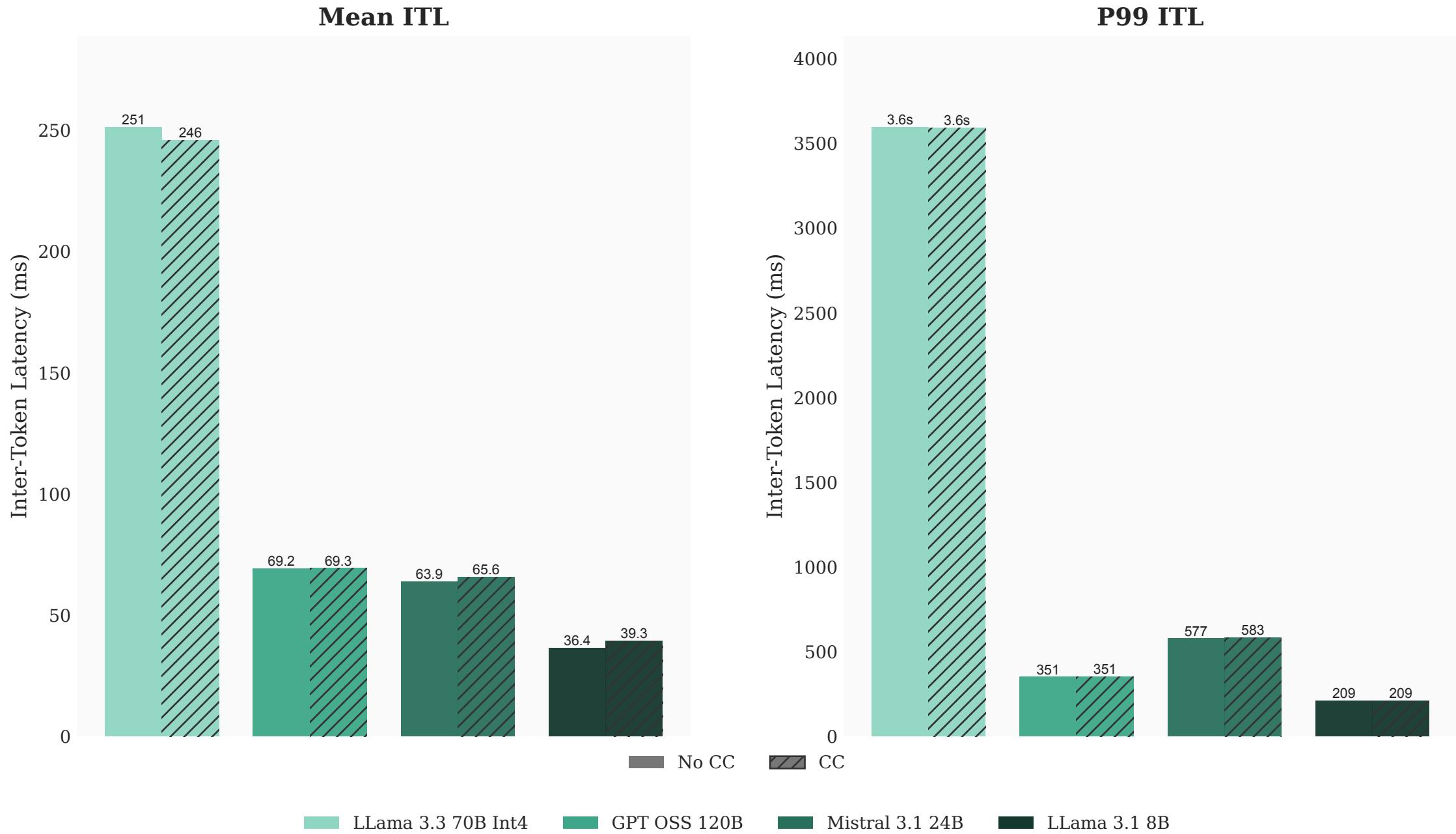
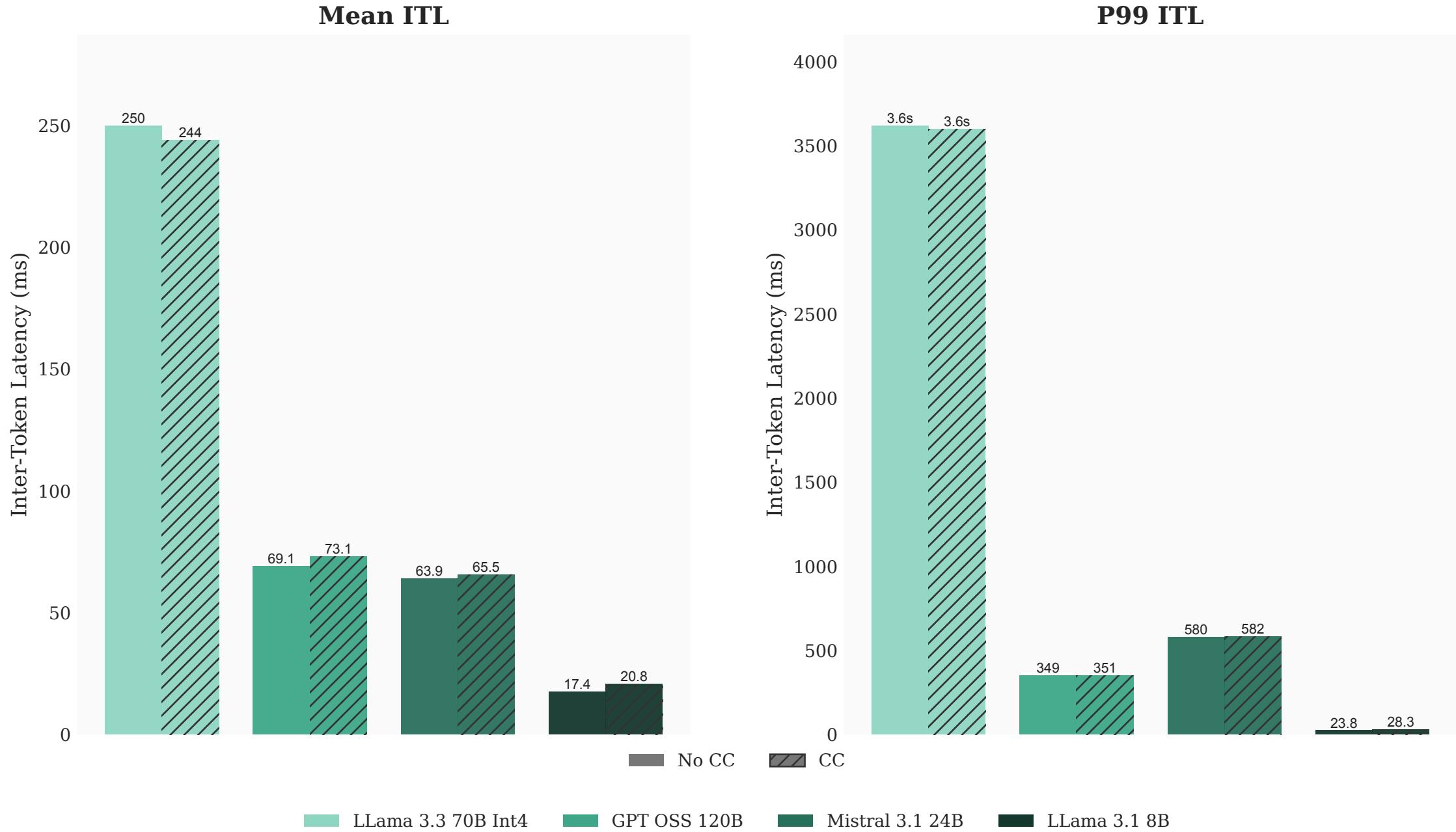


# Random (1500 $\Rightarrow$ 250) (Request Rate 100)

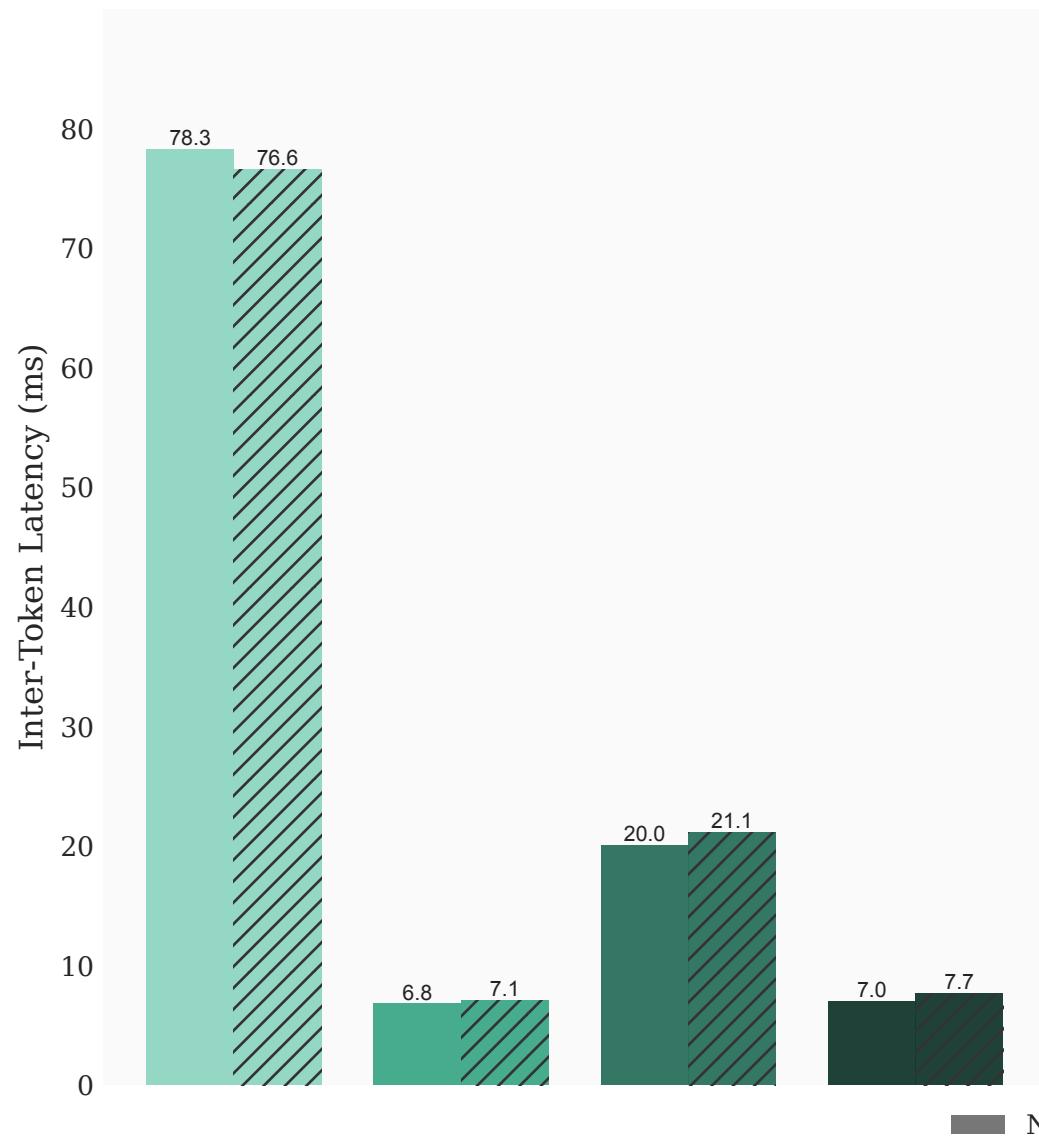


# Random (1500 $\Rightarrow$ 250) (Request Rate 50)

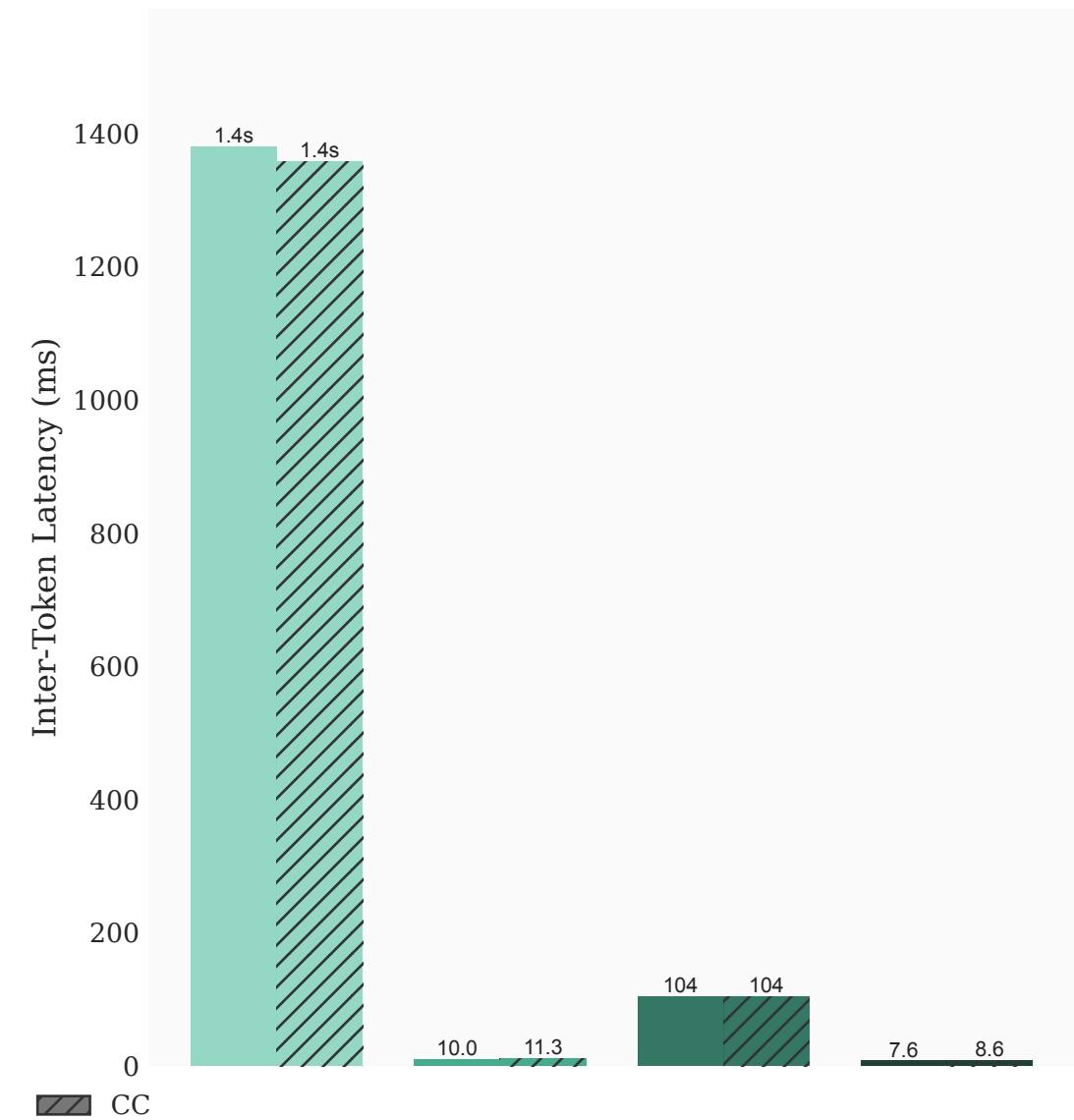


# Random (1500 $\Rightarrow$ 250) (Request Rate 1)

## Mean ITL



## P99 ITL



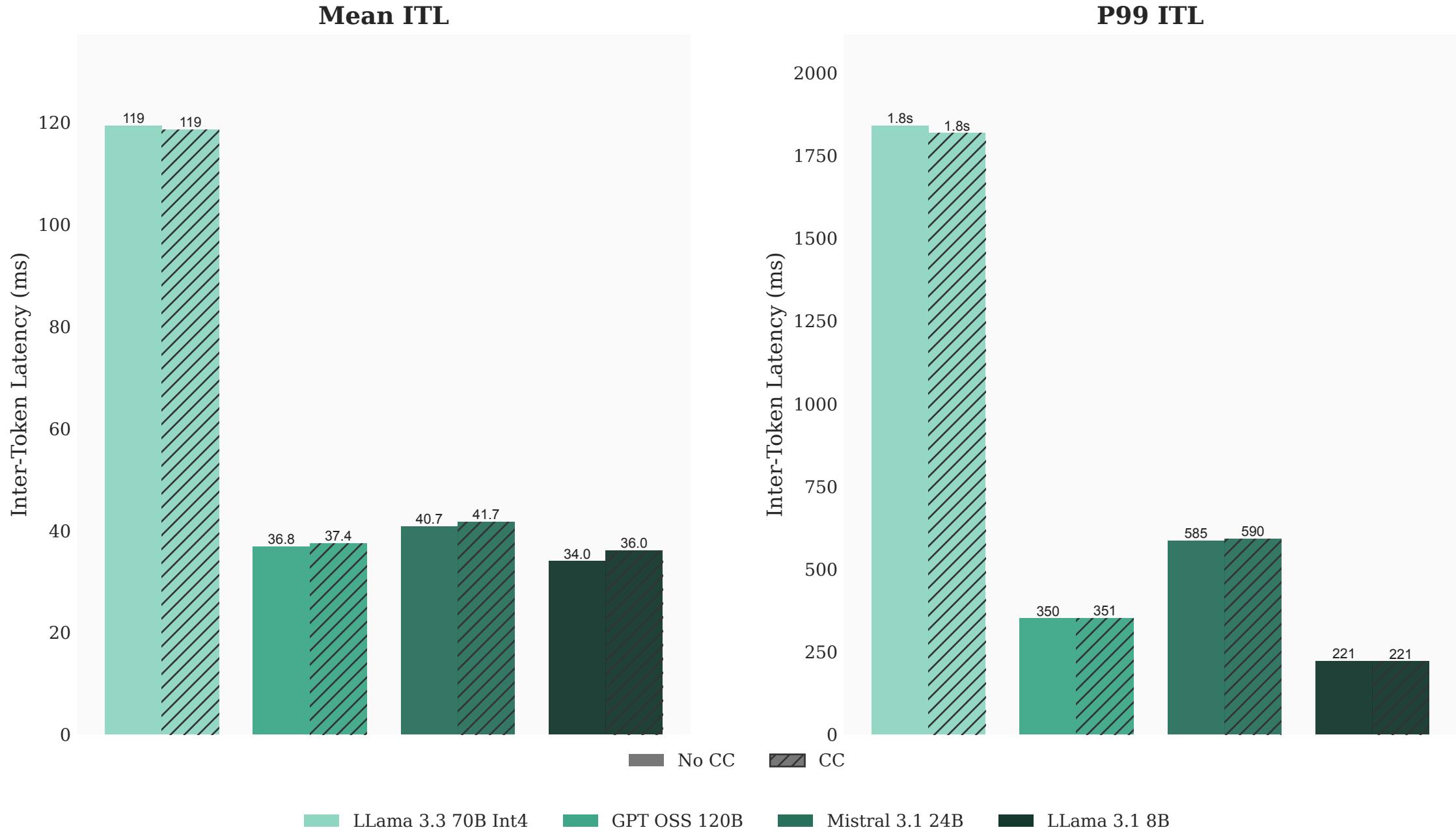
■ LLama 3.3 70B Int4

■ GPT OSS 120B

■ Mistral 3.1 24B

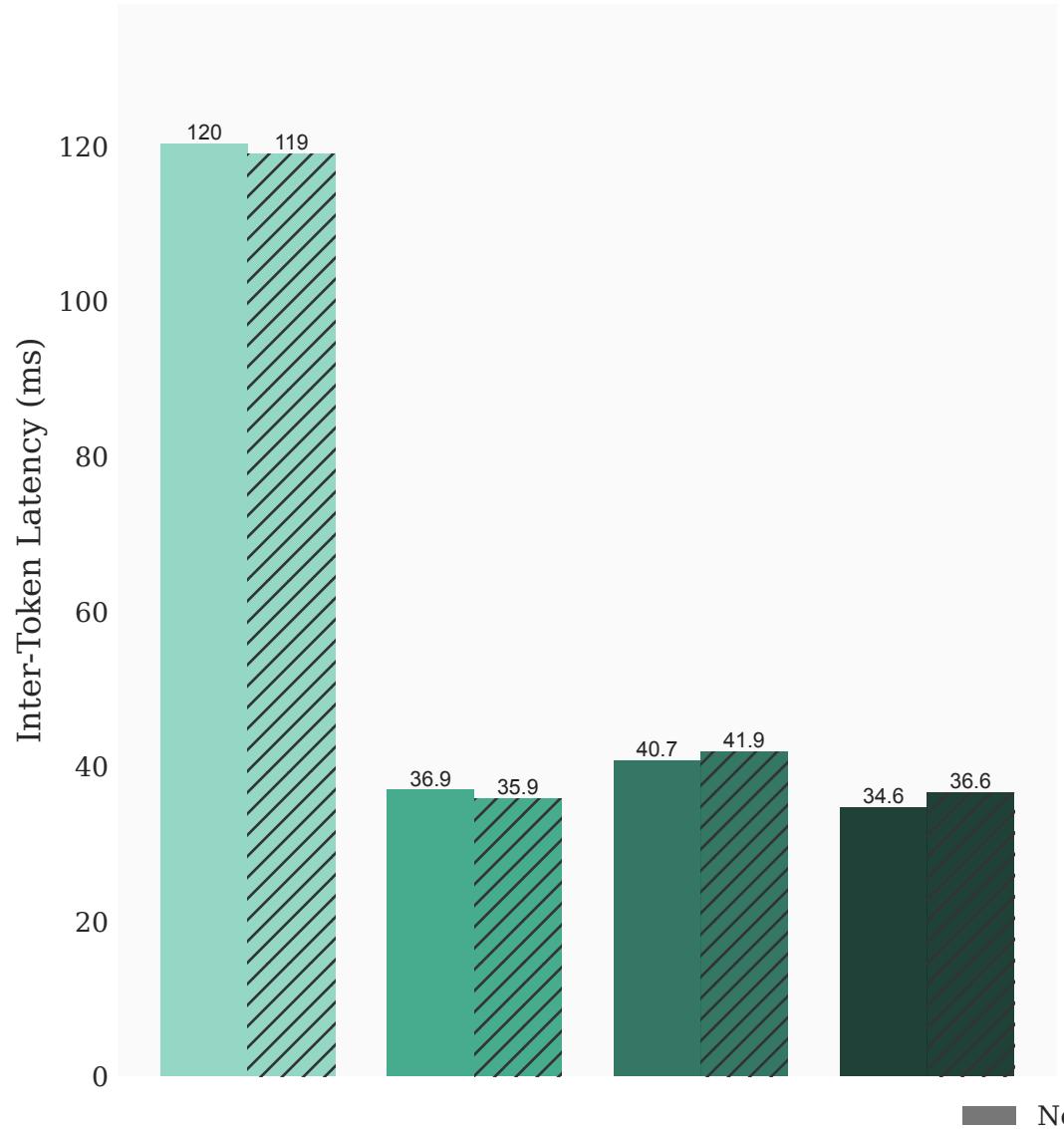
■ LLama 3.1 8B

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

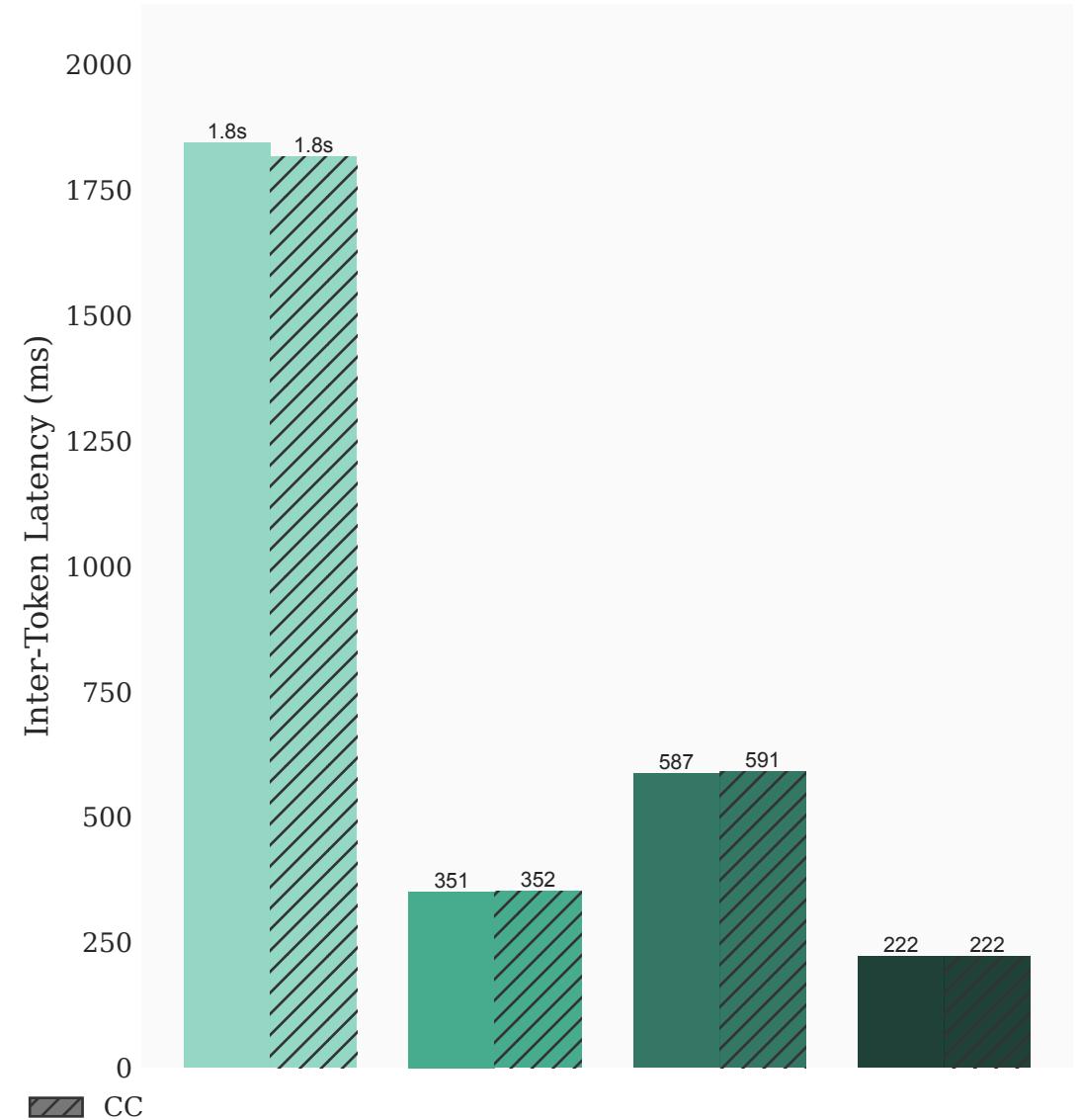


# Random (4000 $\Rightarrow$ 1000) (Request Rate 50)

Mean ITL



P99 ITL



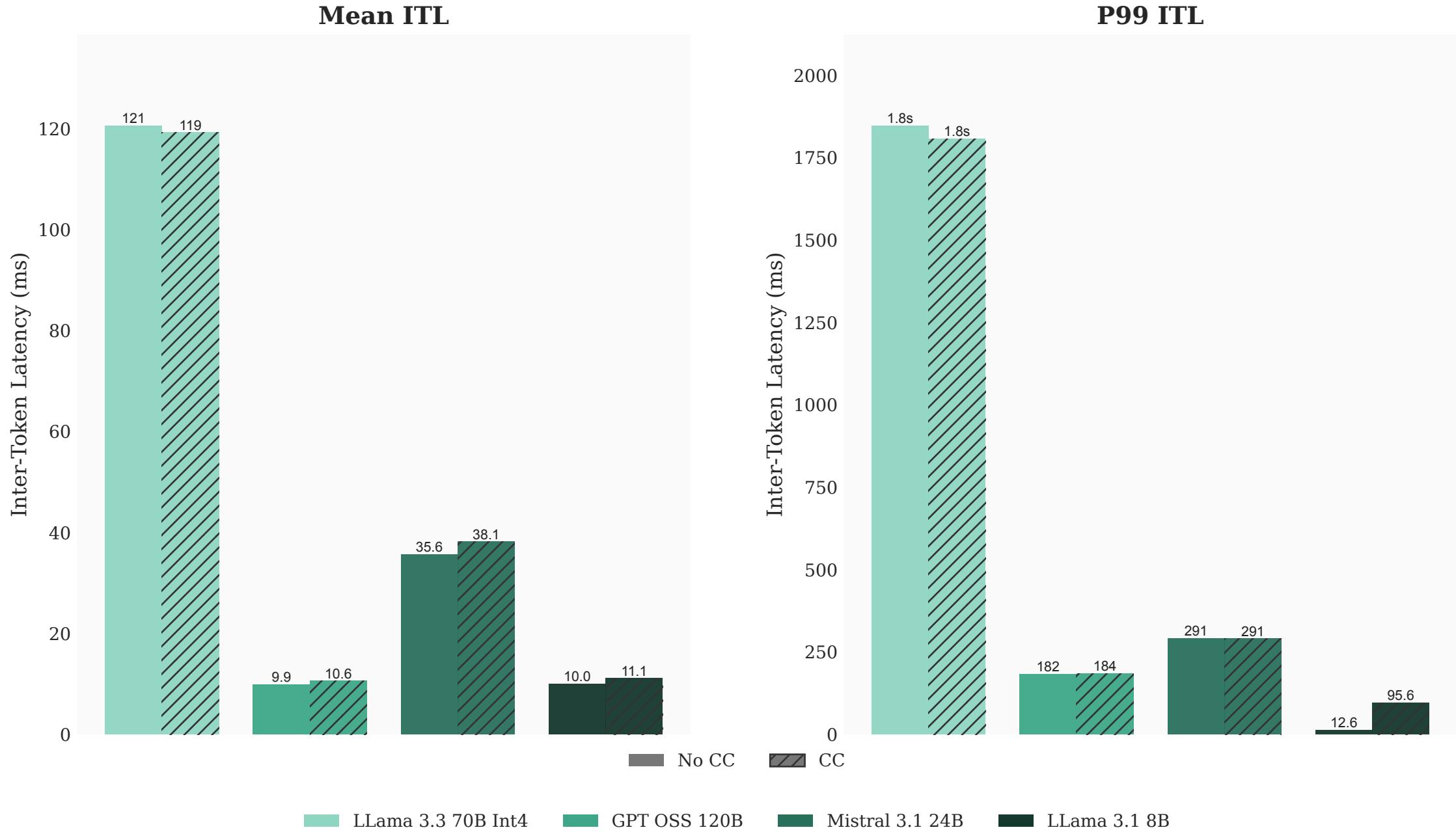
LLama 3.3 70B Int4

GPT OSS 120B

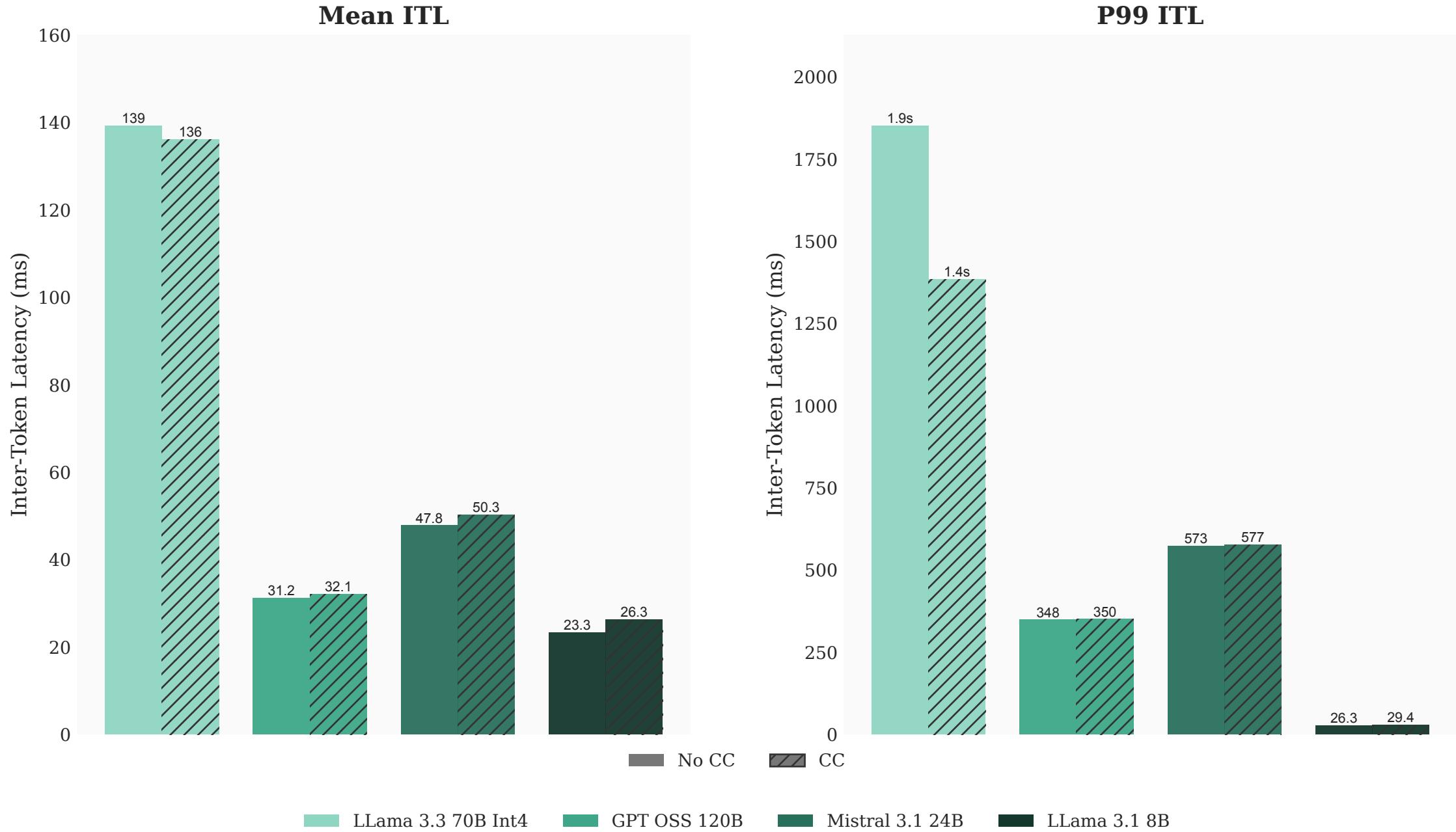
Mistral 3.1 24B

LLama 3.1 8B

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

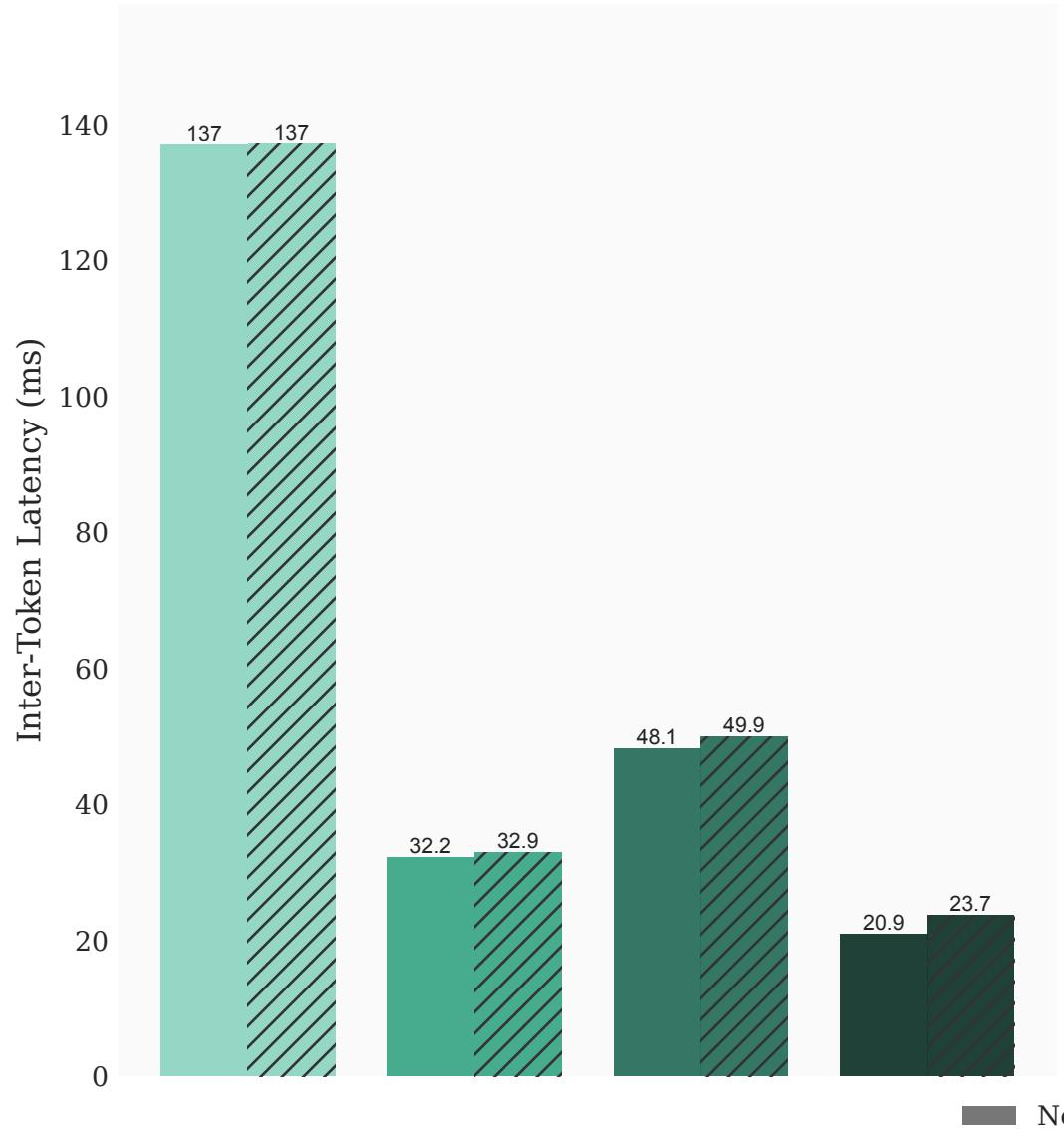


# Random (1000 $\Rightarrow$ 1000) (Request Rate 100)

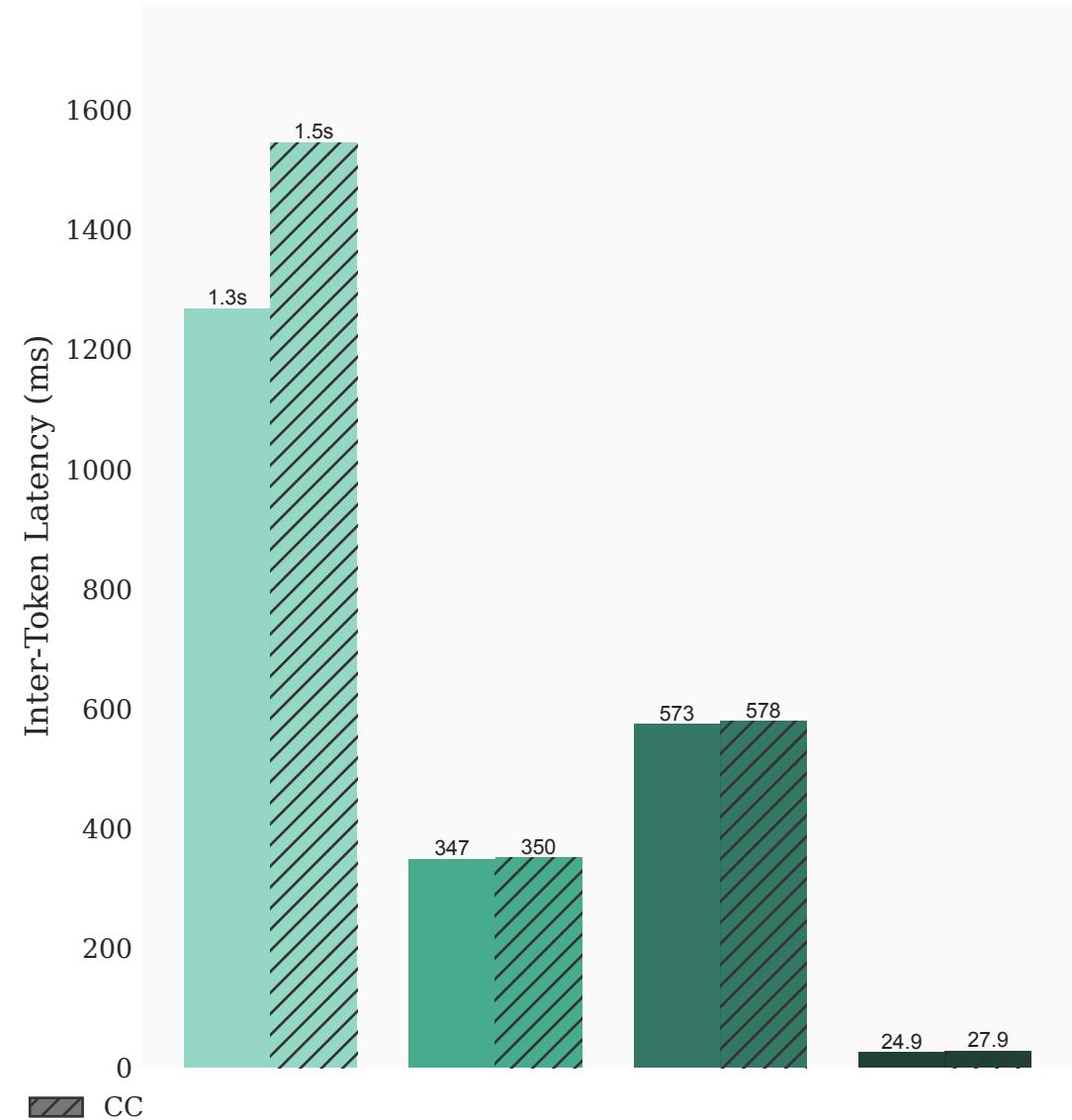


# Random (1000 $\Rightarrow$ 1000) (Request Rate 50)

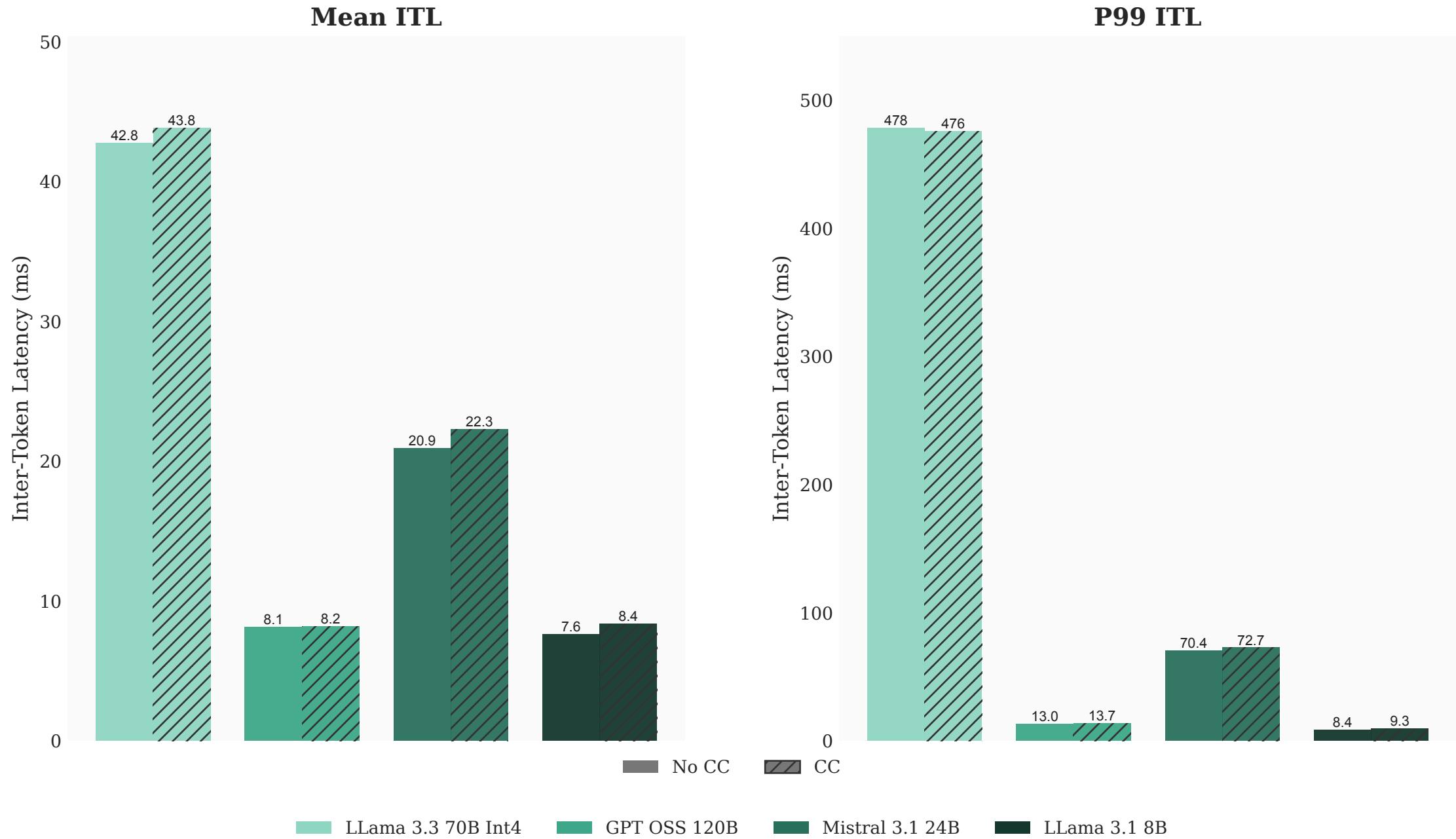
Mean ITL



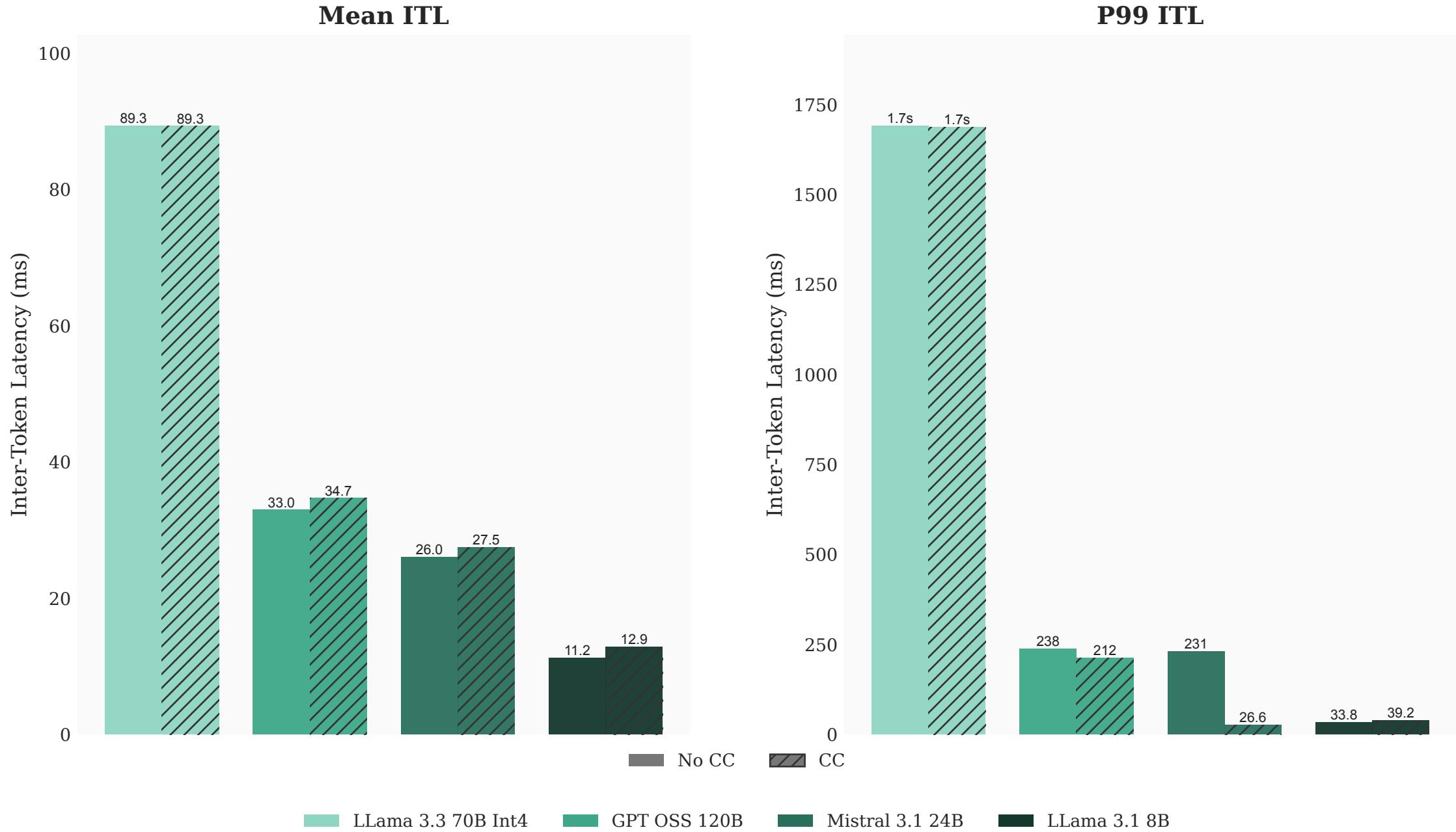
P99 ITL



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

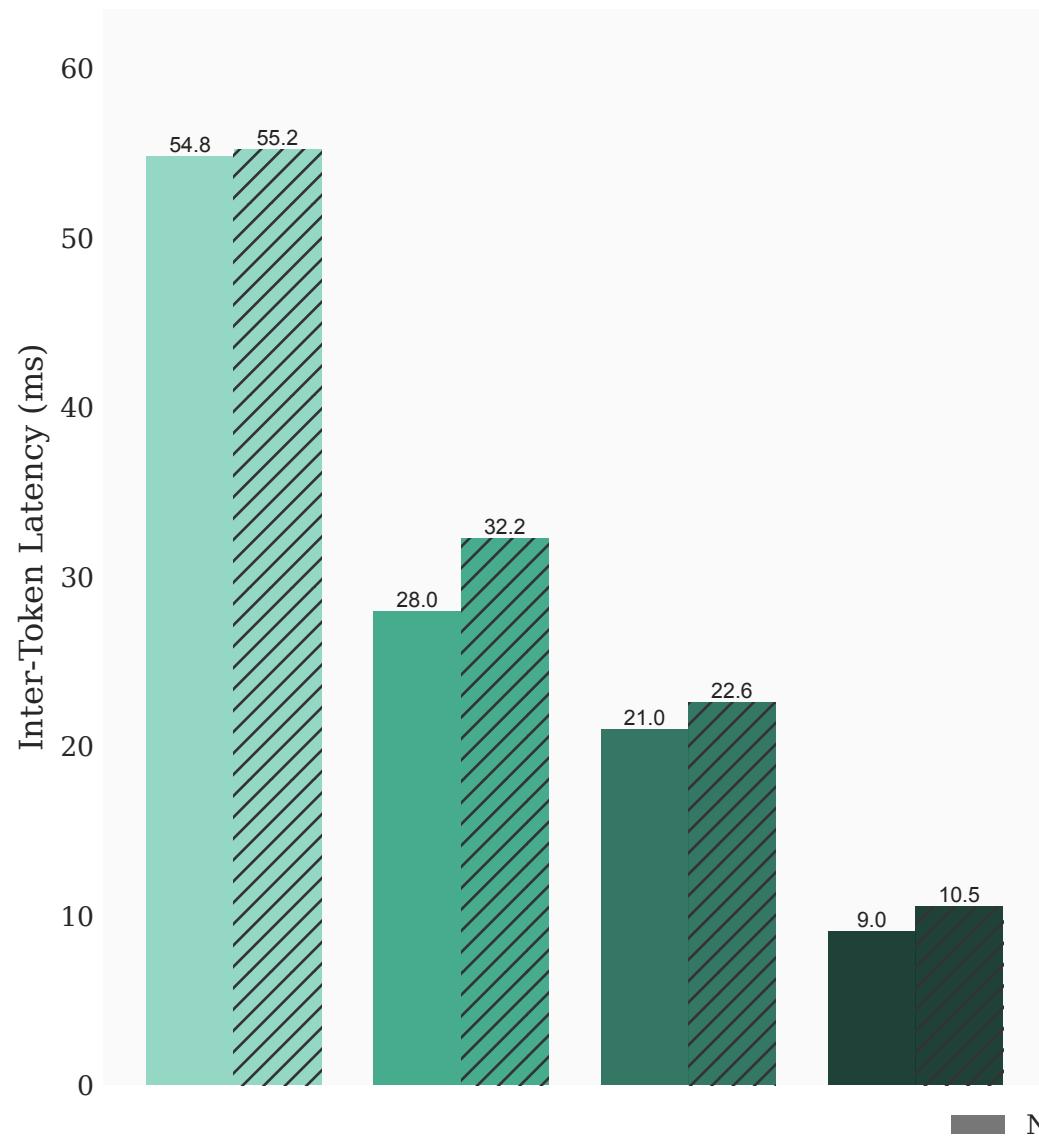


# ShareGPT (Request Rate 100)

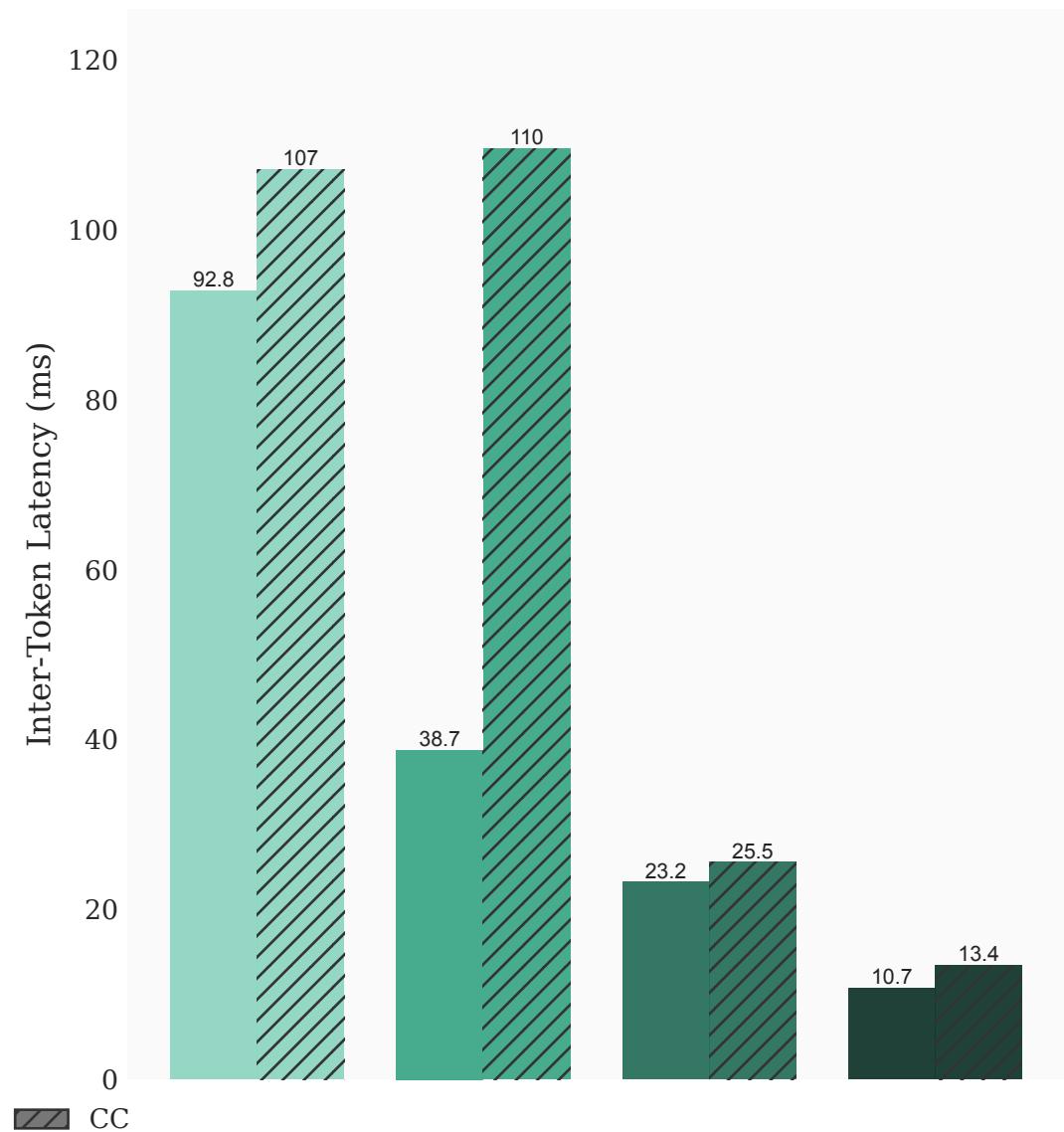


# ShareGPT (Request Rate 50)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

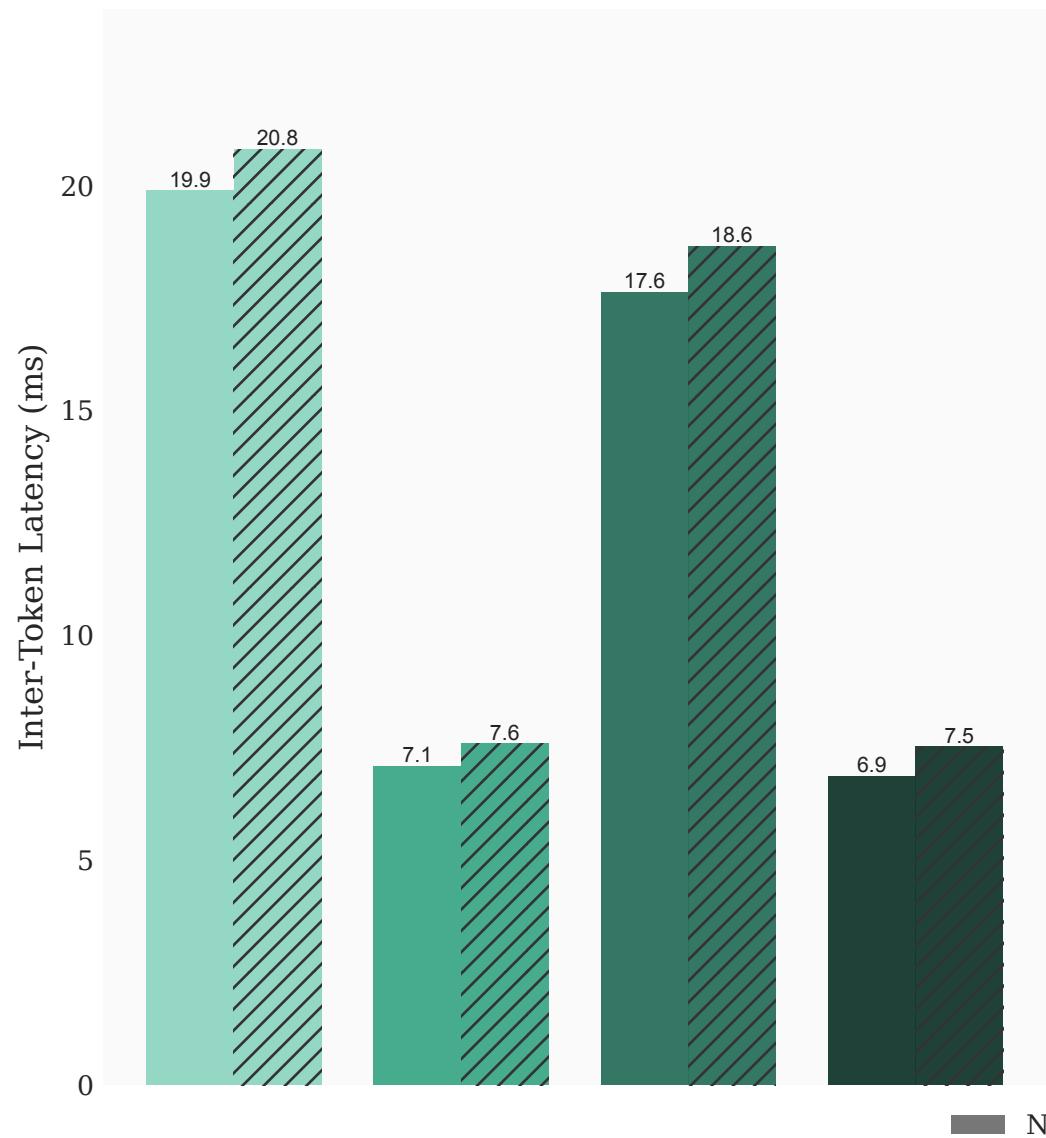
■ GPT OSS 120B

■ Mistral 3.1 24B

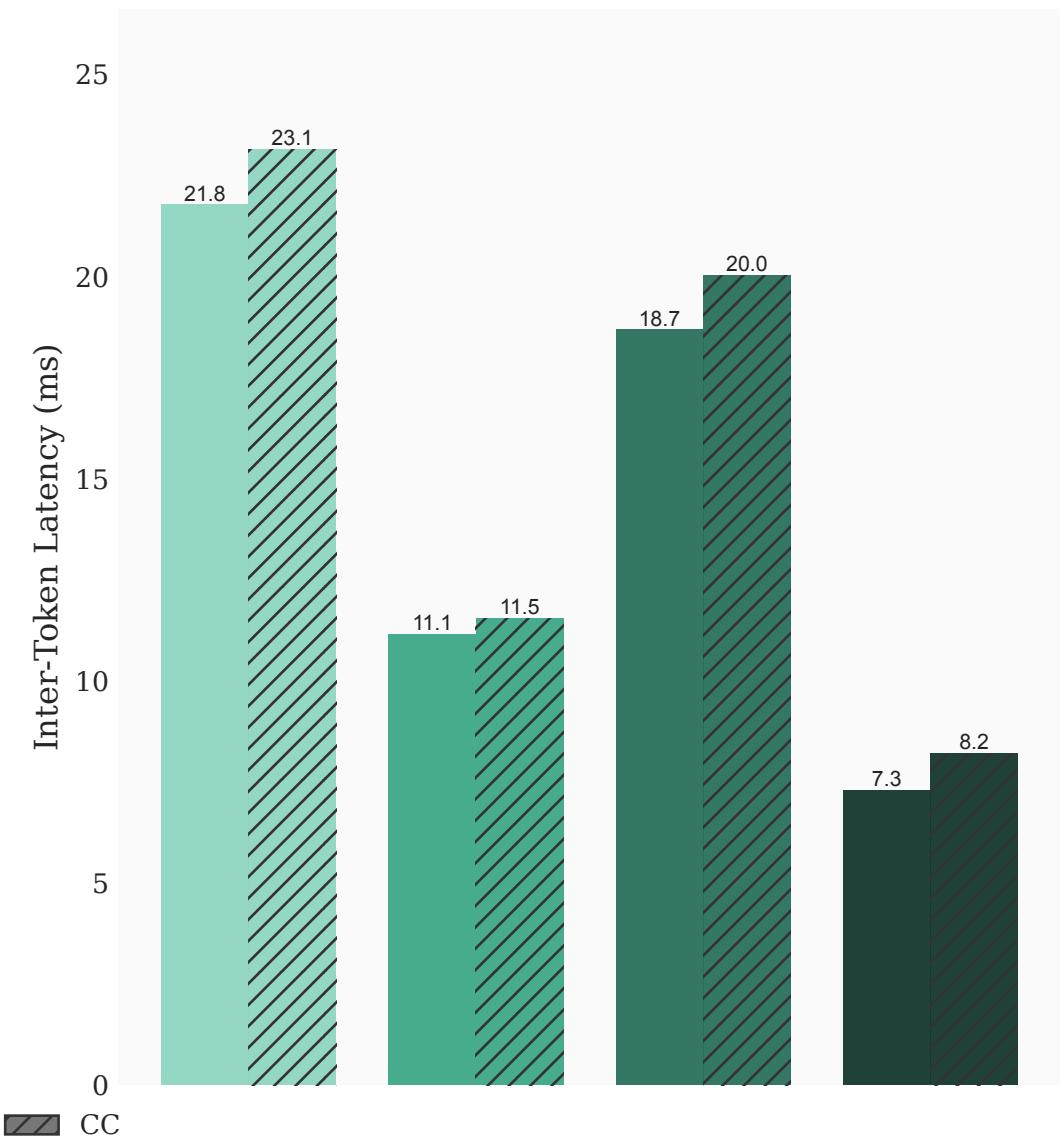
■ LLama 3.1 8B

# ShareGPT (Request Rate 1)

## Mean ITL



## P99 ITL



■ LLama 3.3 70B Int4

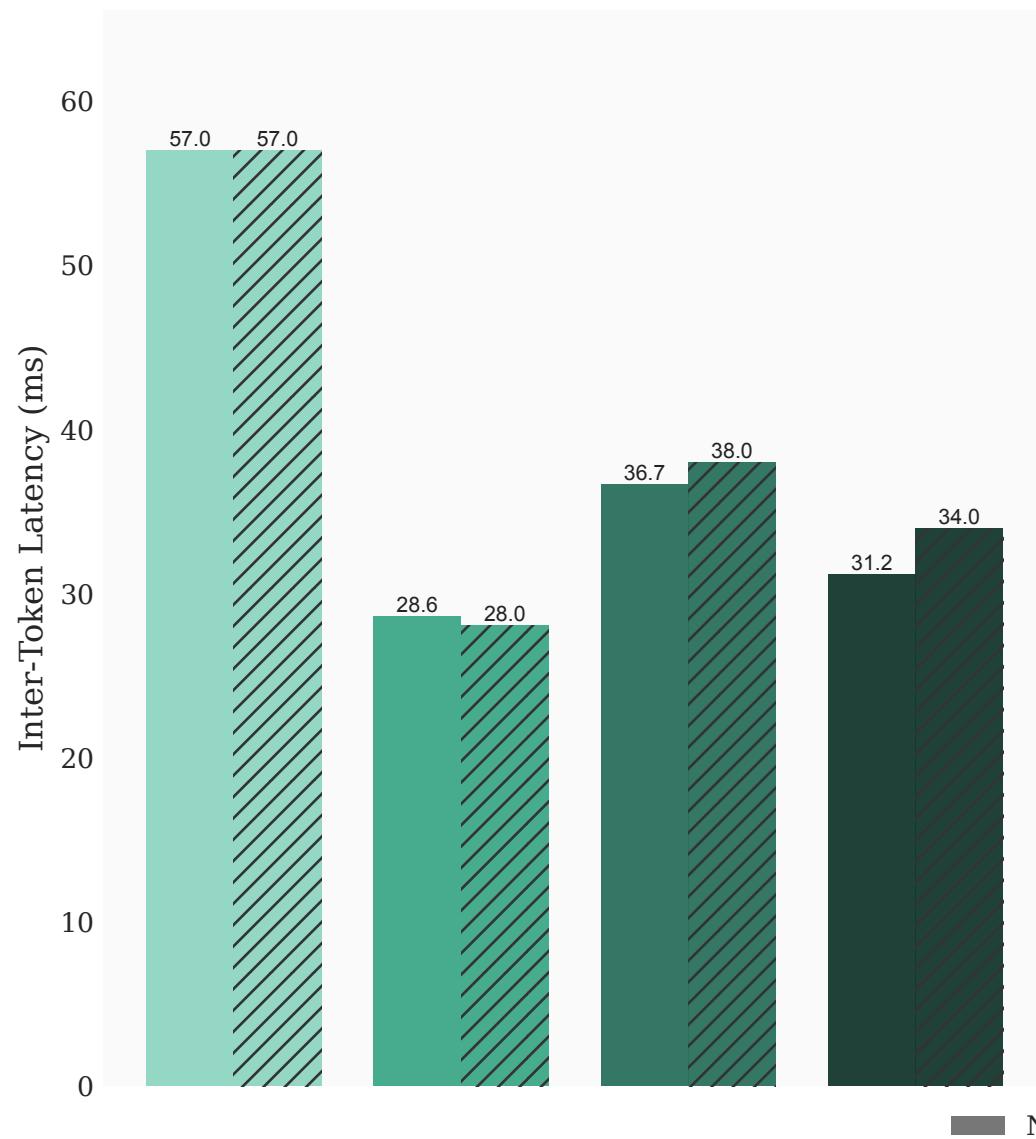
■ GPT OSS 120B

■ Mistral 3.1 24B

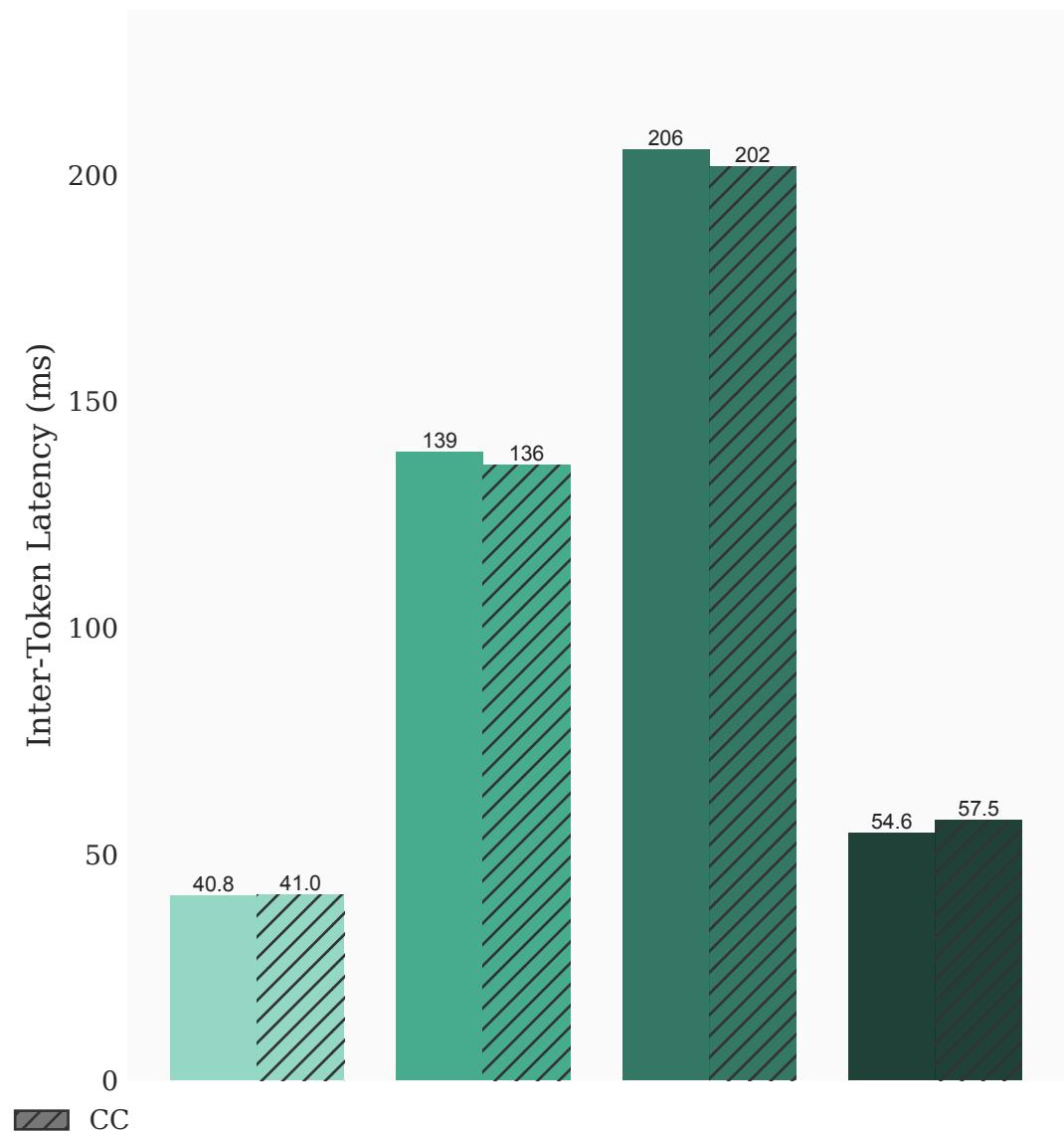
■ LLama 3.1 8B

# Edit 10K Characters (Request Rate 100)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

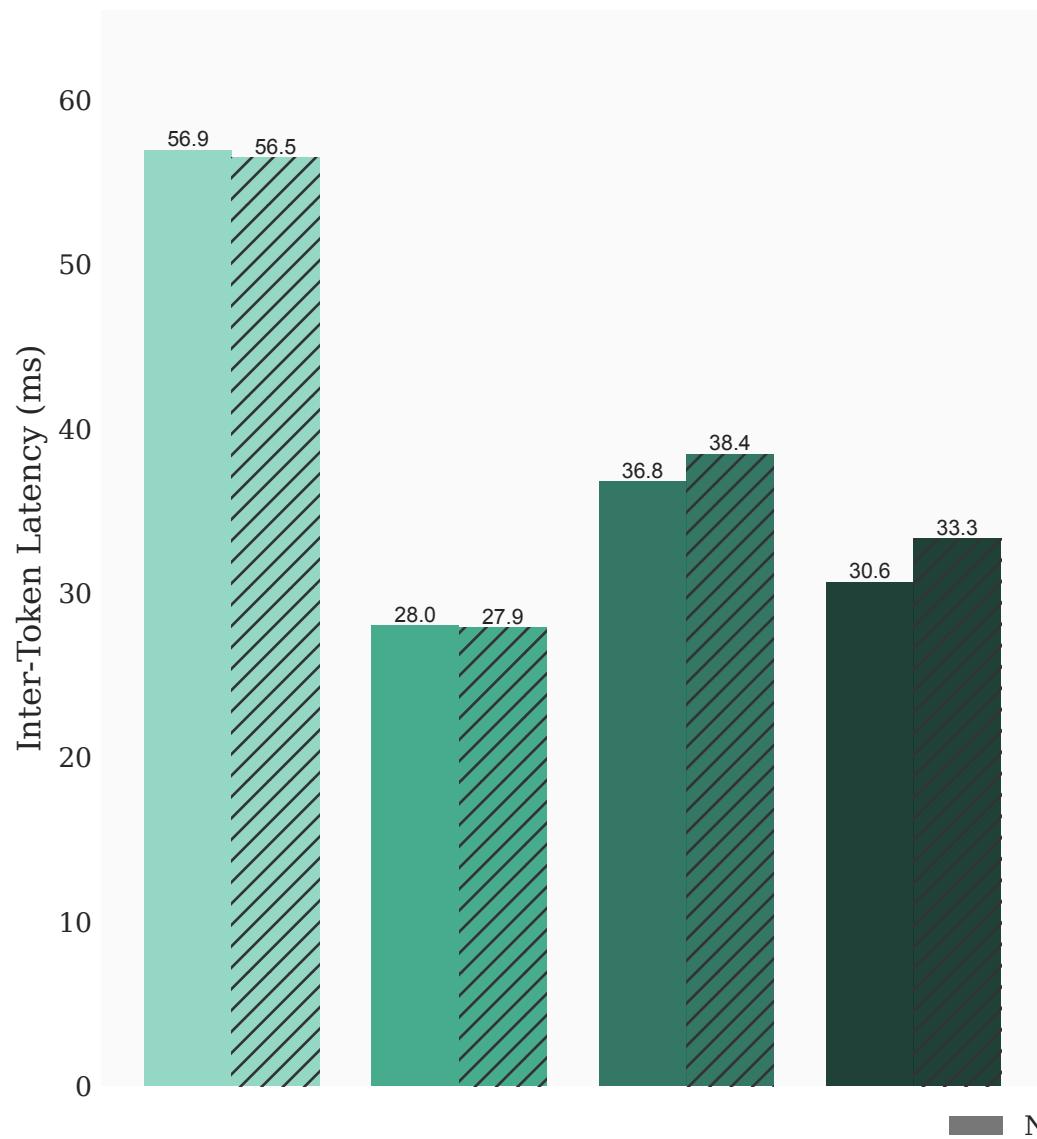
■ GPT OSS 120B

■ Mistral 3.1 24B

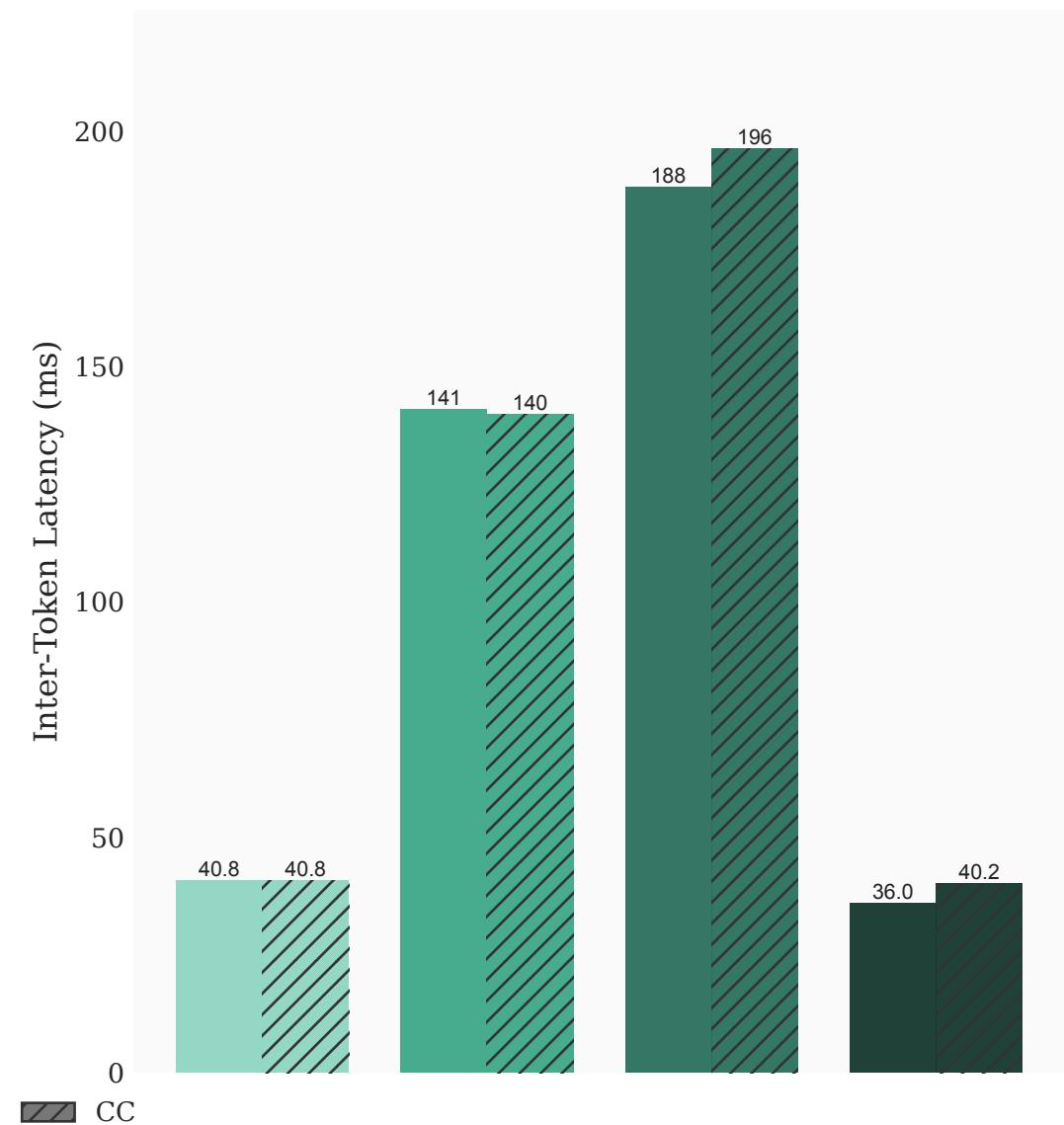
■ LLama 3.1 8B

# Edit 10K Characters (Request Rate 50)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

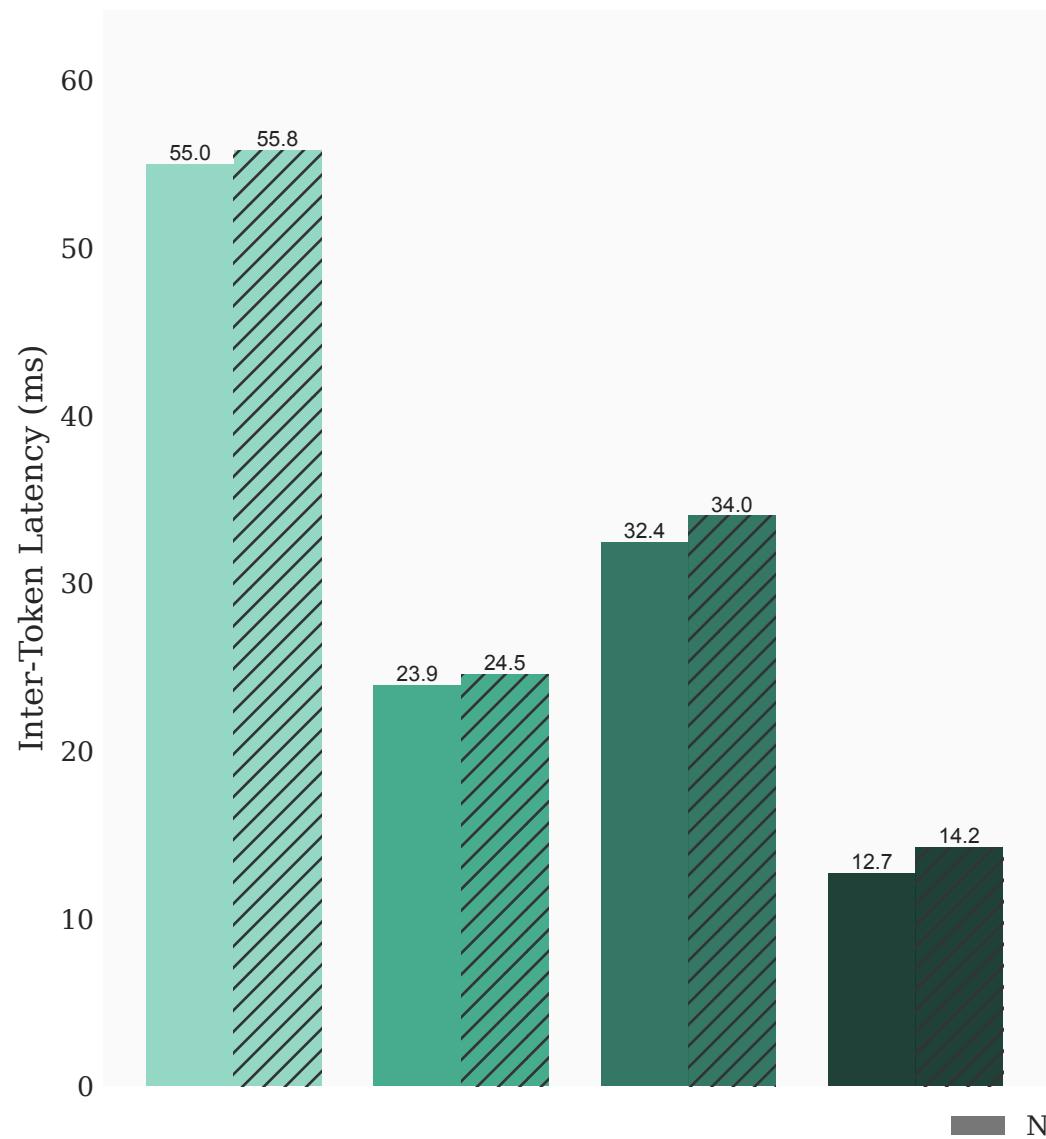
■ GPT OSS 120B

■ Mistral 3.1 24B

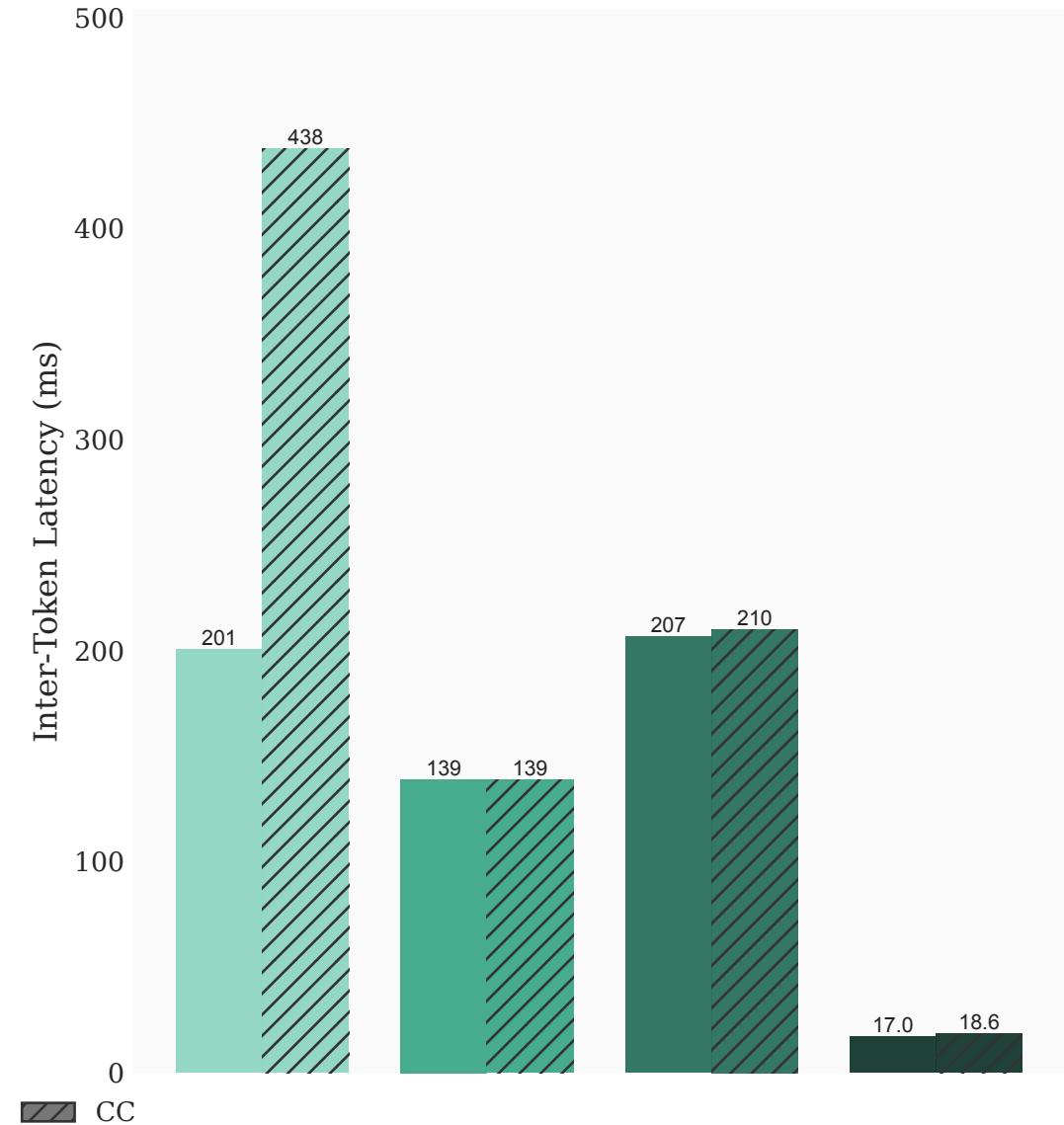
■ LLama 3.1 8B

# Edit 10K Characters (Request Rate 1)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

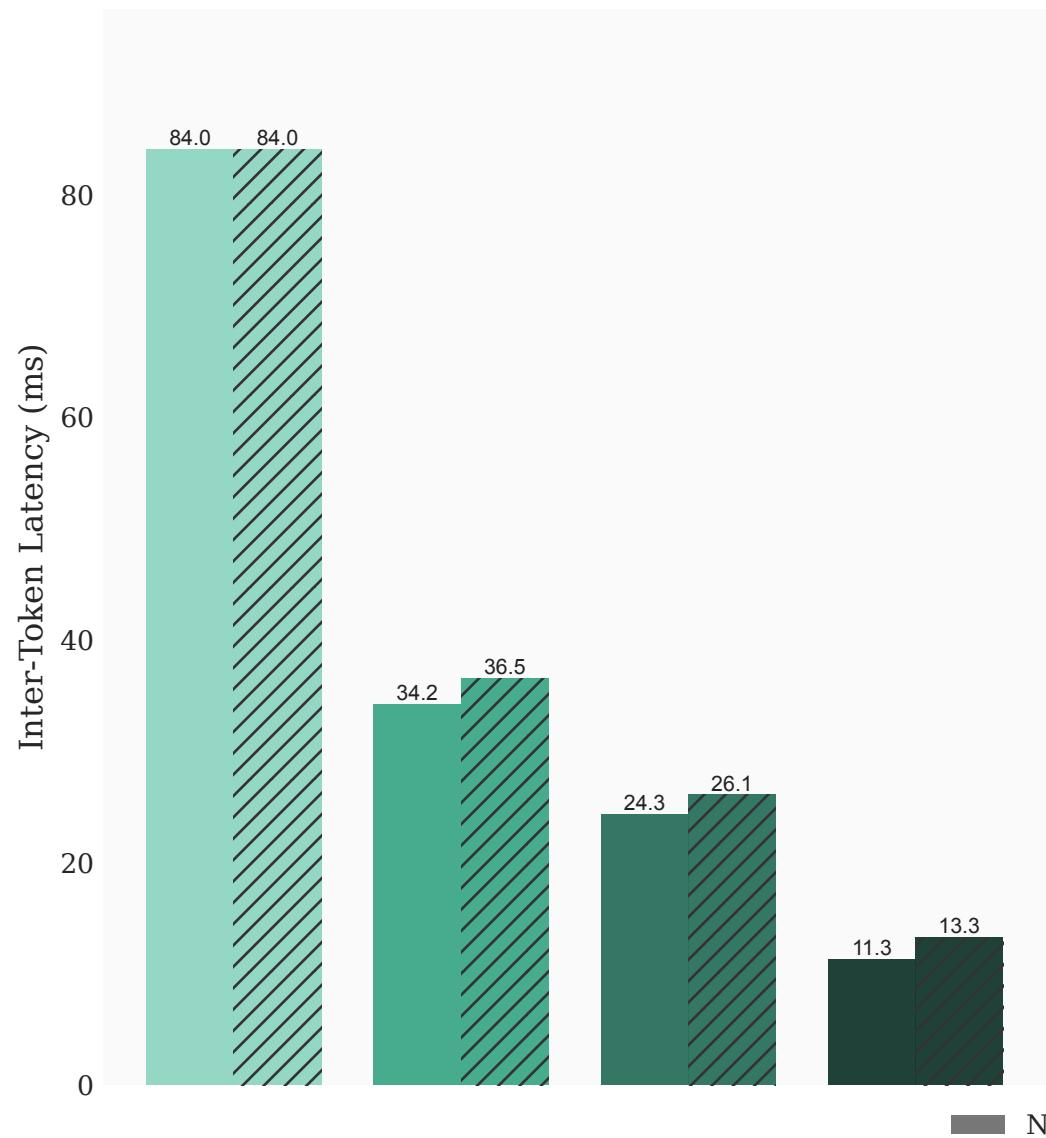
■ GPT OSS 120B

■ Mistral 3.1 24B

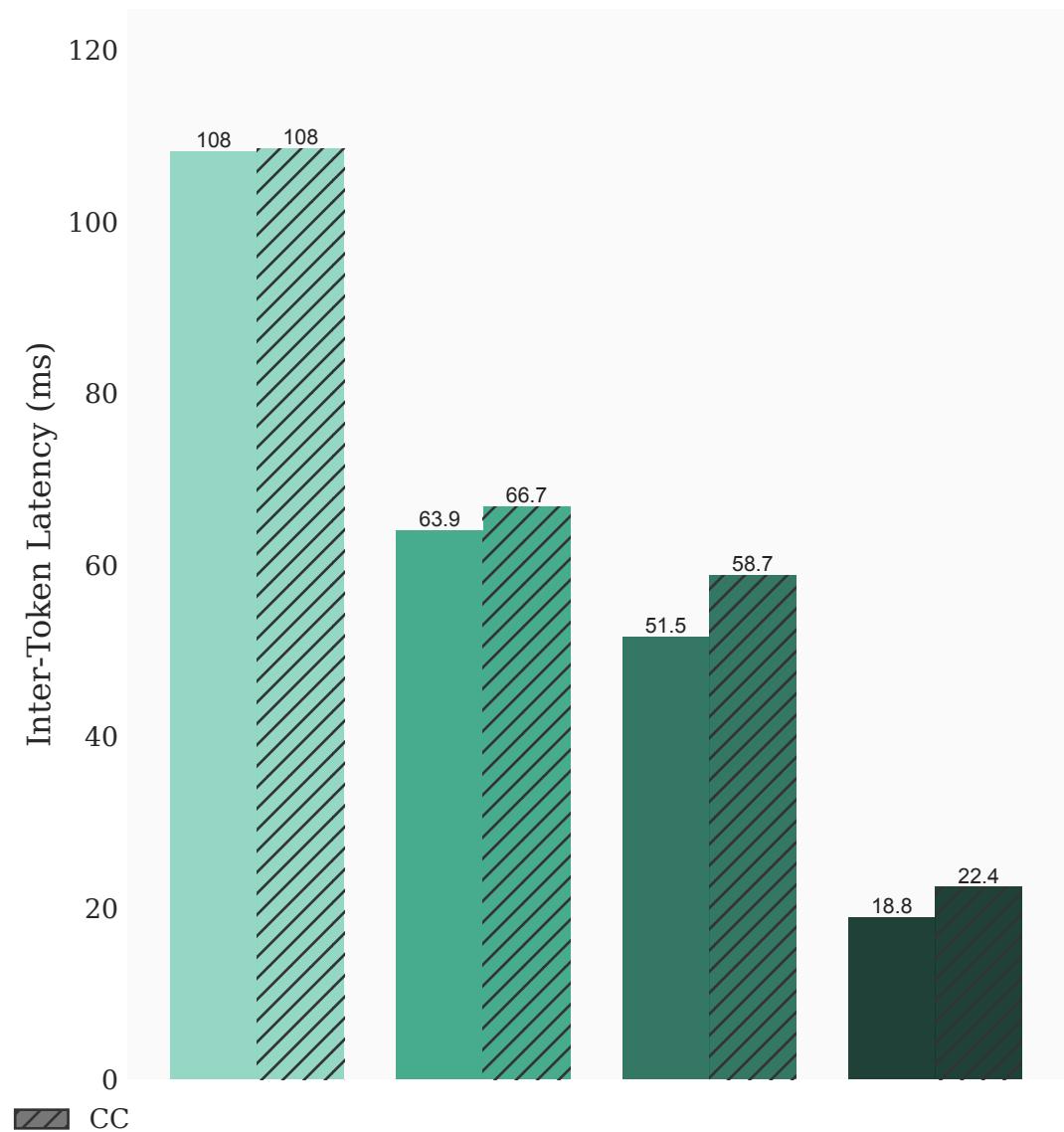
■ LLama 3.1 8B

# Numina Math (Request Rate 100)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

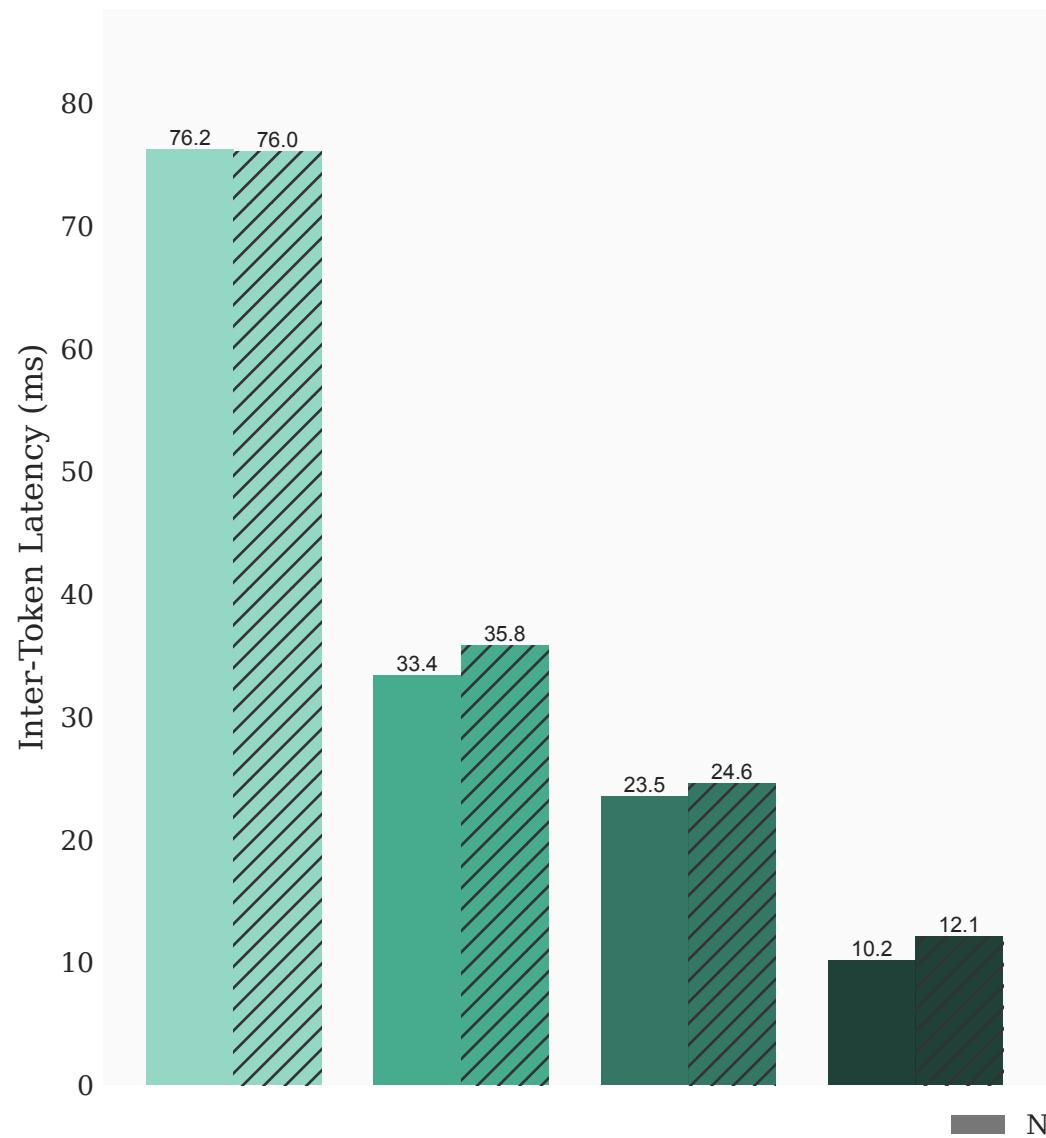
■ GPT OSS 120B

■ Mistral 3.1 24B

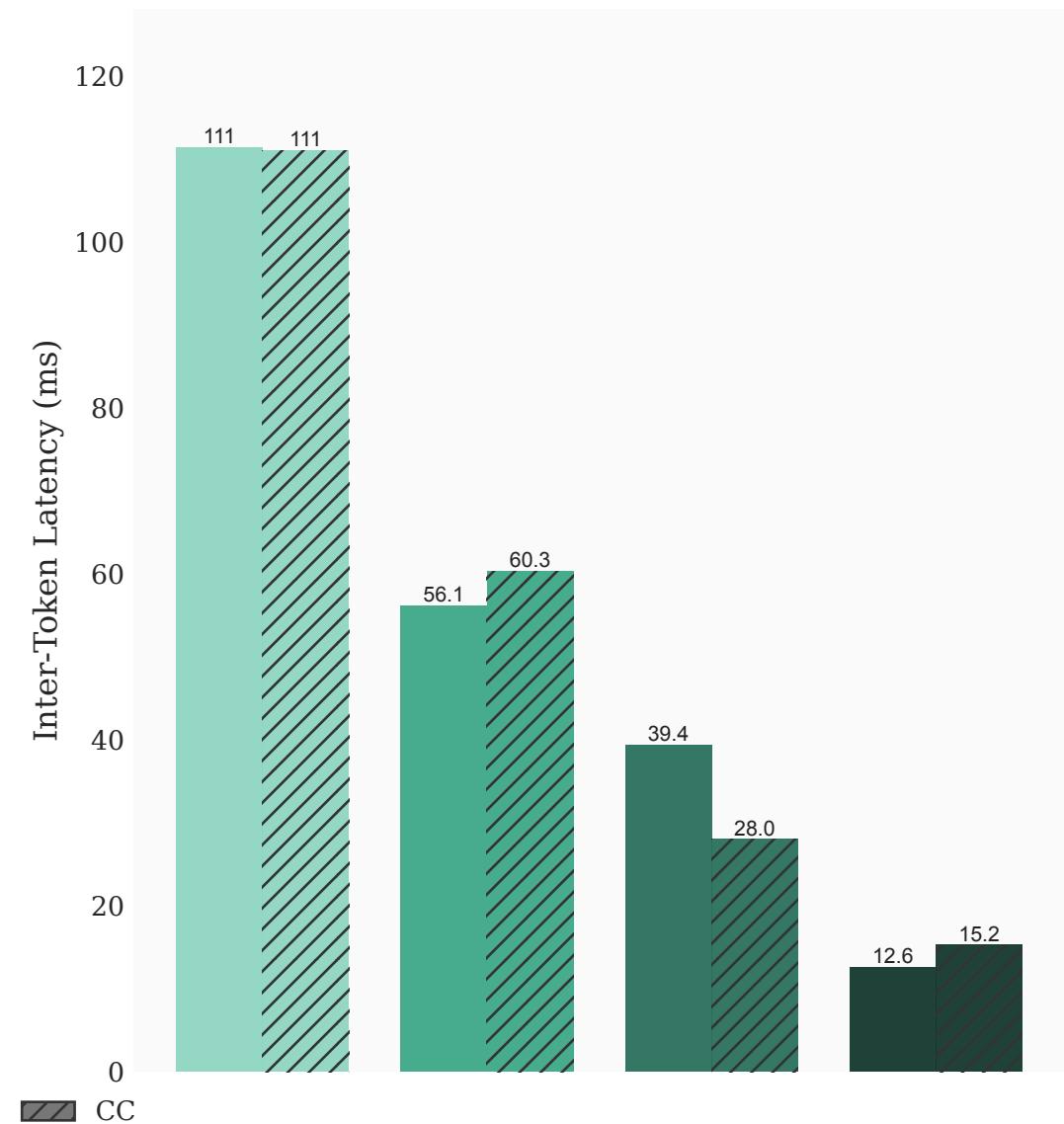
■ LLama 3.1 8B

# Numina Math (Request Rate 50)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

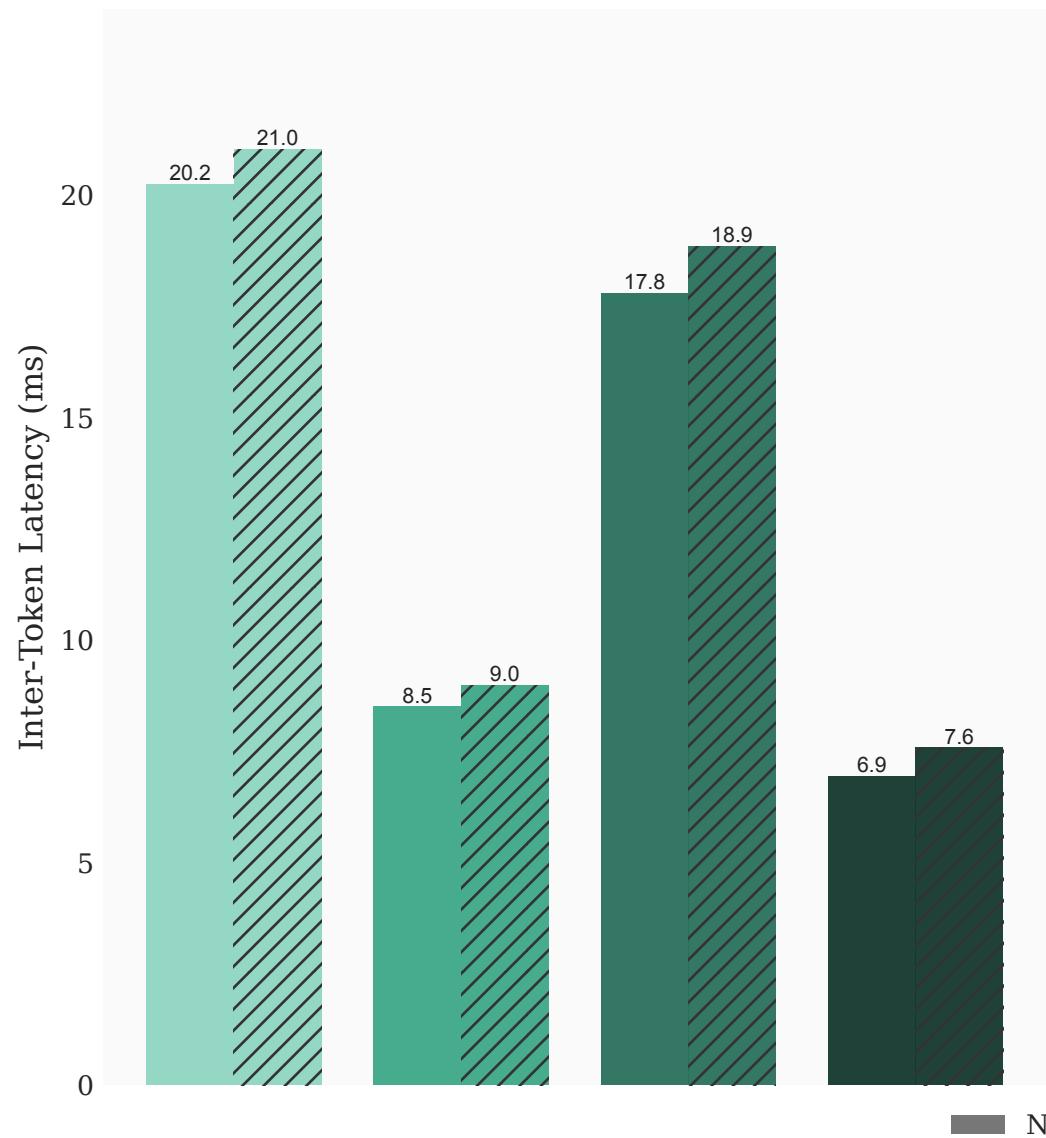
■ GPT OSS 120B

■ Mistral 3.1 24B

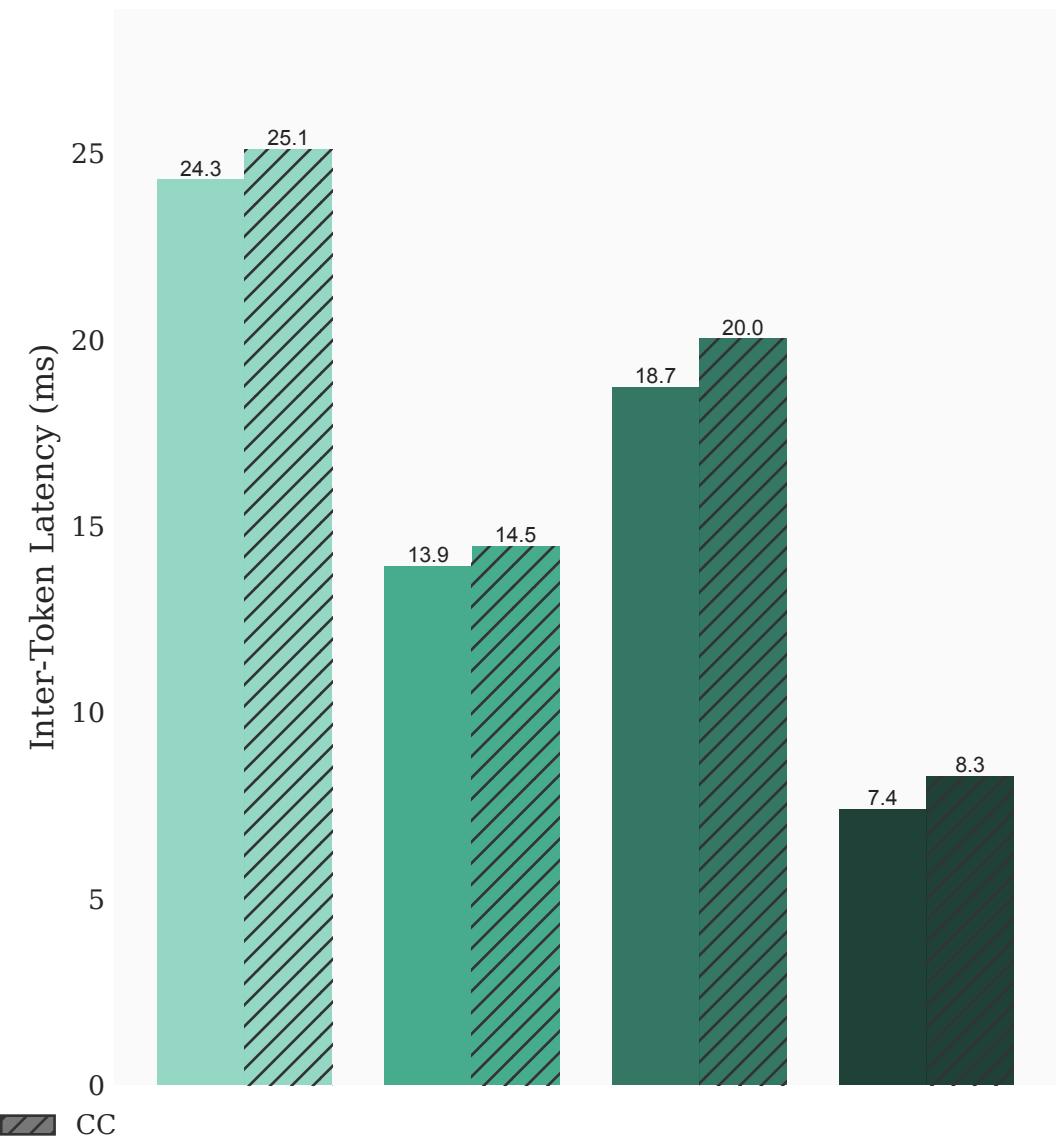
■ LLama 3.1 8B

# Numina Math (Request Rate 1)

Mean ITL



P99 ITL



■ LLama 3.3 70B Int4

■ GPT OSS 120B

■ Mistral 3.1 24B

■ LLama 3.1 8B