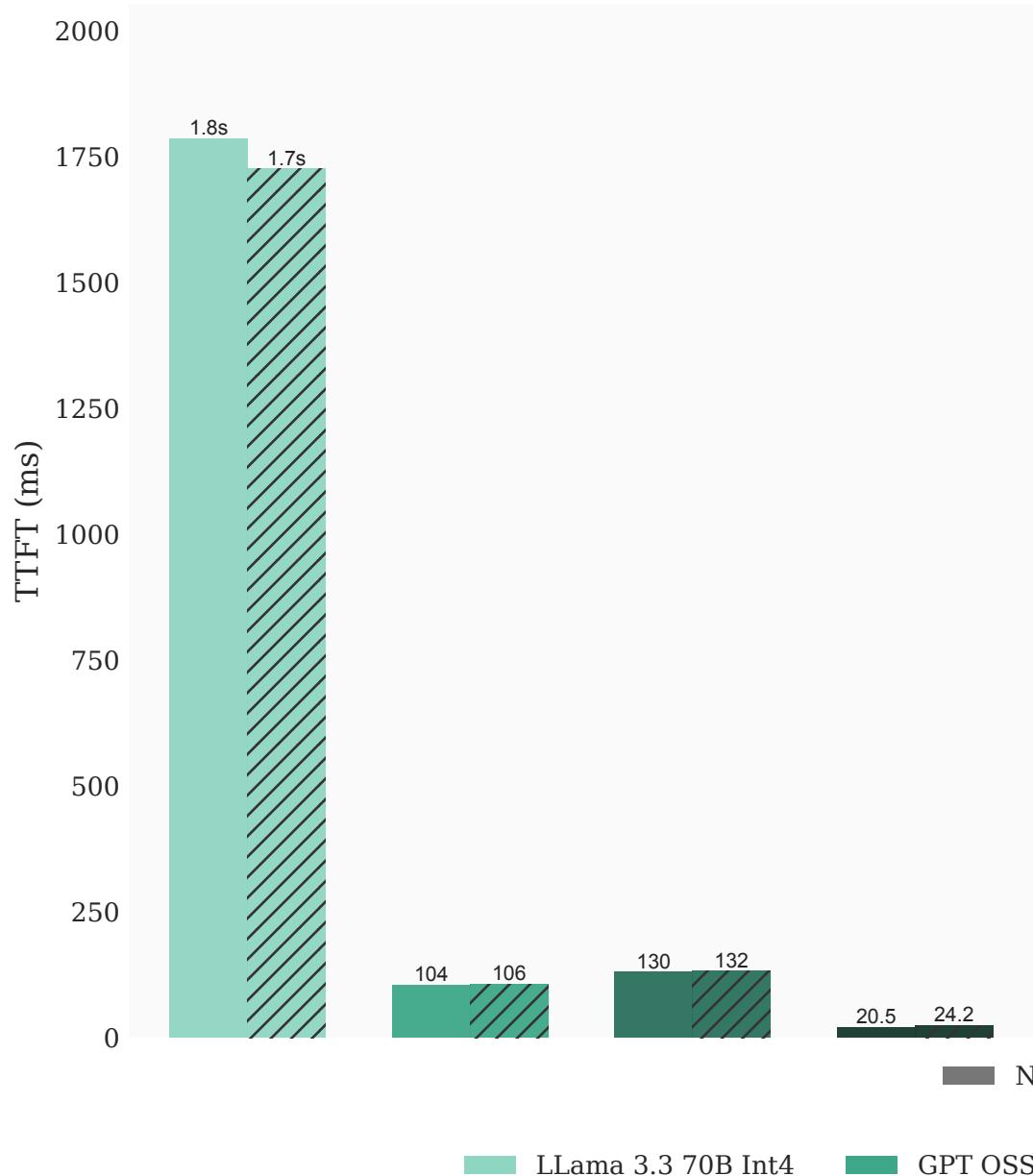


Time to First Token (P99)

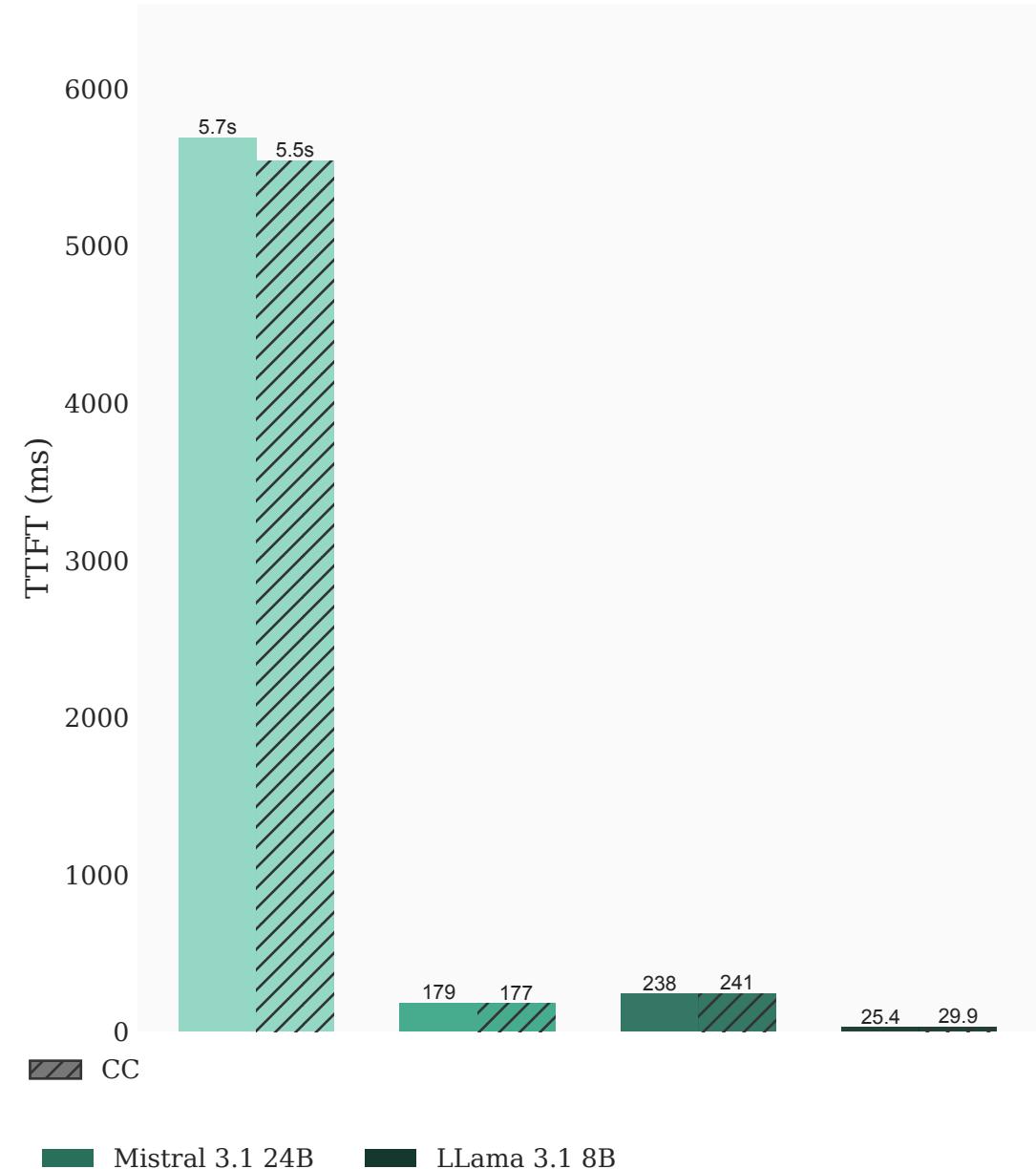


Random (1500 ⇒ 250) (Request Rate 1)

Time to First Token (Mean)

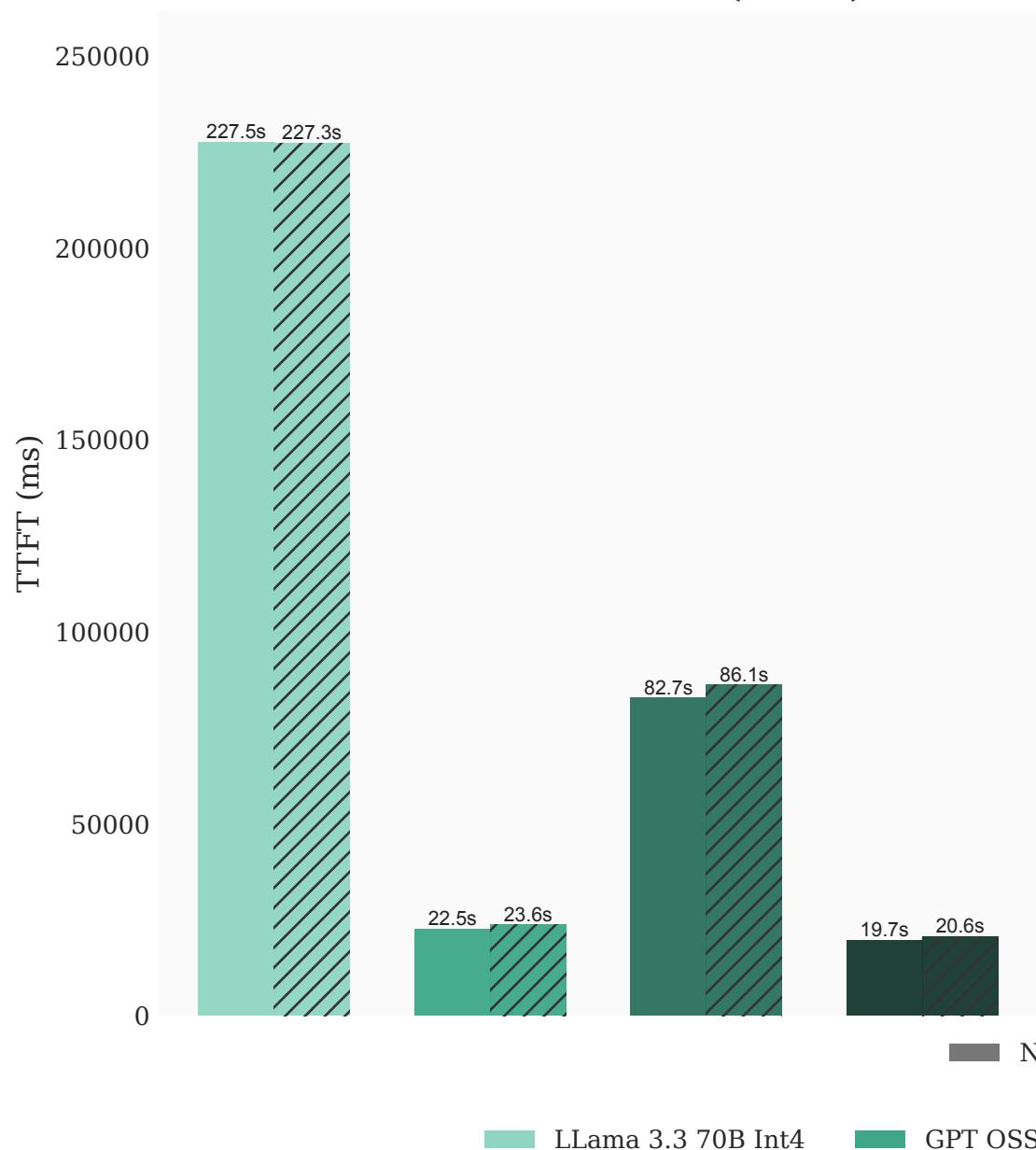


Time to First Token (P99)

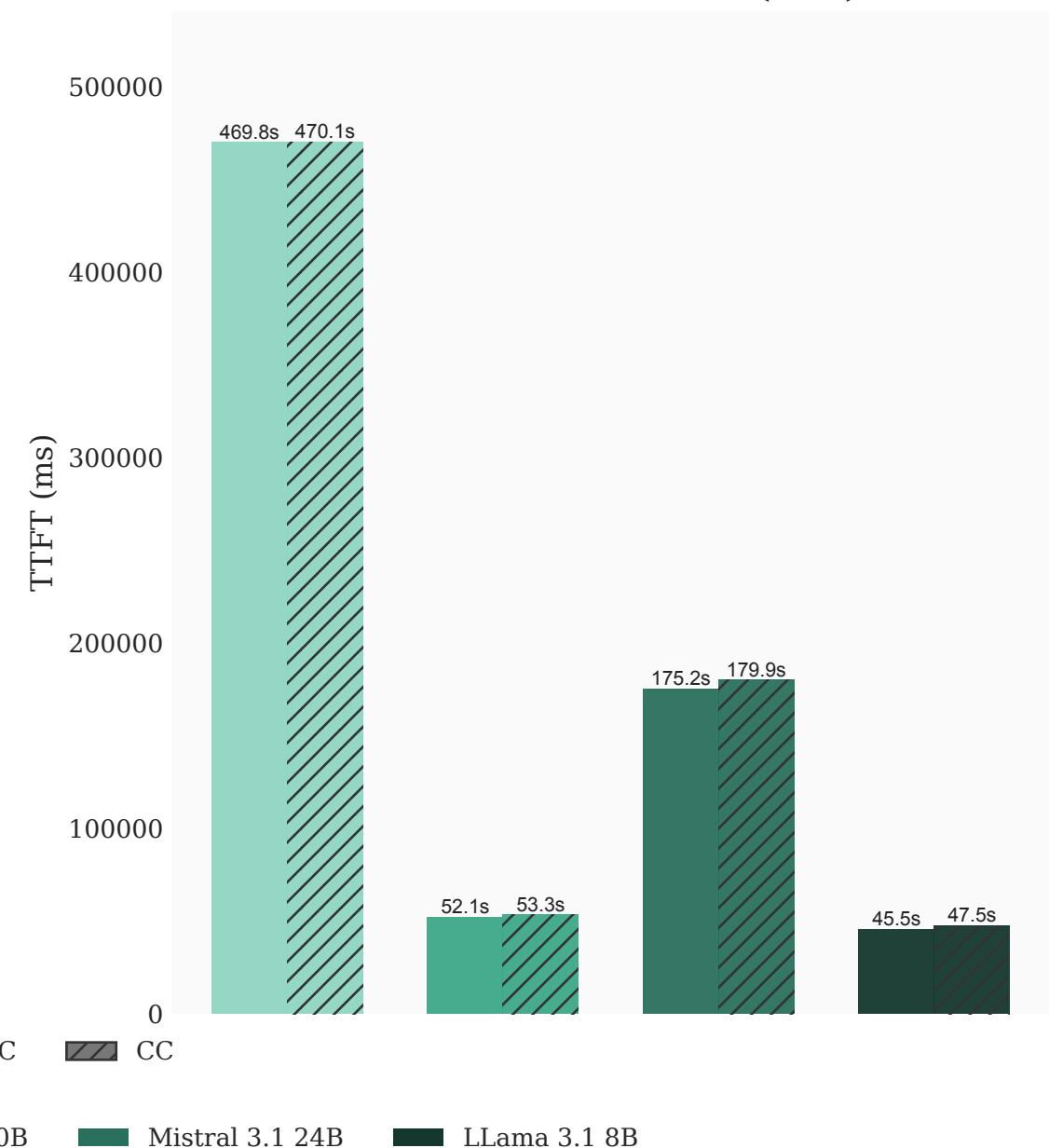


Random (4000 \Rightarrow 1000) (Request Rate 100)

Time to First Token (Mean)

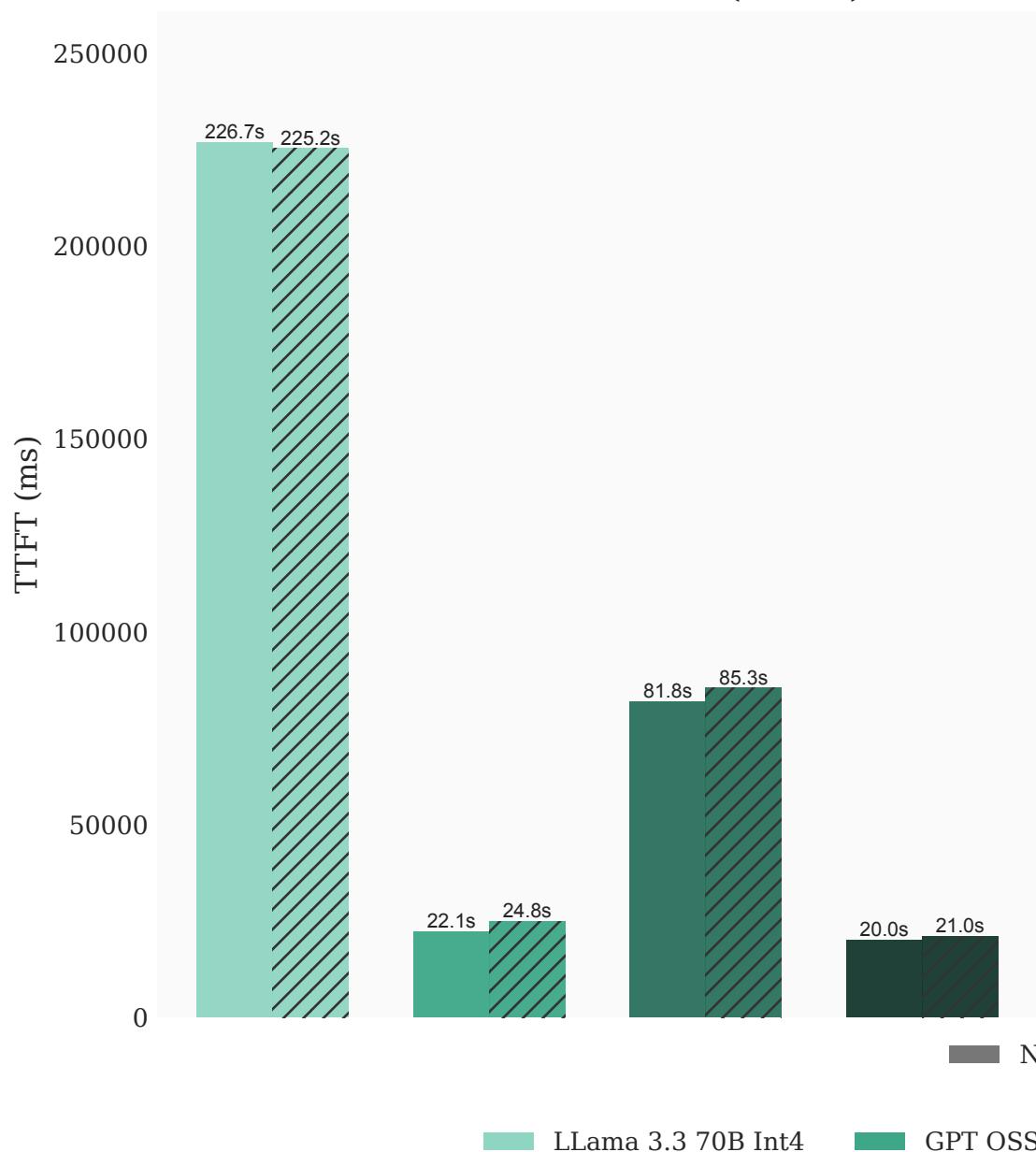


Time to First Token (P99)

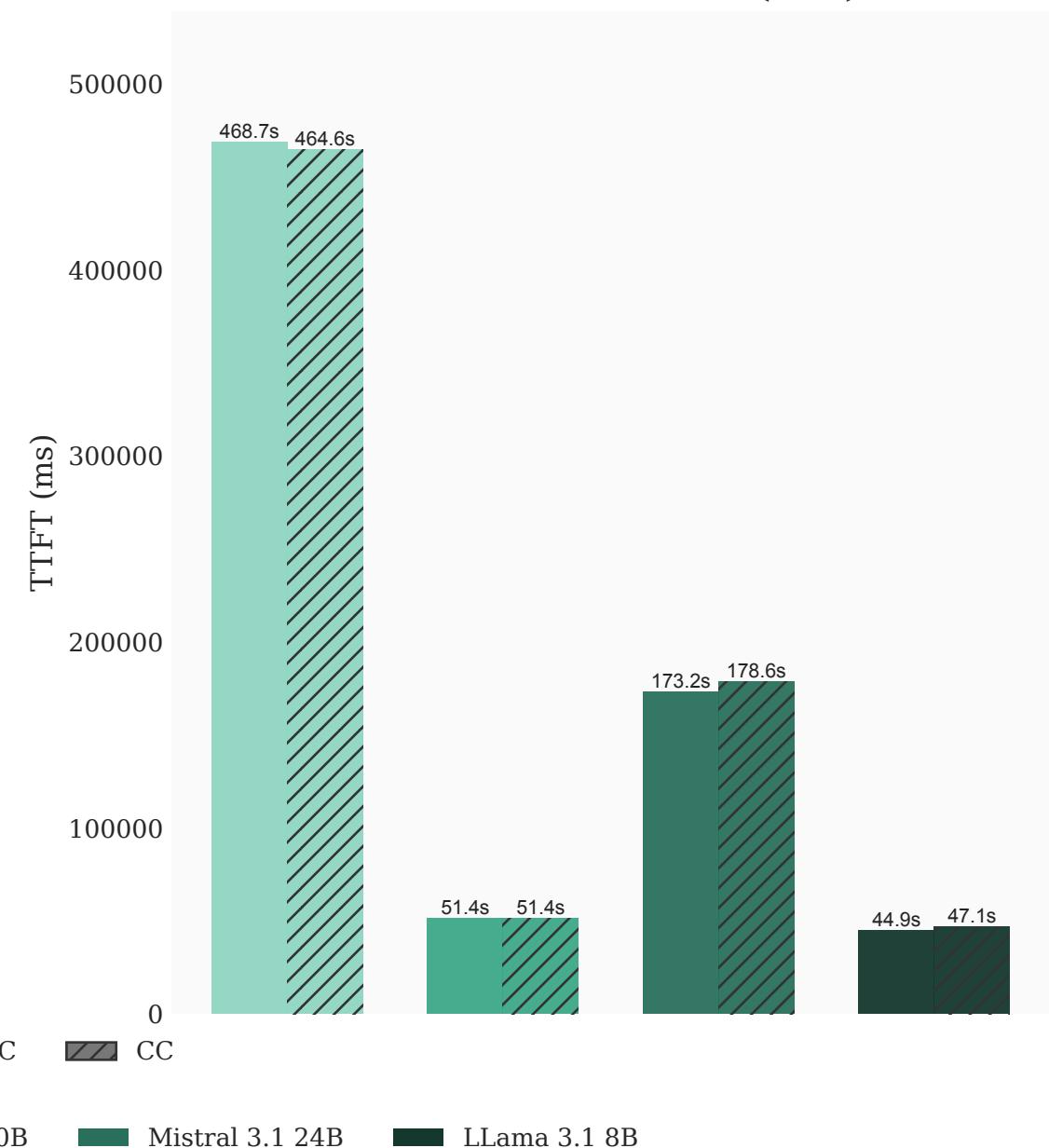


Random (4000 ⇒ 1000) (Request Rate 50)

Time to First Token (Mean)

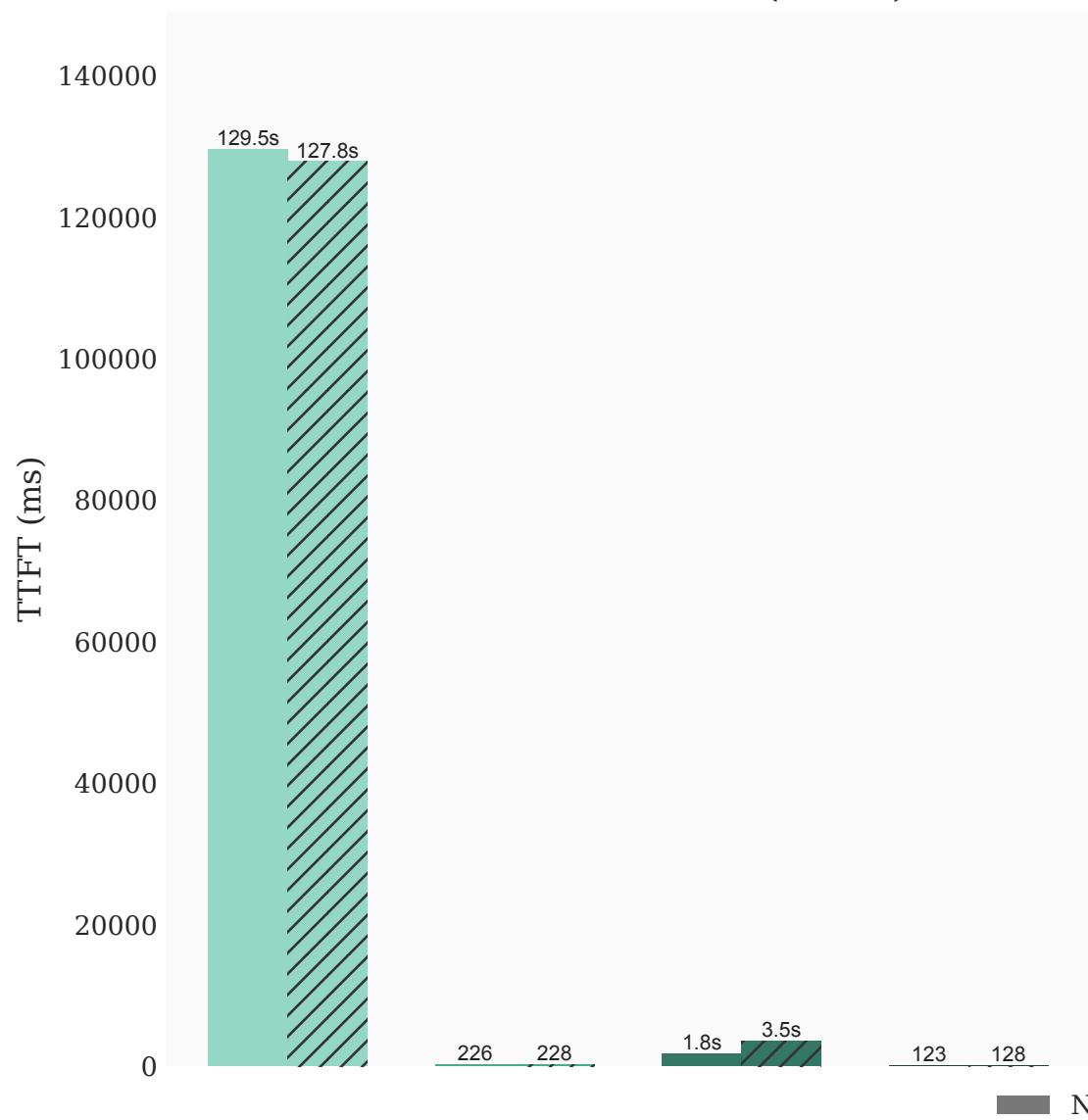


Time to First Token (P99)

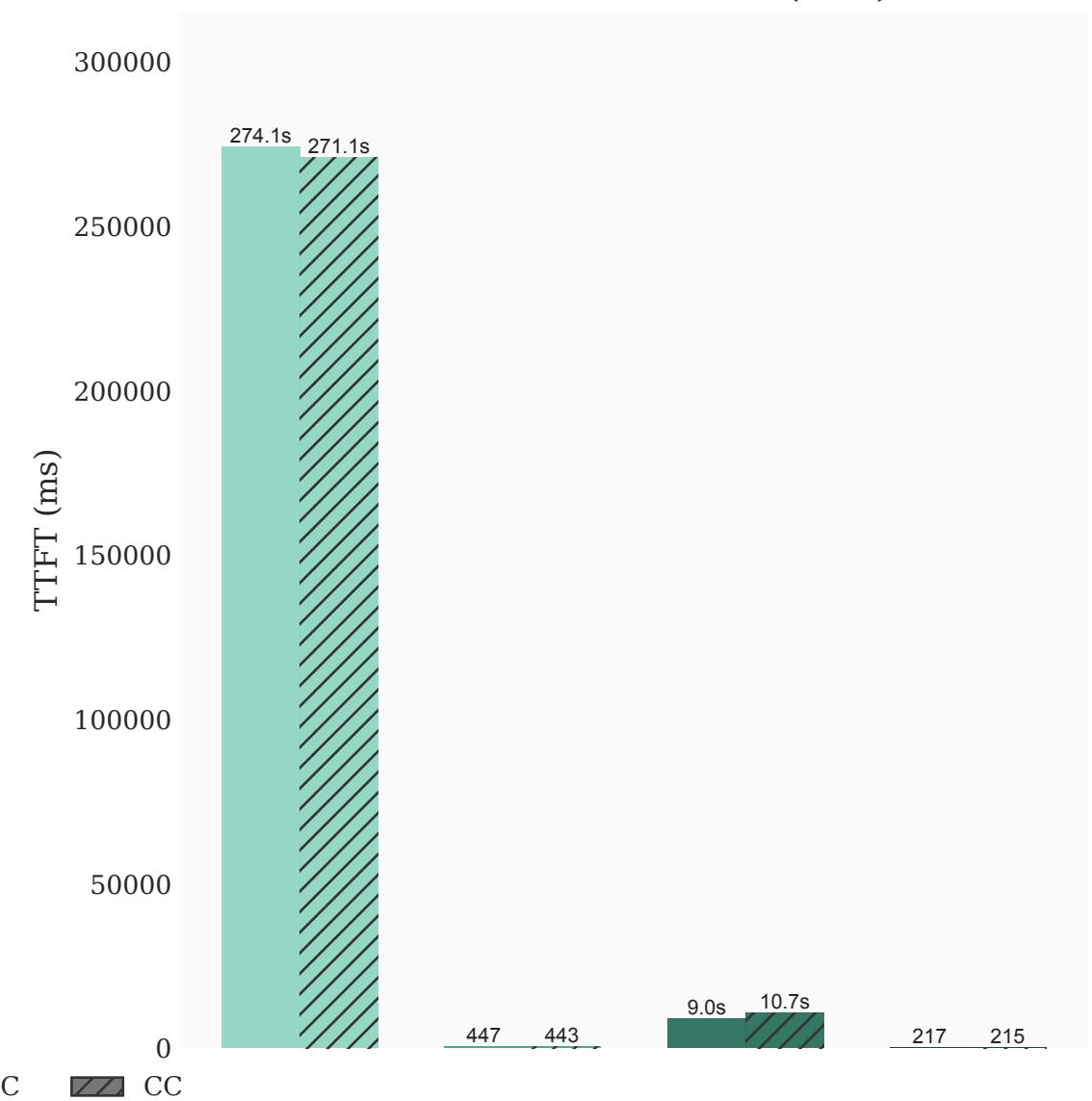


Random (4000 \Rightarrow 1000) (Request Rate 1)

Time to First Token (Mean)



Time to First Token (P99)



■ LLama 3.3 70B Int4

■ GPT OSS 120B

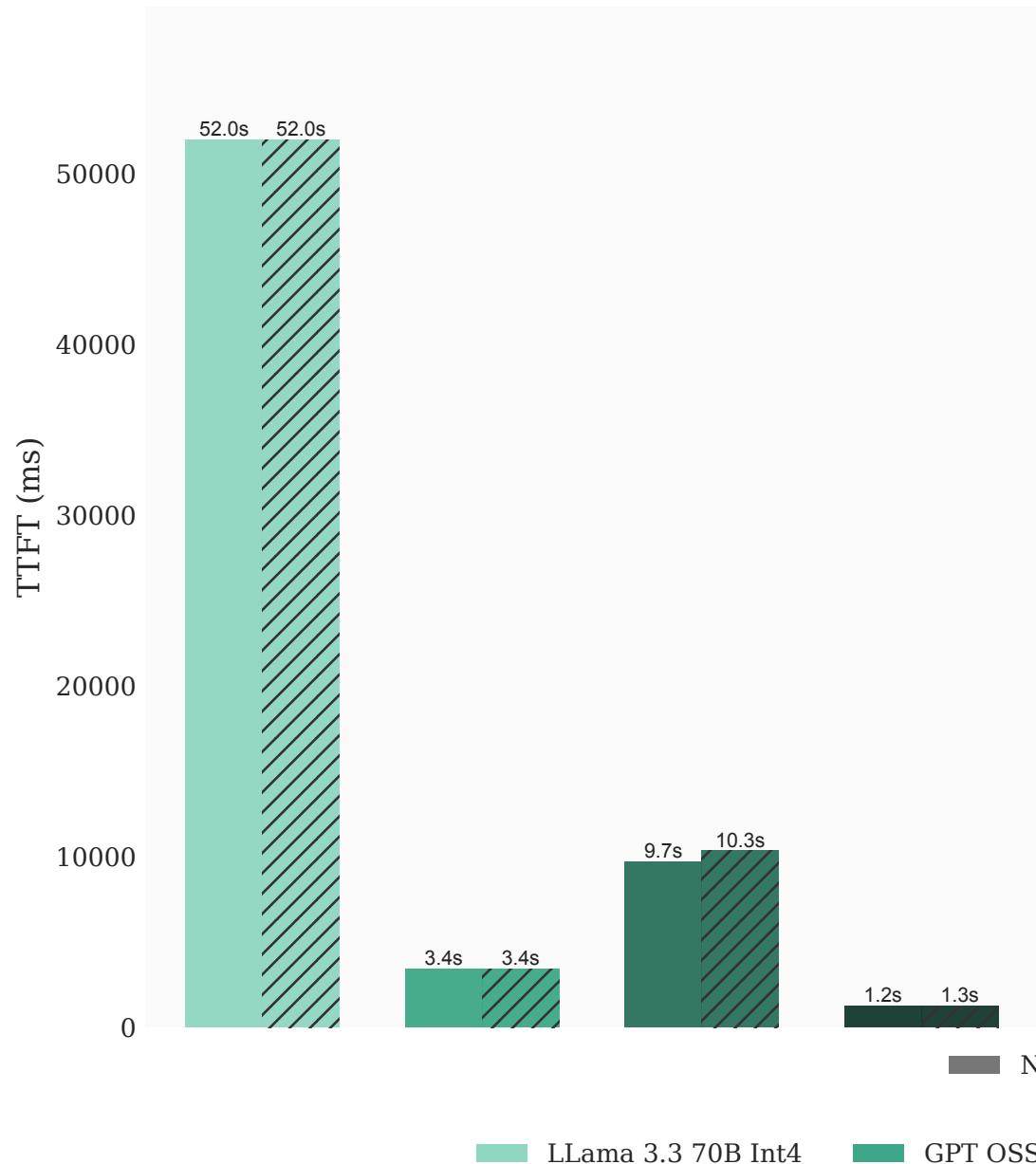
■ Mistral 3.1 24B

■ LLama 3.1 8B

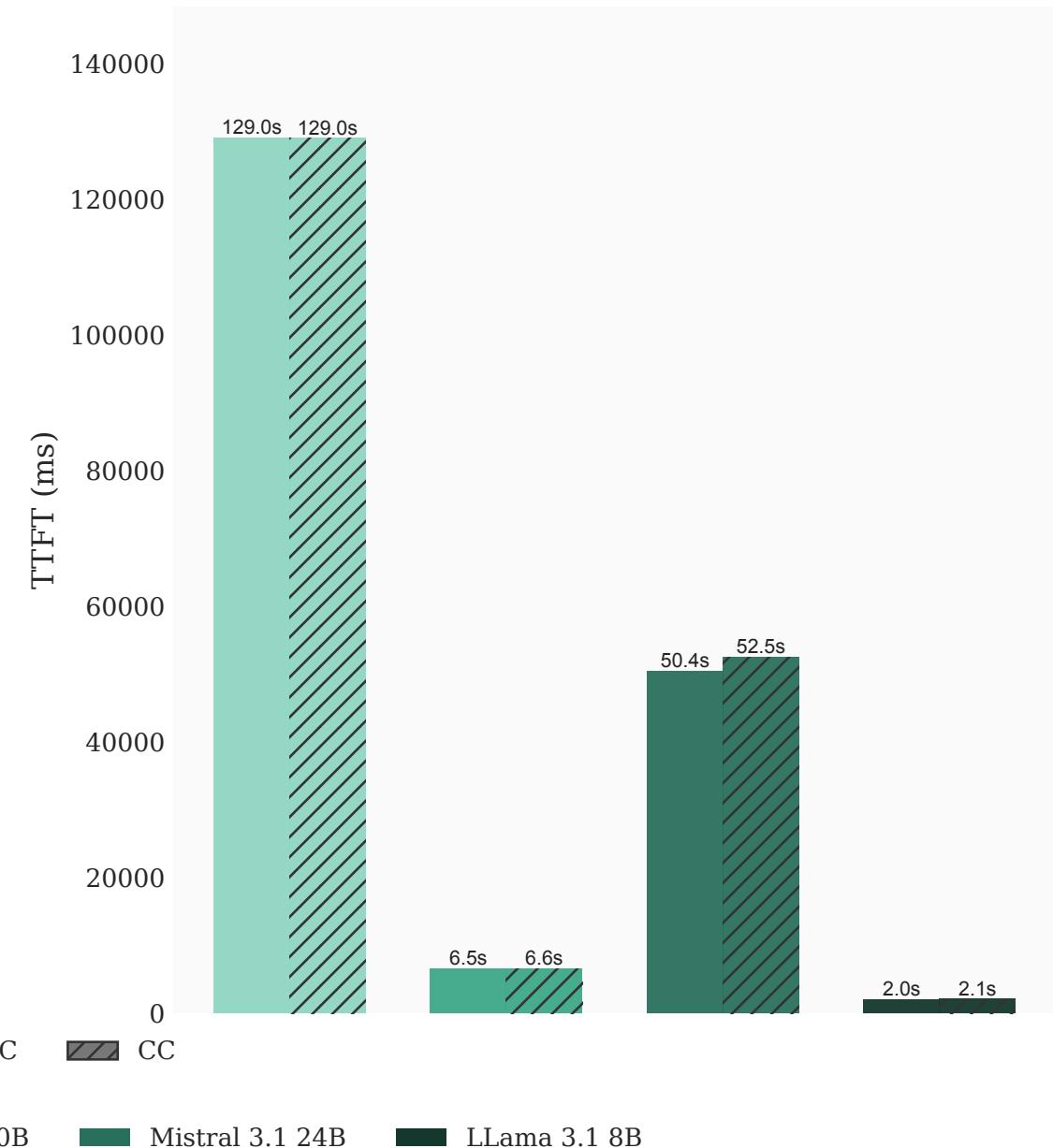
■ No CC ■ CC

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

Time to First Token (Mean)

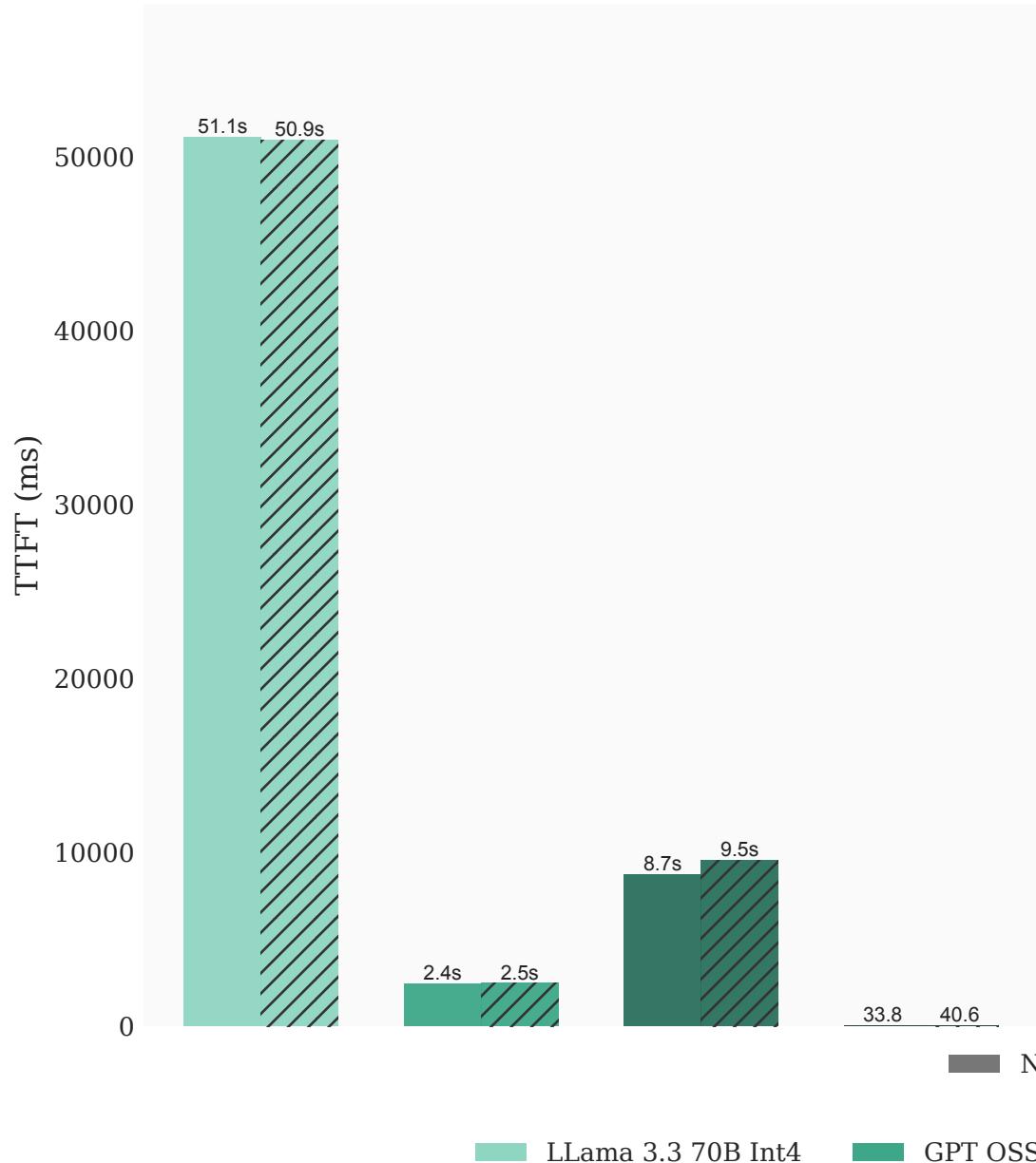


Time to First Token (P99)

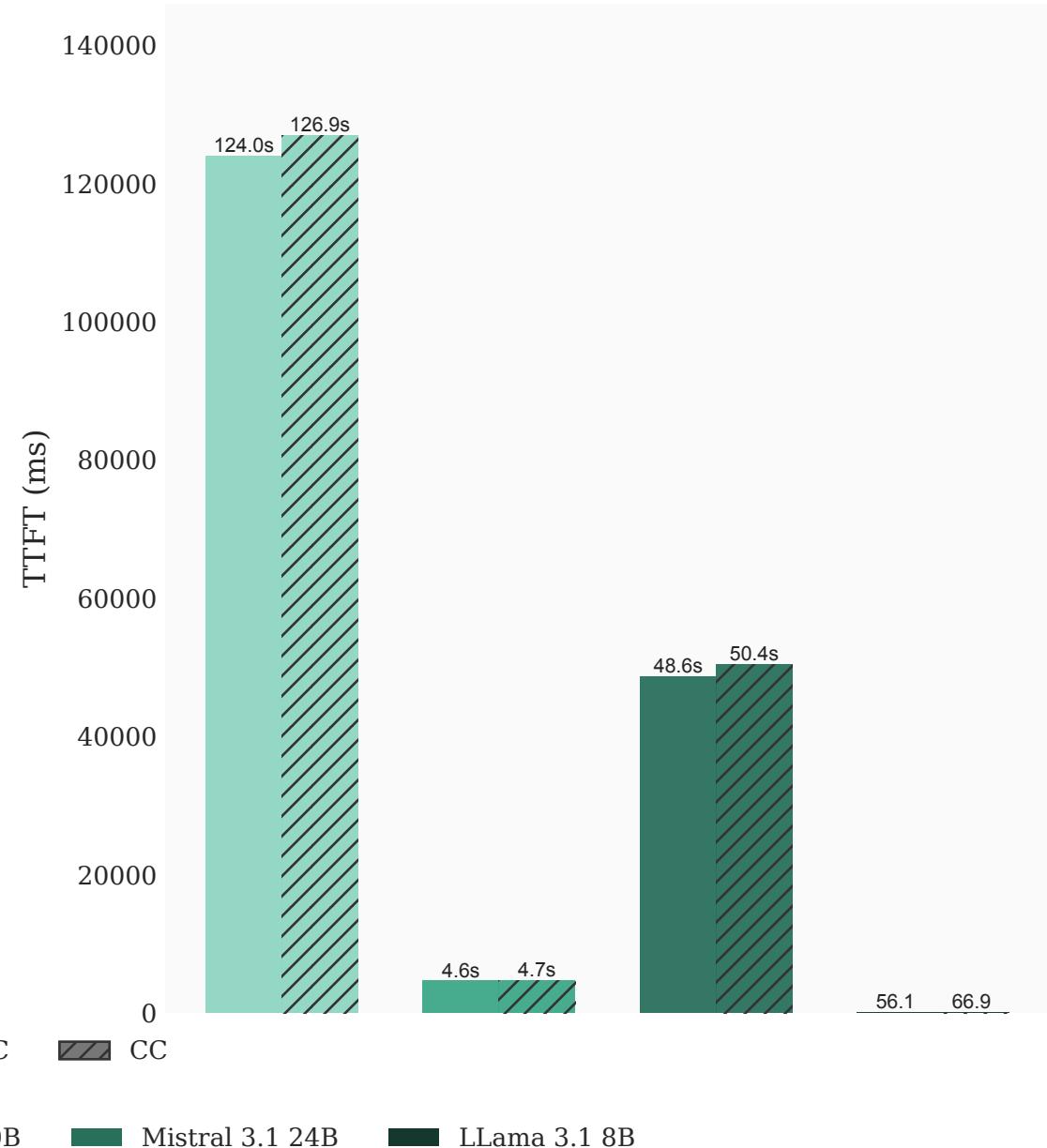


Random (1000 \Rightarrow 1000) (Request Rate 50)

Time to First Token (Mean)

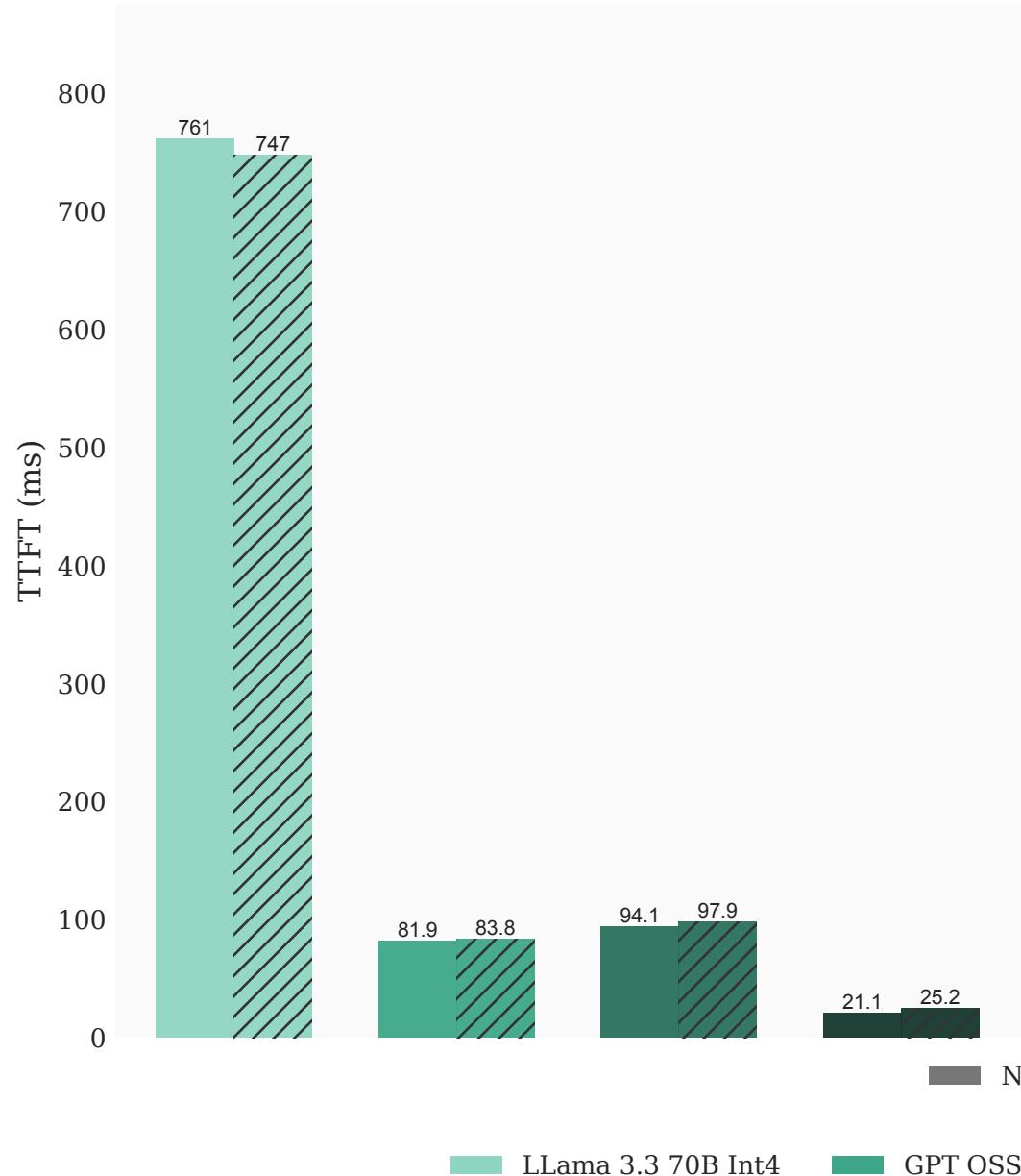


Time to First Token (P99)

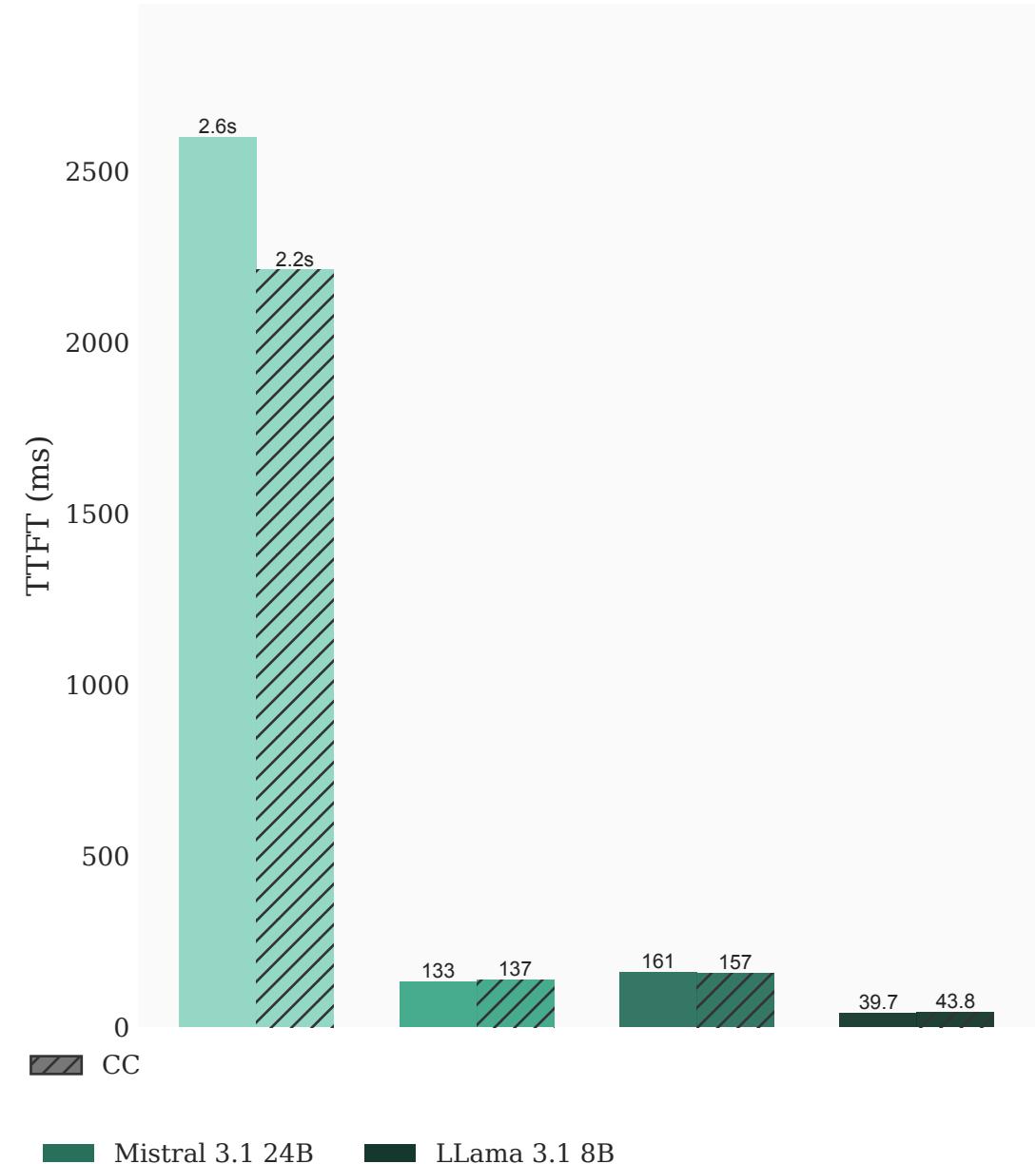


Random (1000 \Rightarrow 1000) (Request Rate 1)

Time to First Token (Mean)

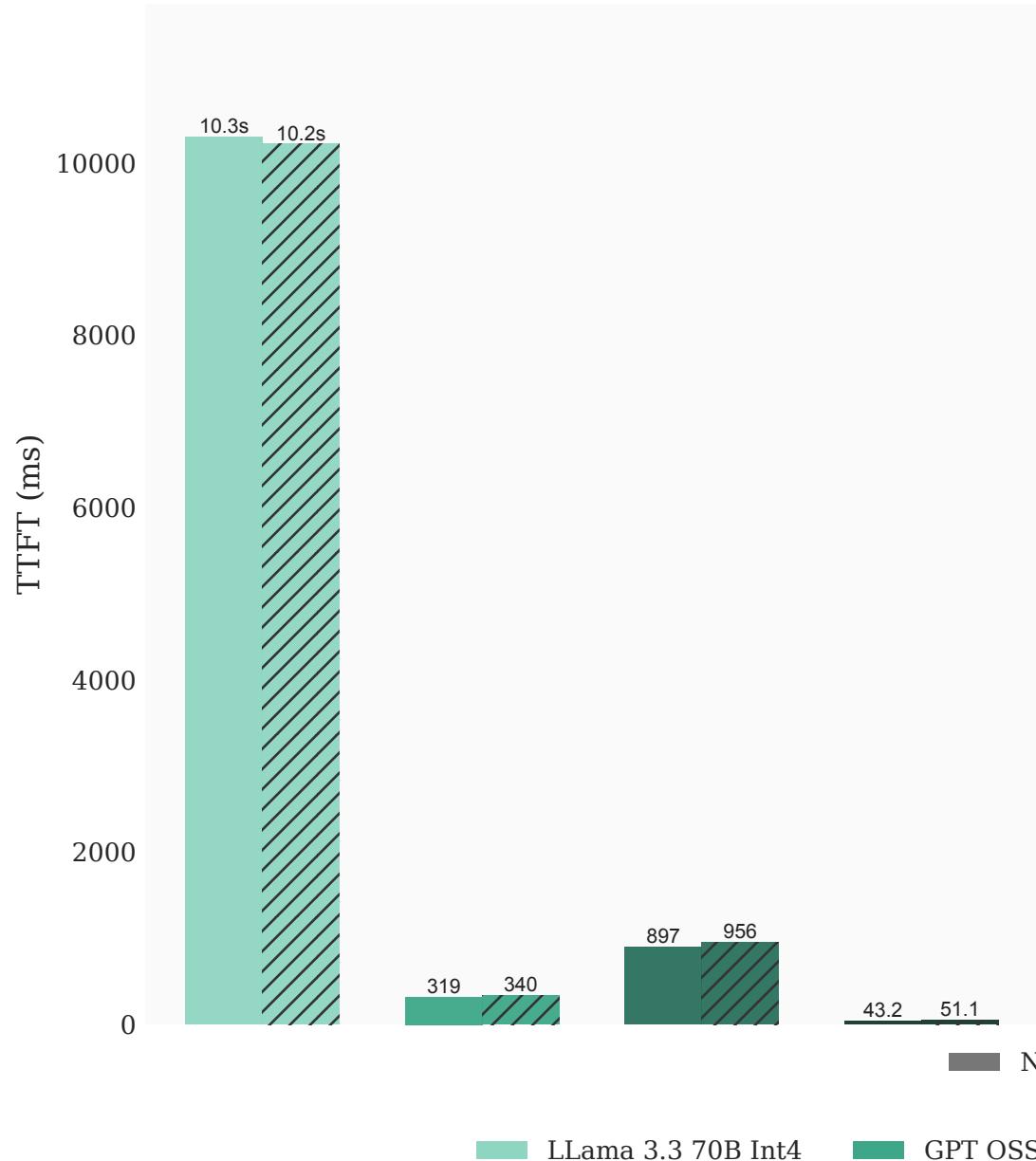


Time to First Token (P99)

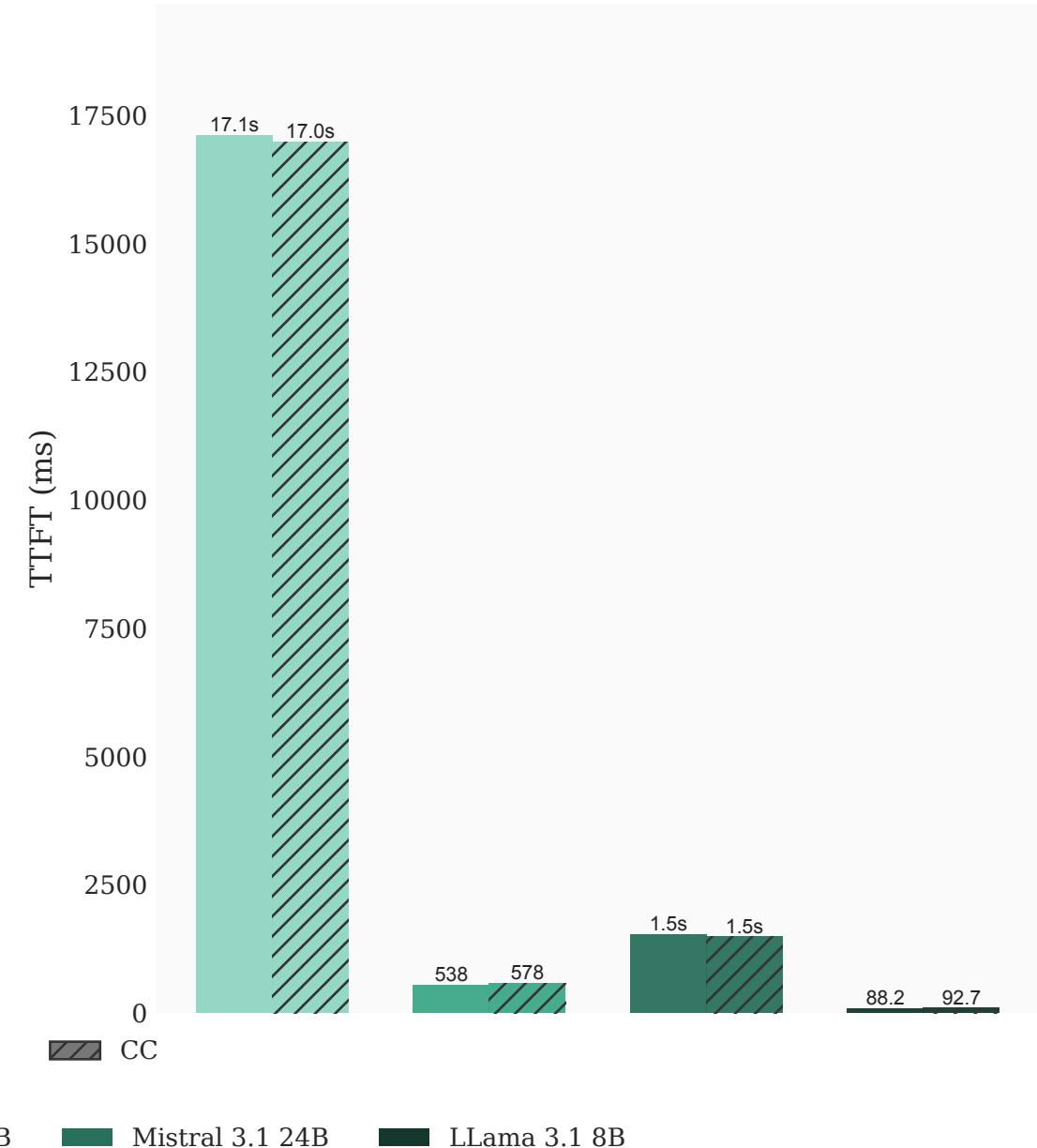


ShareGPT (Request Rate 100)

Time to First Token (Mean)

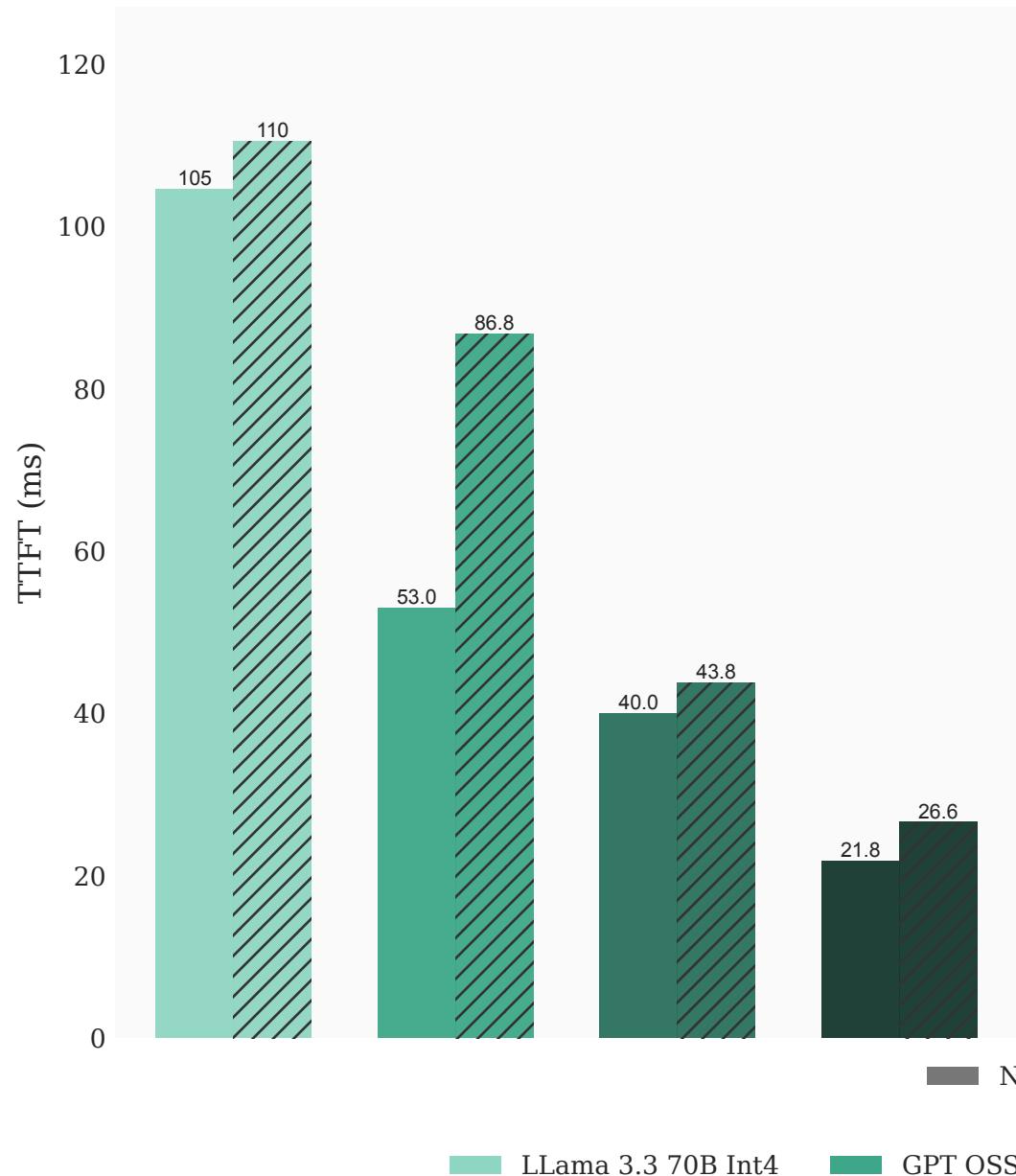


Time to First Token (P99)

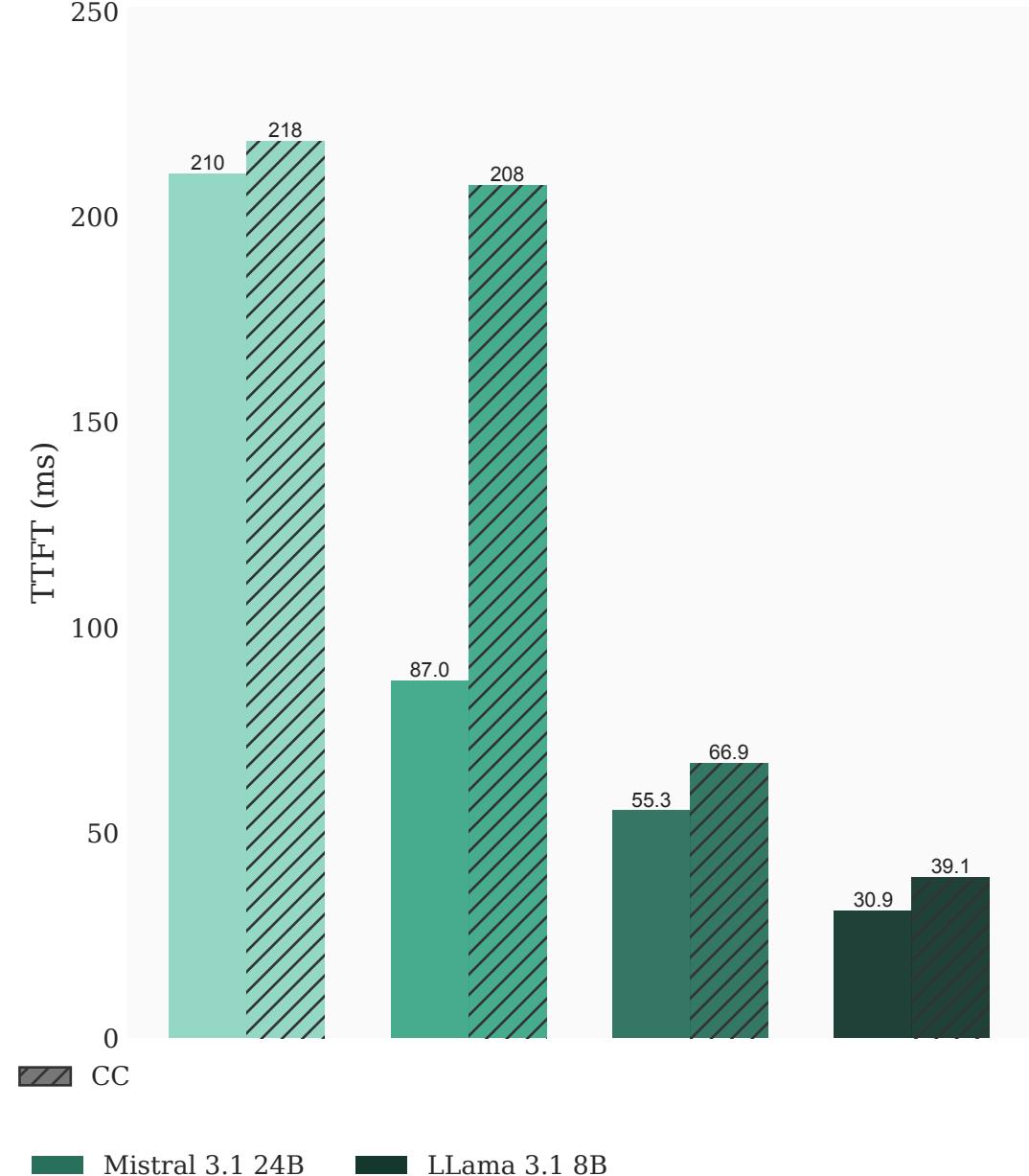


ShareGPT (Request Rate 50)

Time to First Token (Mean)

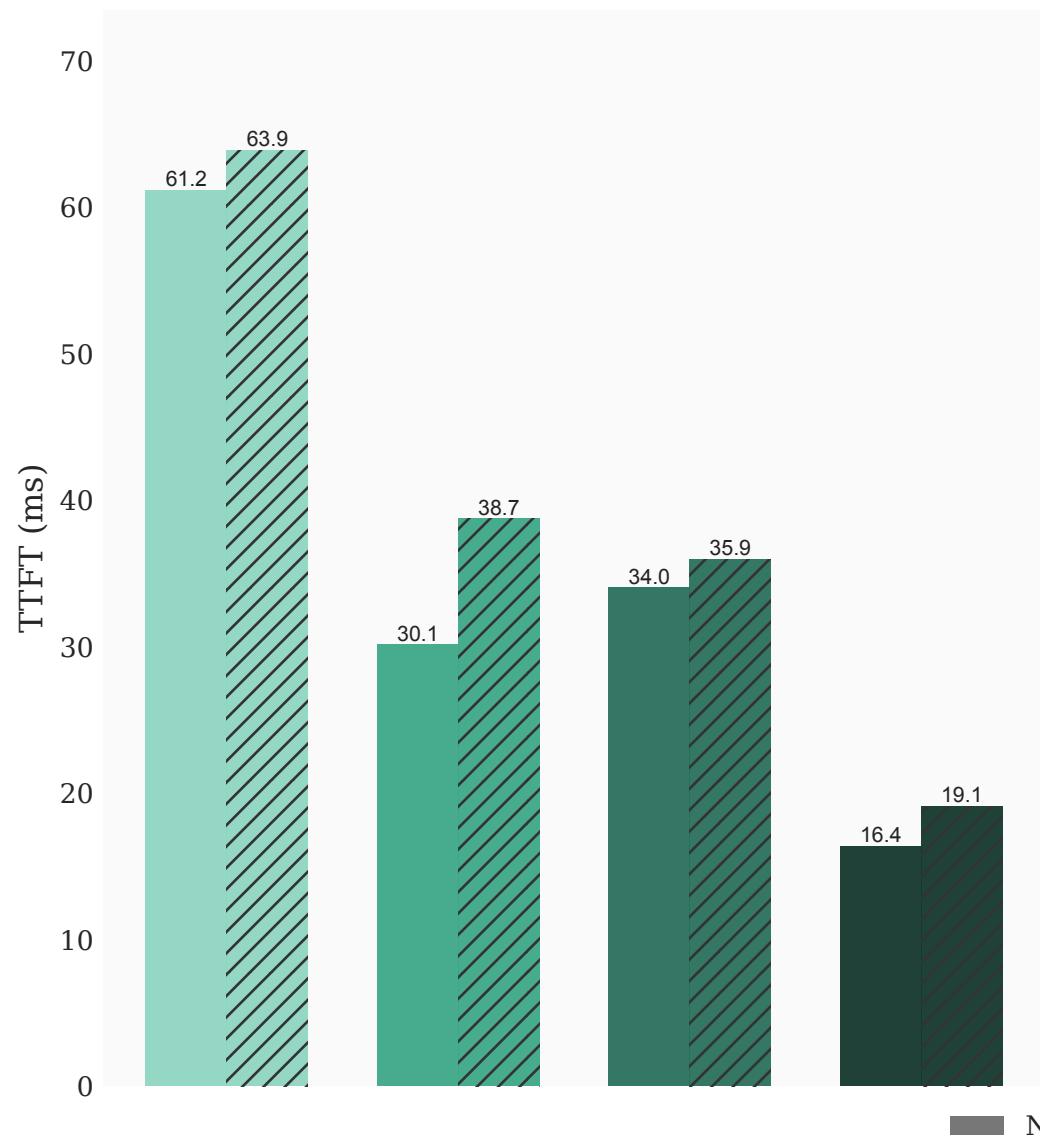


Time to First Token (P99)

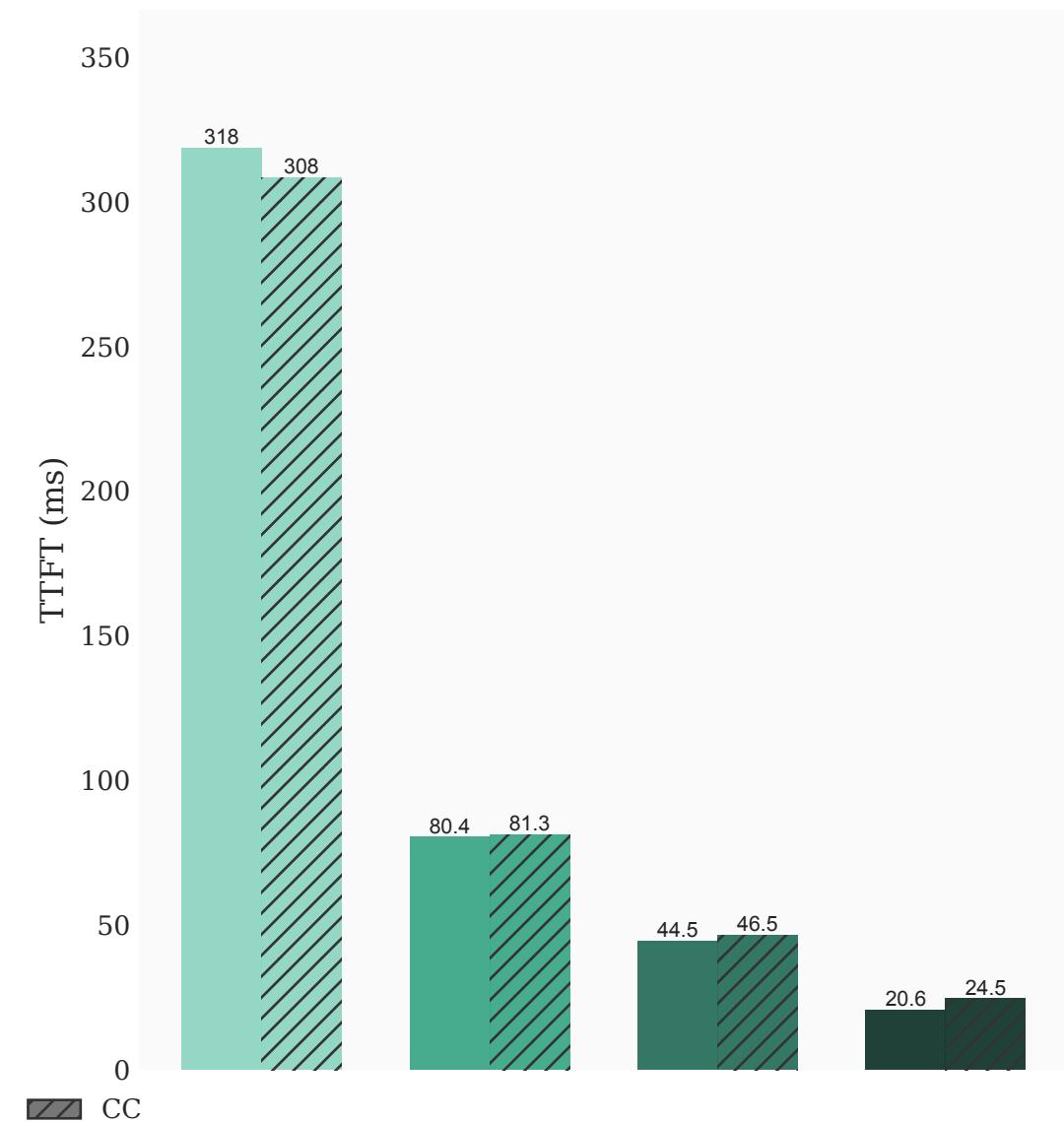


ShareGPT (Request Rate 1)

Time to First Token (Mean)



Time to First Token (P99)



■ Llama 3.3 70B Int4

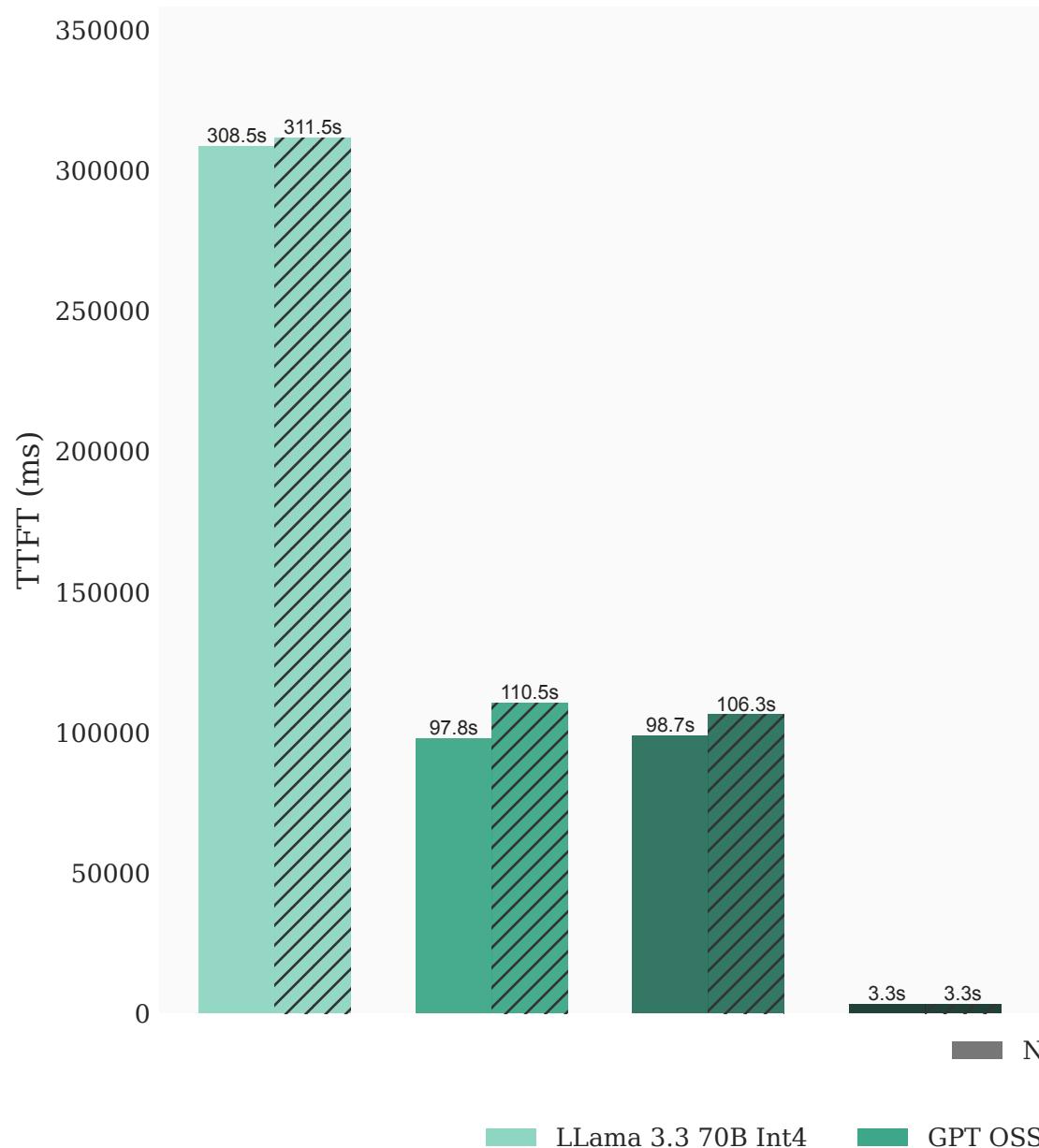
■ GPT OSS 120B

■ Mistral 3.1 24B

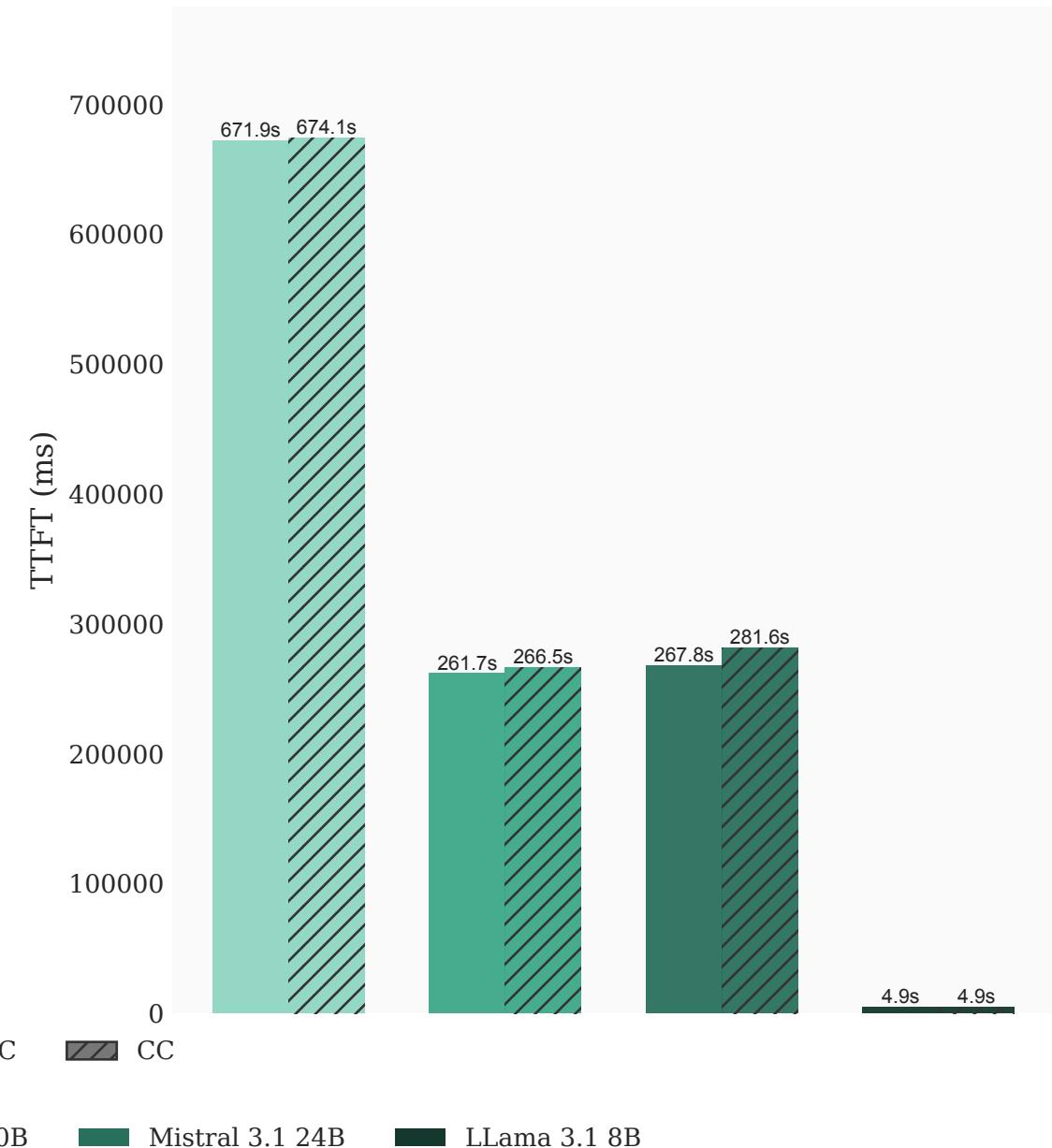
■ Llama 3.1 8B

Edit 10K Characters (Request Rate 100)

Time to First Token (Mean)

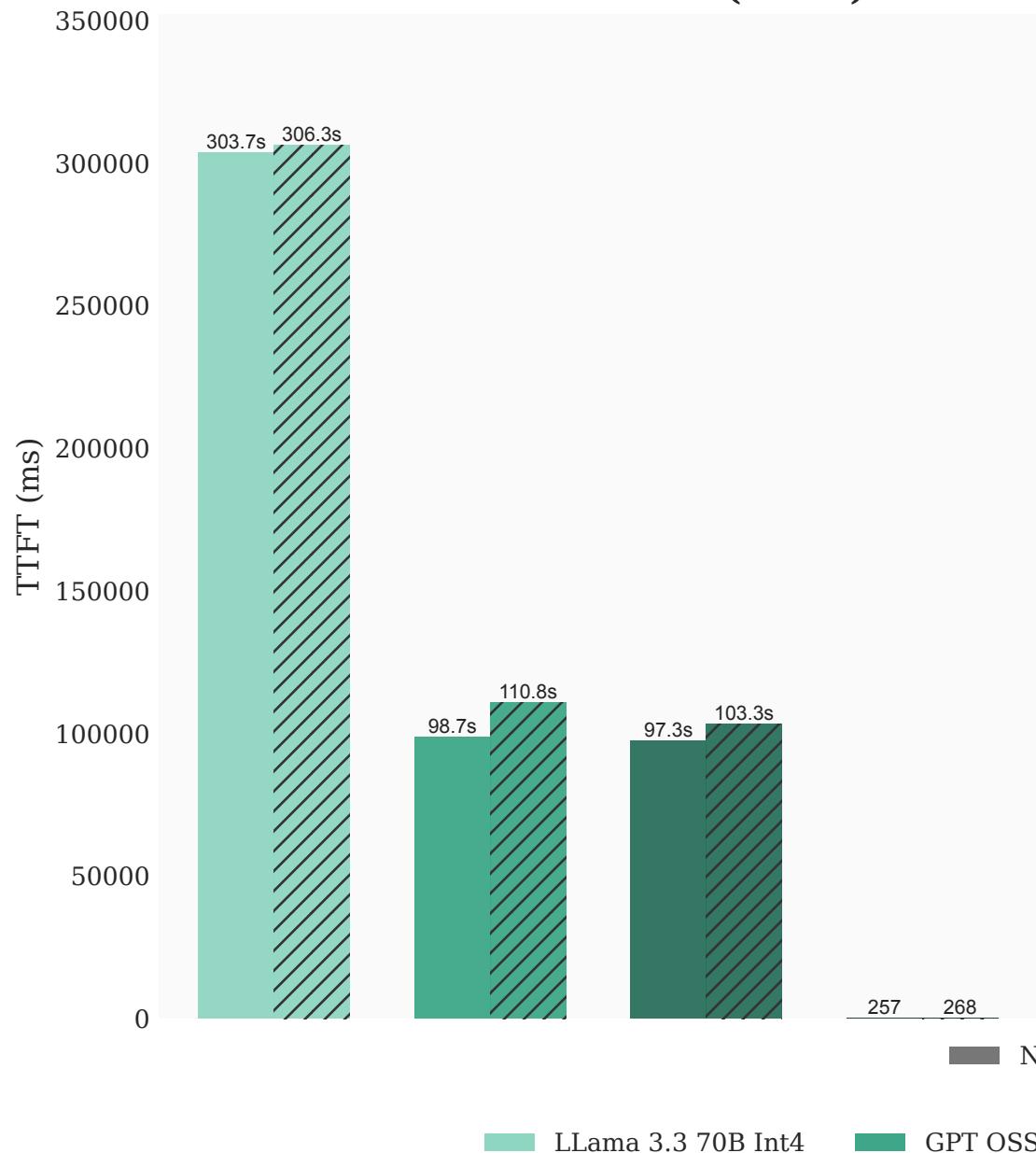


Time to First Token (P99)

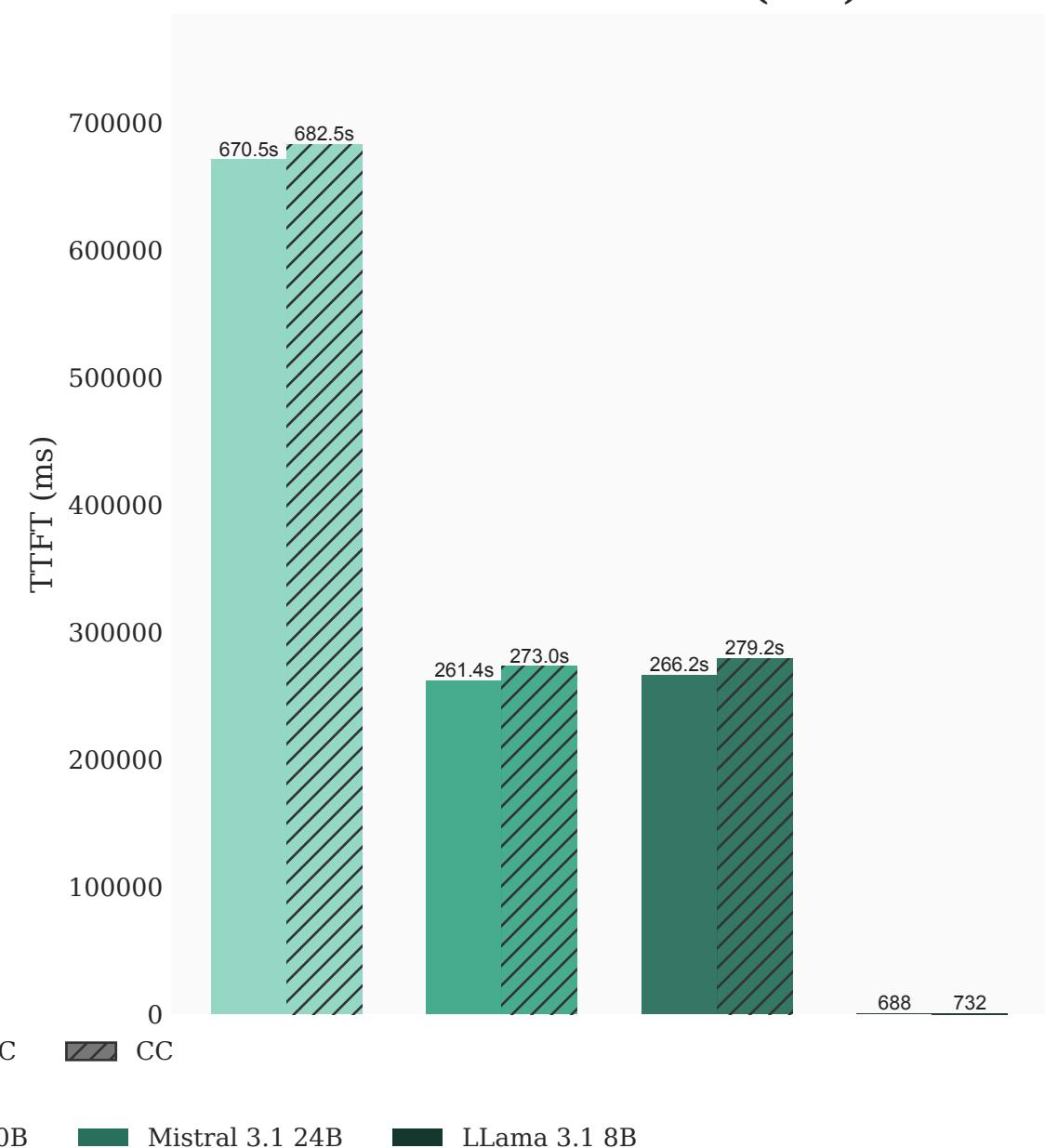


Edit 10K Characters (Request Rate 50)

Time to First Token (Mean)

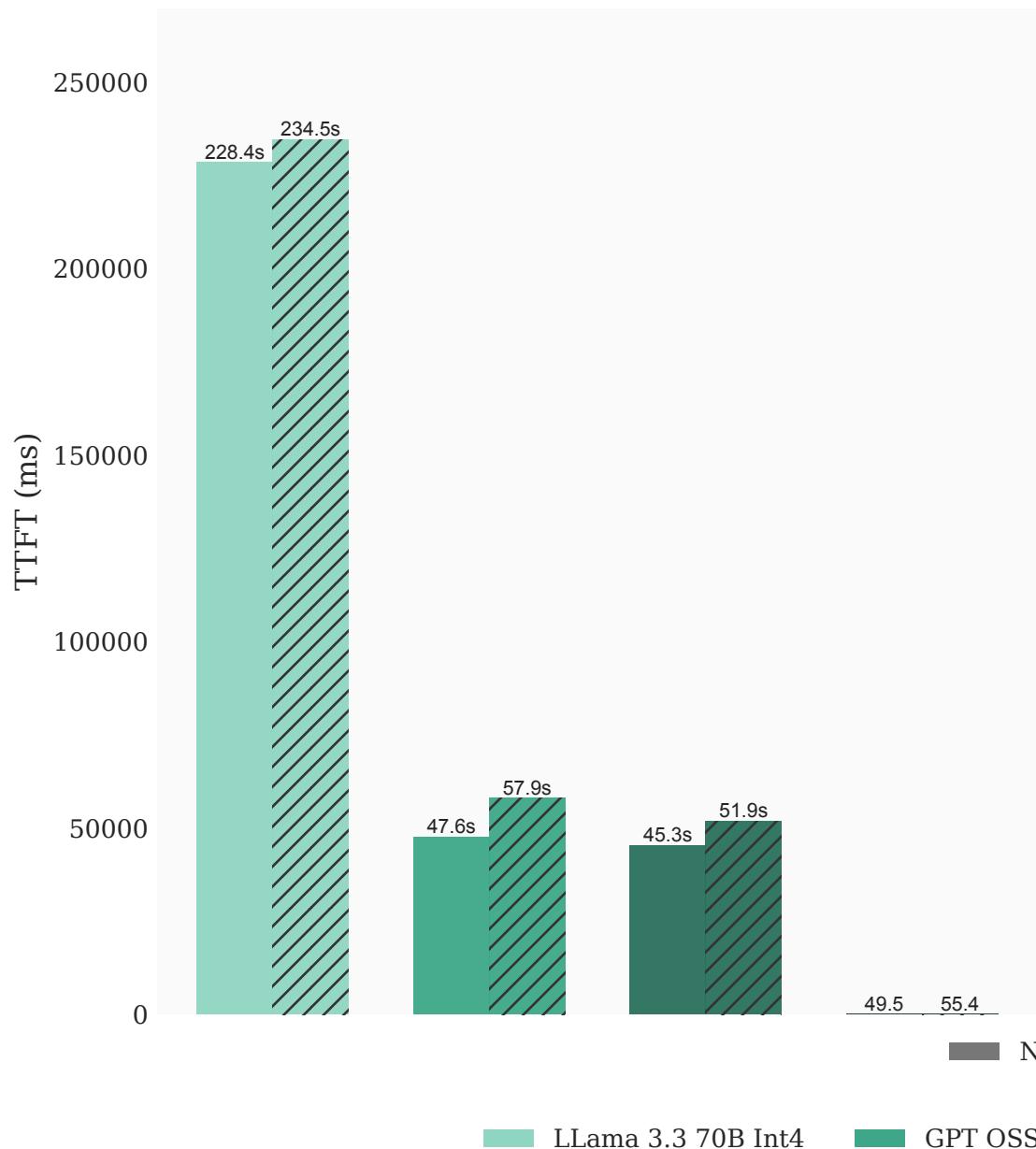


Time to First Token (P99)

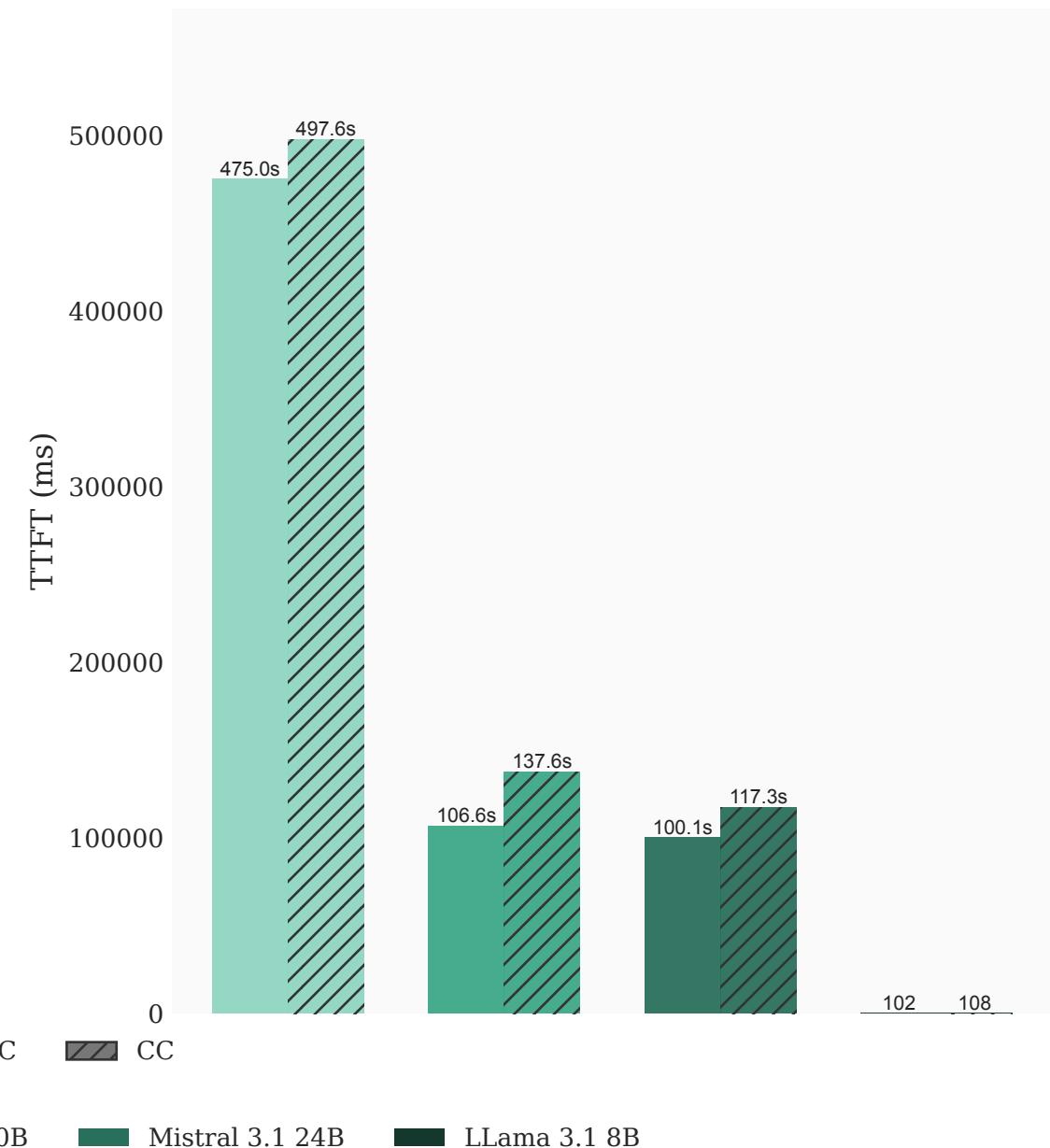


Edit 10K Characters (Request Rate 1)

Time to First Token (Mean)

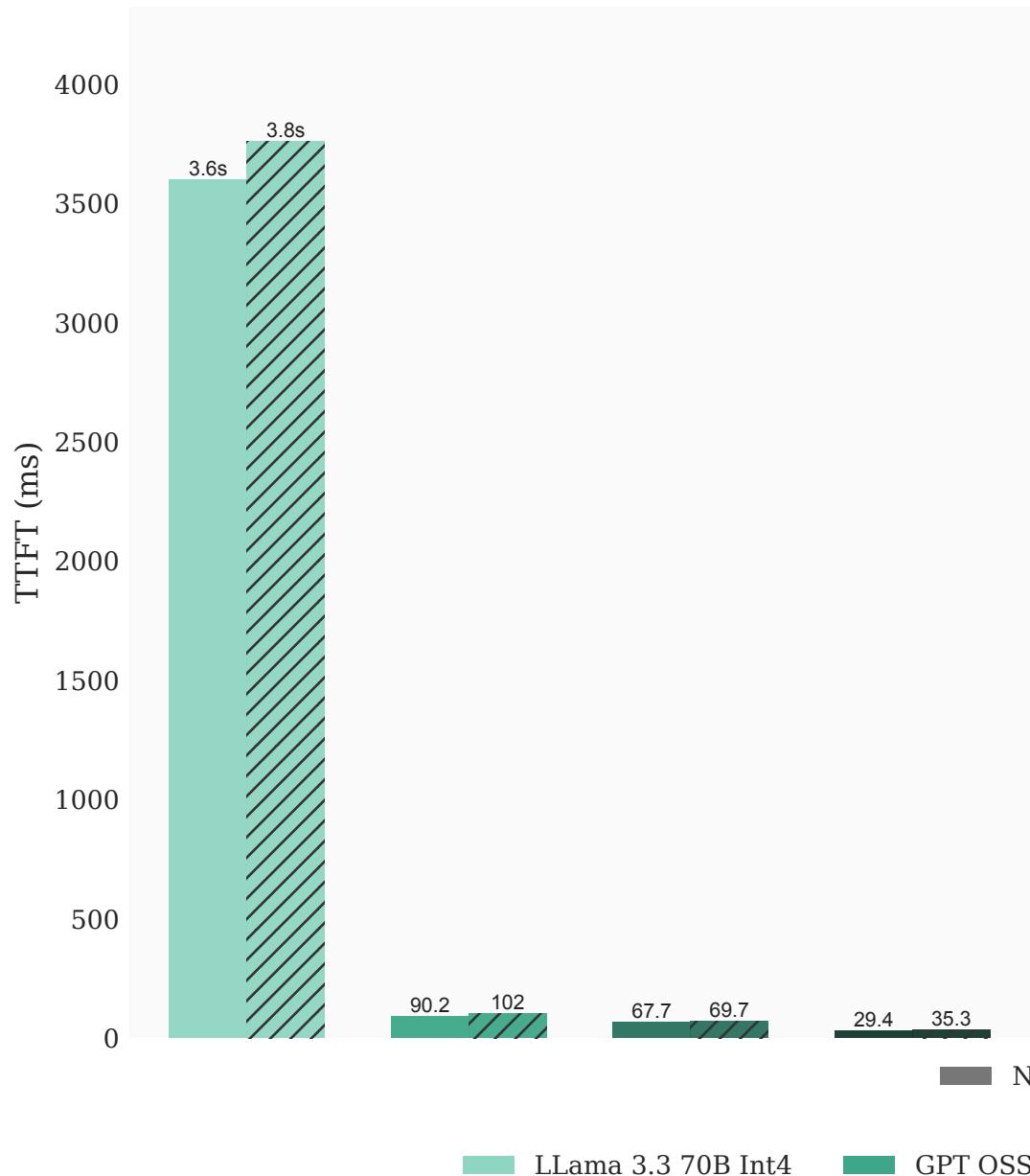


Time to First Token (P99)

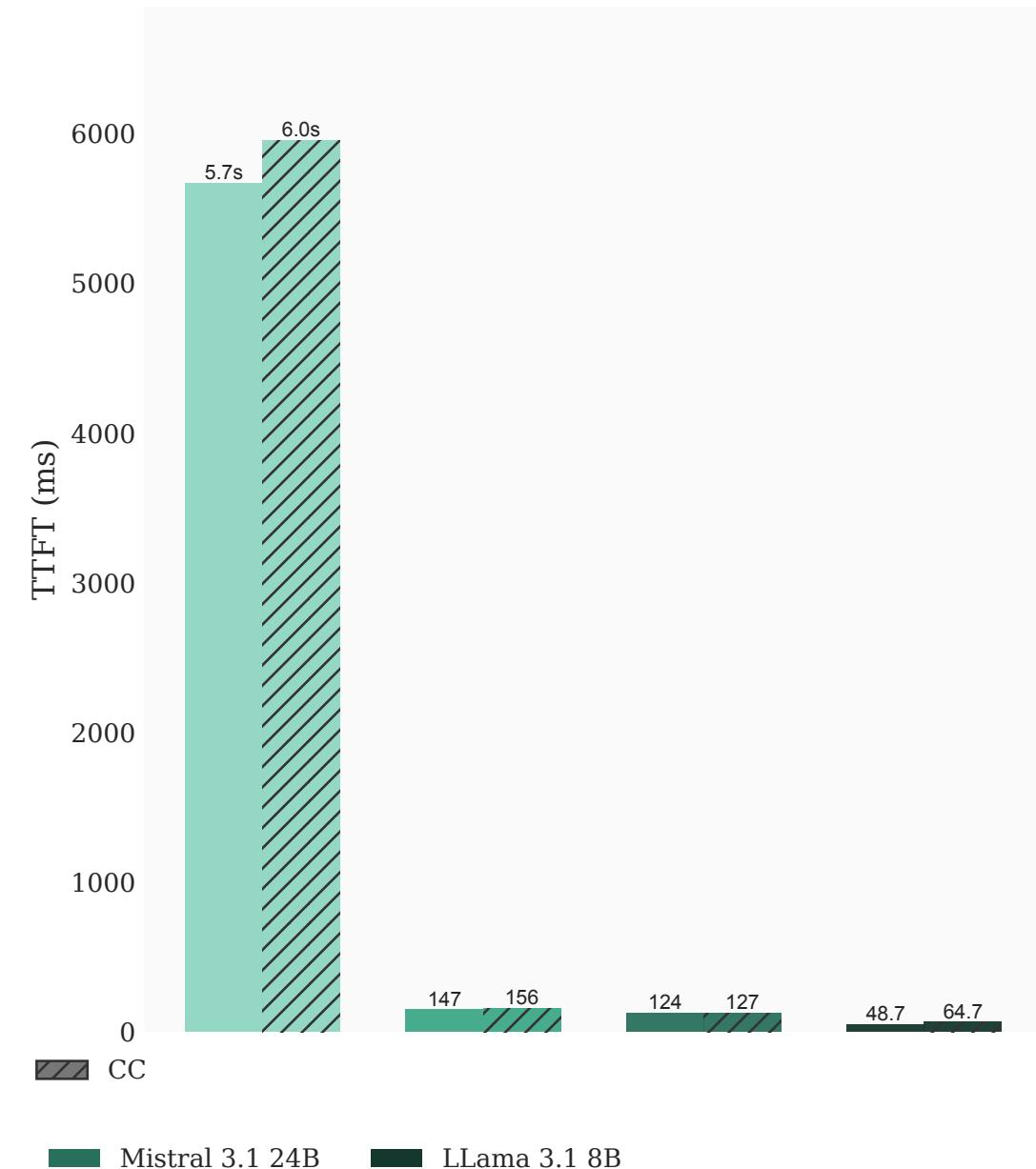


Numina Math (Request Rate 100)

Time to First Token (Mean)

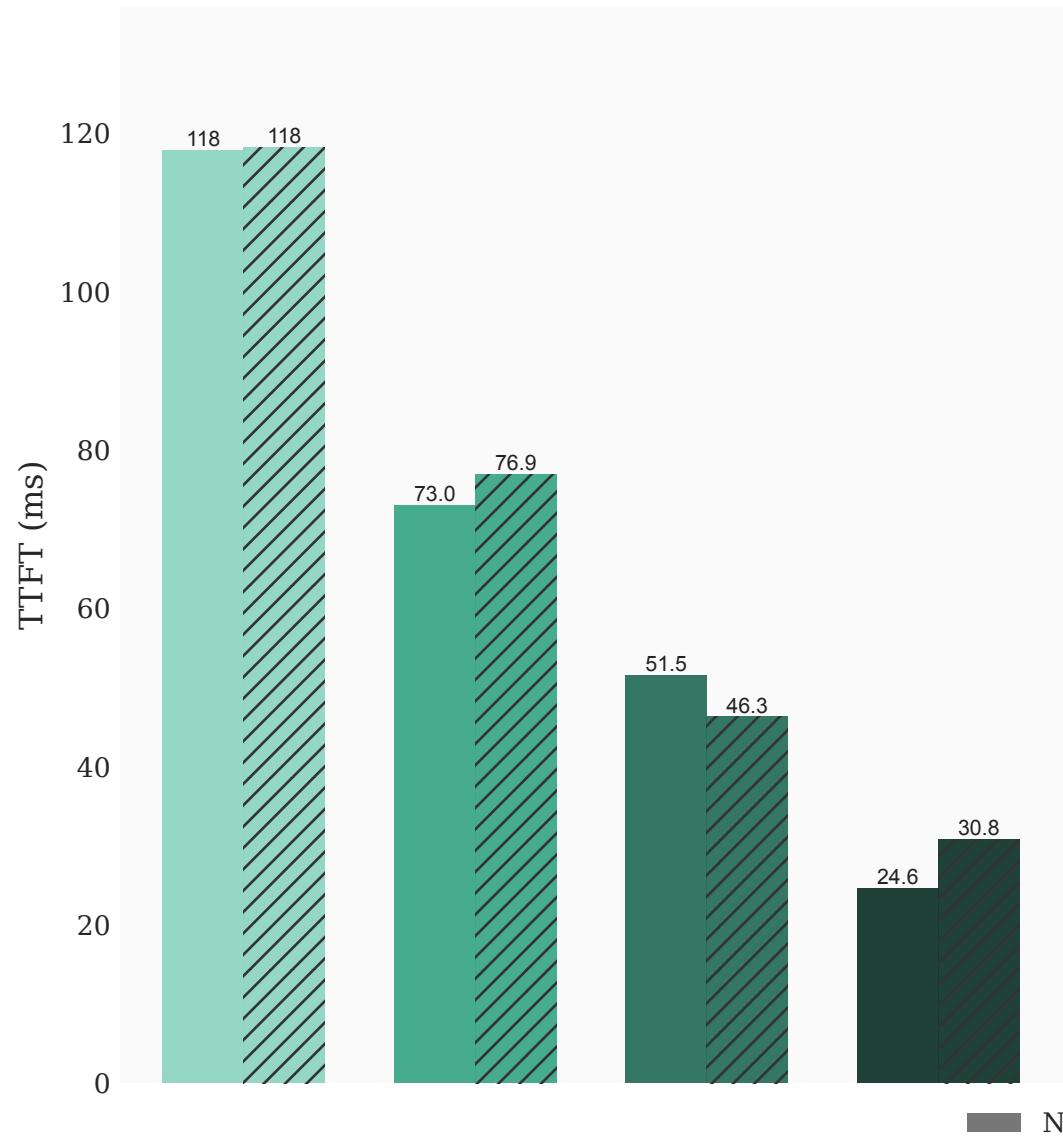


Time to First Token (P99)

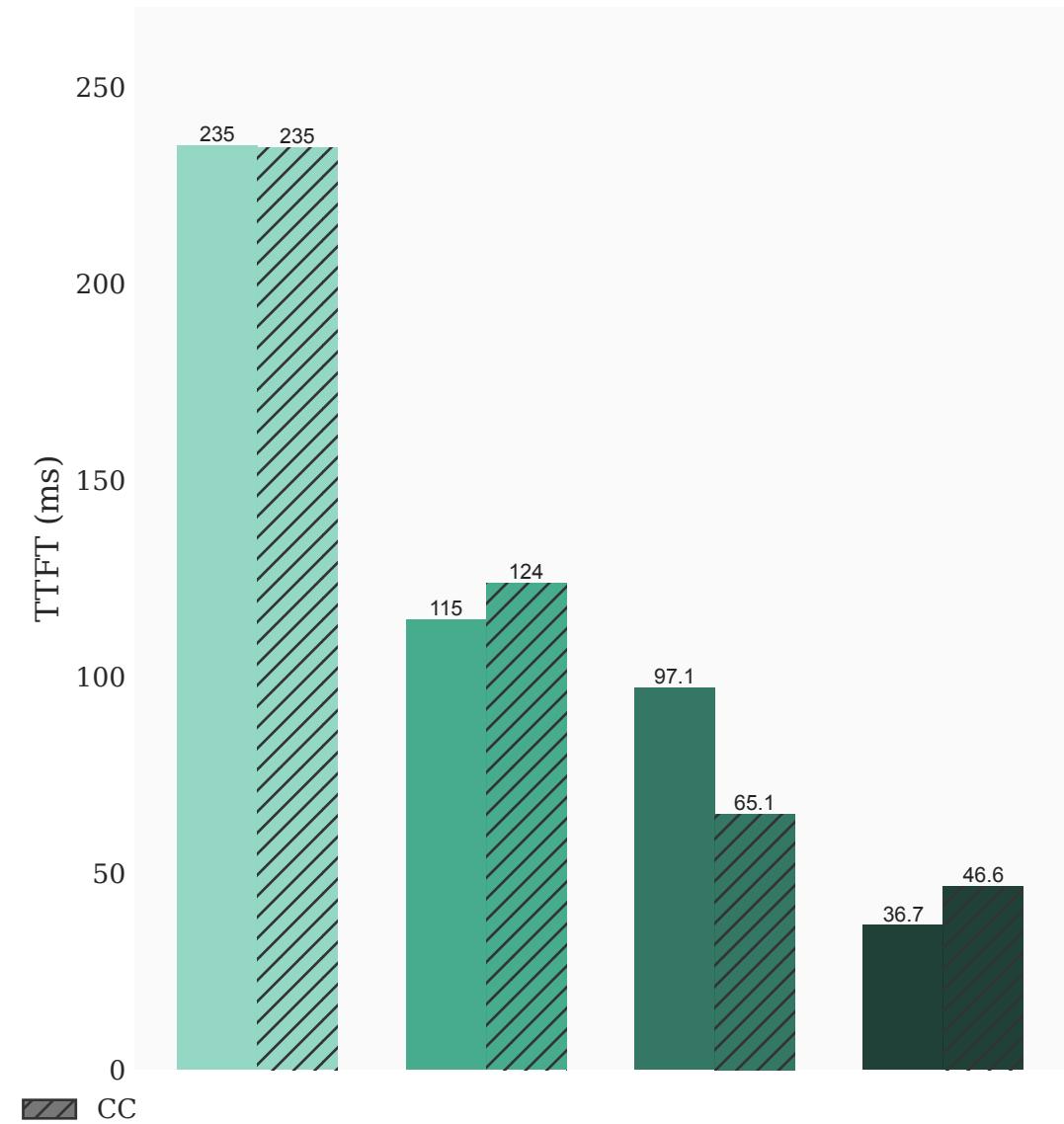


Numina Math (Request Rate 50)

Time to First Token (Mean)



Time to First Token (P99)



LLama 3.3 70B Int4

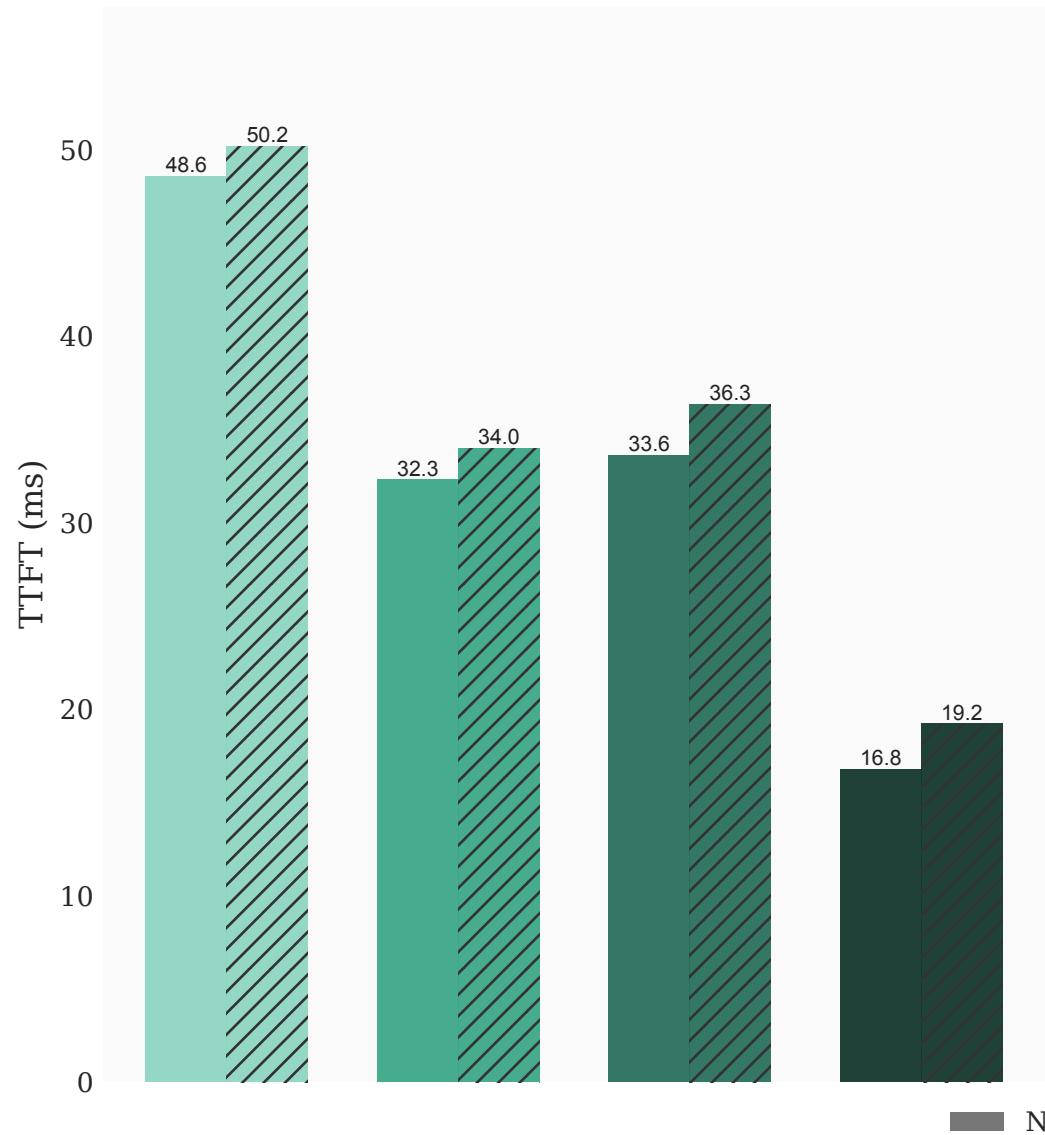
GPT OSS 120B

Mistral 3.1 24B

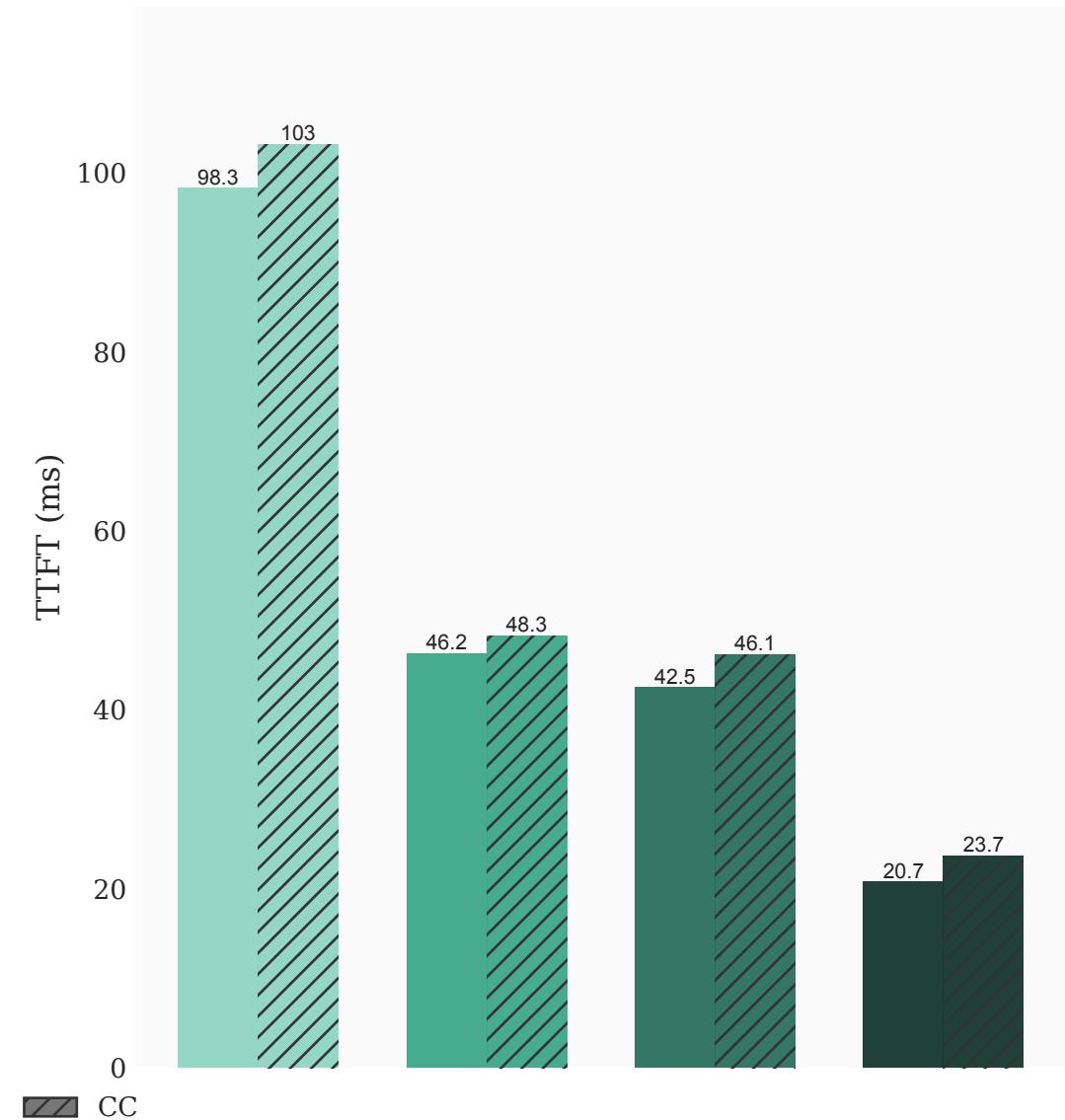
LLama 3.1 8B

Numina Math (Request Rate 1)

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