Random (1500 ⇒ 250) (100 Concurrent Requests)

Time to First Token (Mean)

Y-axis: TTFT (ms)

Bar labels: 2.8s, 2.8s, 13.5s, 13.1s, 5.7s, 5.5s, 75.3s, 75.9s

Time to First Token (P99)

Y-axis: TTFT (ms)

Bar labels: 5.5s, 5.5s, 28.3s, 27.7s, 11.8s, 10.9s, 152.5s, 155.6s

Legend: ▨ CC   ▬ No CC

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

Random (1500 ⇒ 250) (50 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)
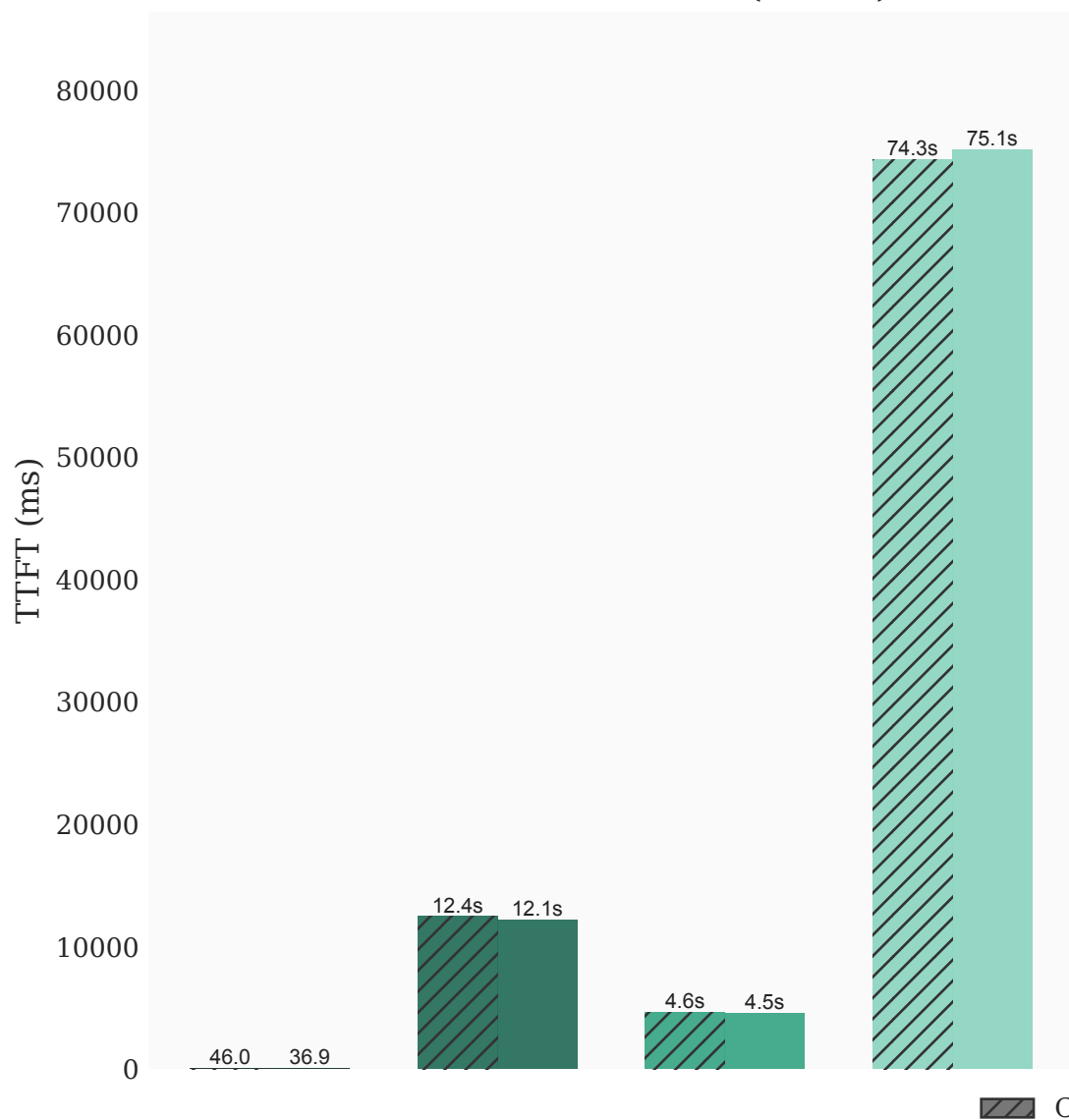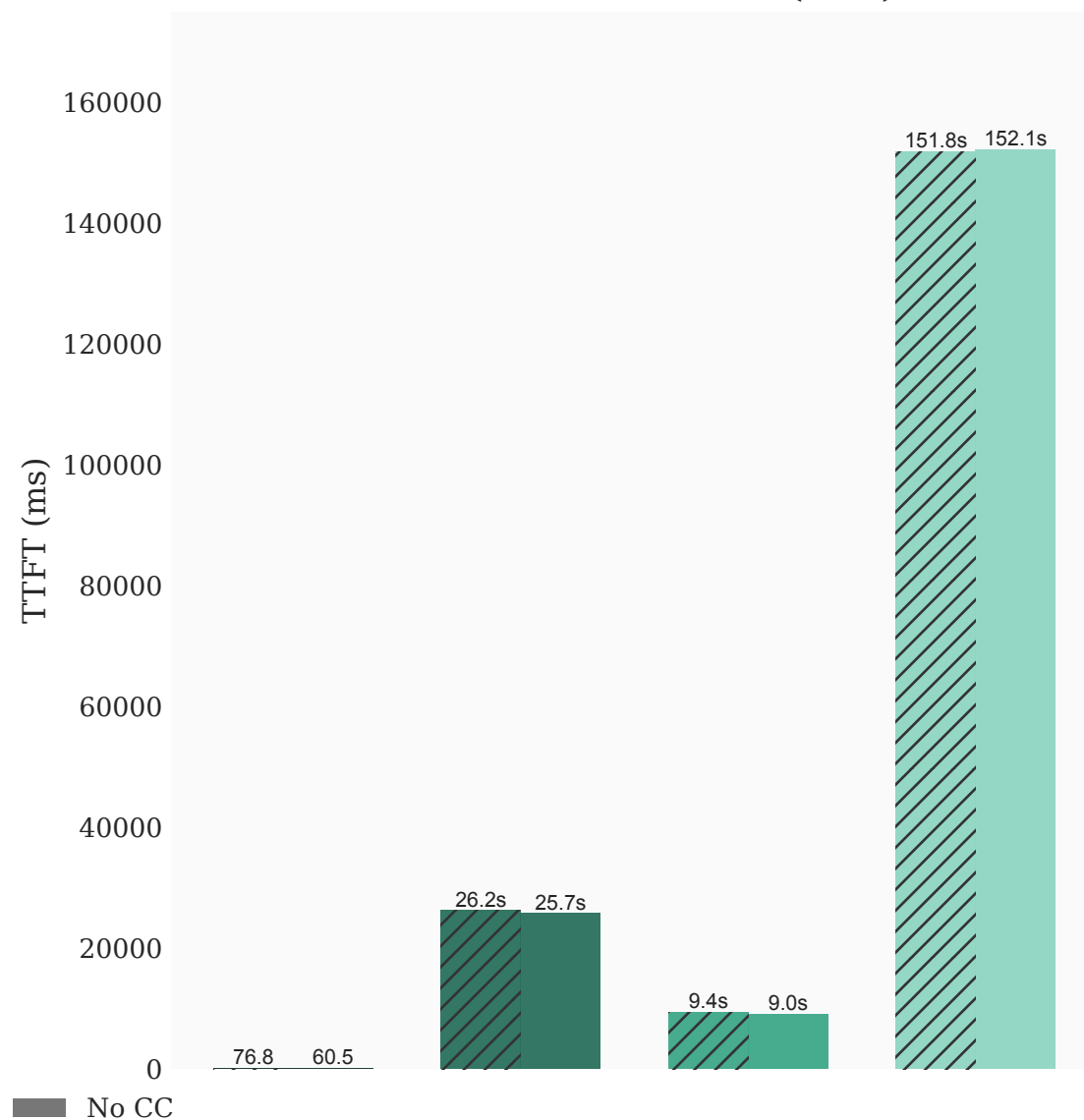
- 46.0   36.9
- 12.4s   12.1s
- 4.6s   4.5s
- 74.3s   75.1s

Time to First Token (P99)

TTFT (ms)

- 76.8   60.5
- 26.2s   25.7s
- 9.4s   9.0s
- 151.8s 152.1s

⧄ CC   ▬ No CC

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

Random (1500 ⇒ 250) (1 Concurrent Requests)

Time to First Token (Mean)
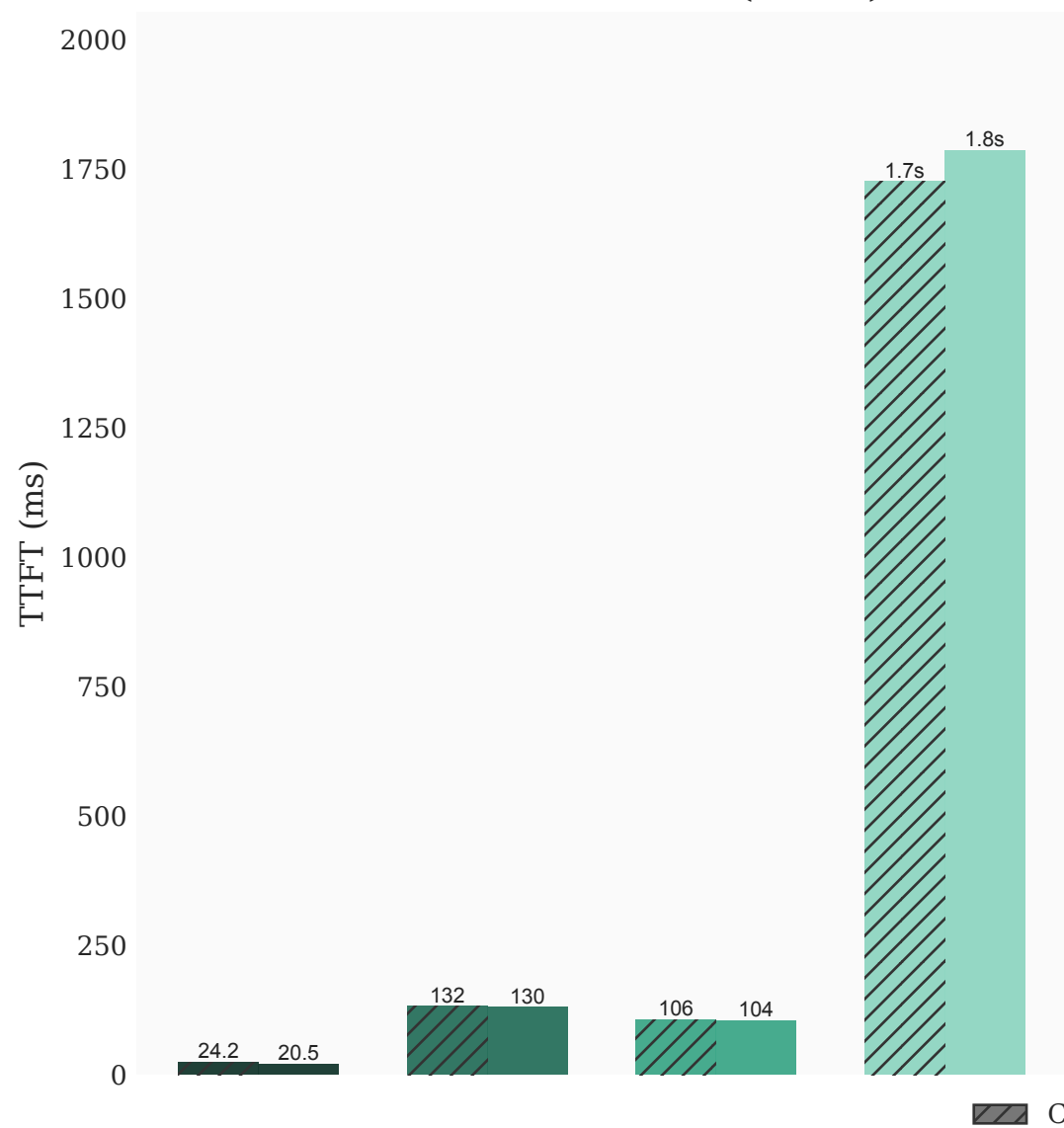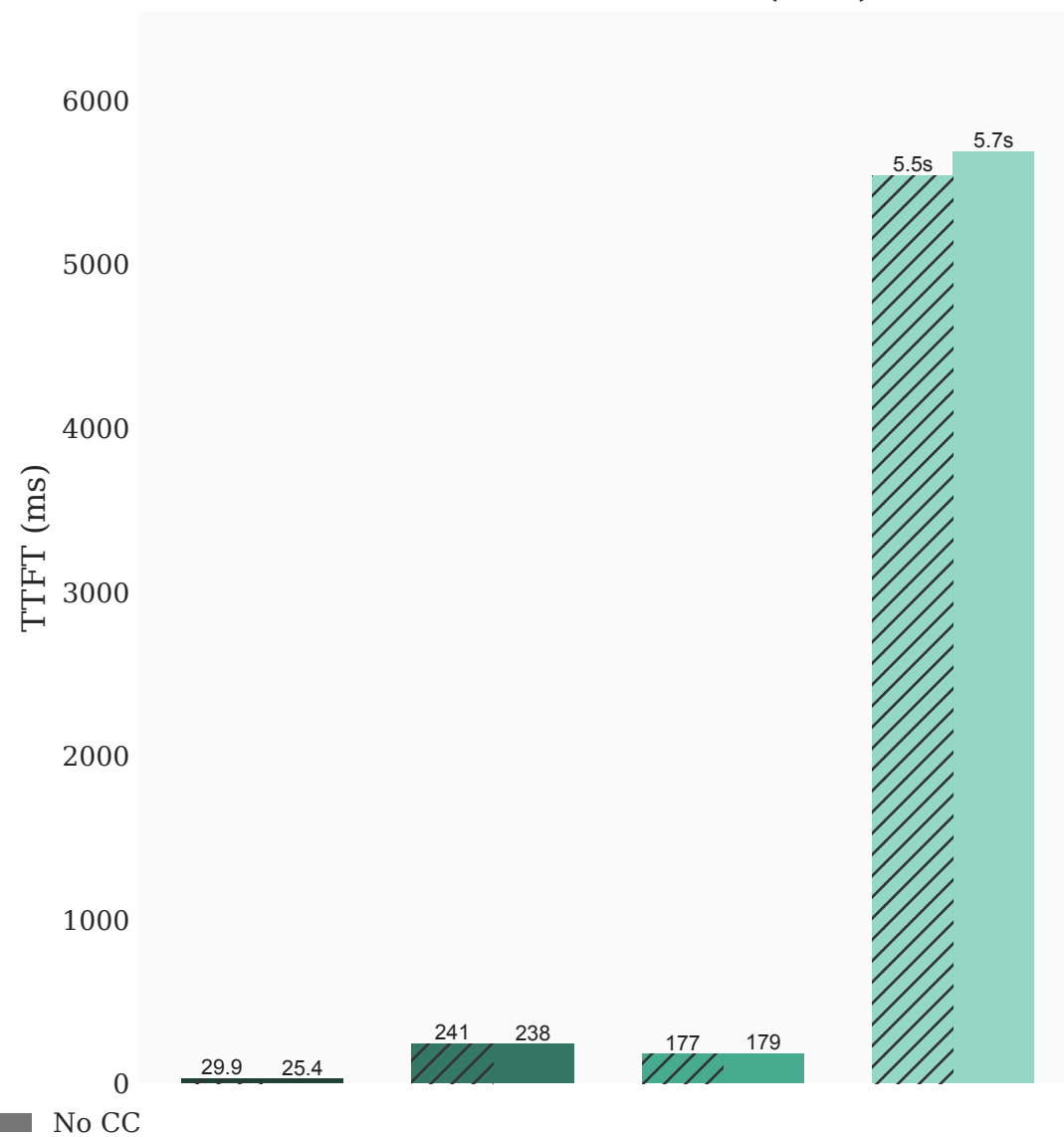
TTFT (ms)

- 24.2 | 20.5 — LLama 3.1 8B
- 132 | 130 — Mistral 3.1 24B
- 106 | 104 — GPT OSS 120B
- 1.7s | 1.8s — LLama 3.3 70B Int4

Time to First Token (P99)

TTFT (ms)

- 29.9 | 25.4 — LLama 3.1 8B
- 241 | 238 — Mistral 3.1 24B
- 177 | 179 — GPT OSS 120B
- 5.5s | 5.7s — LLama 3.3 70B Int4
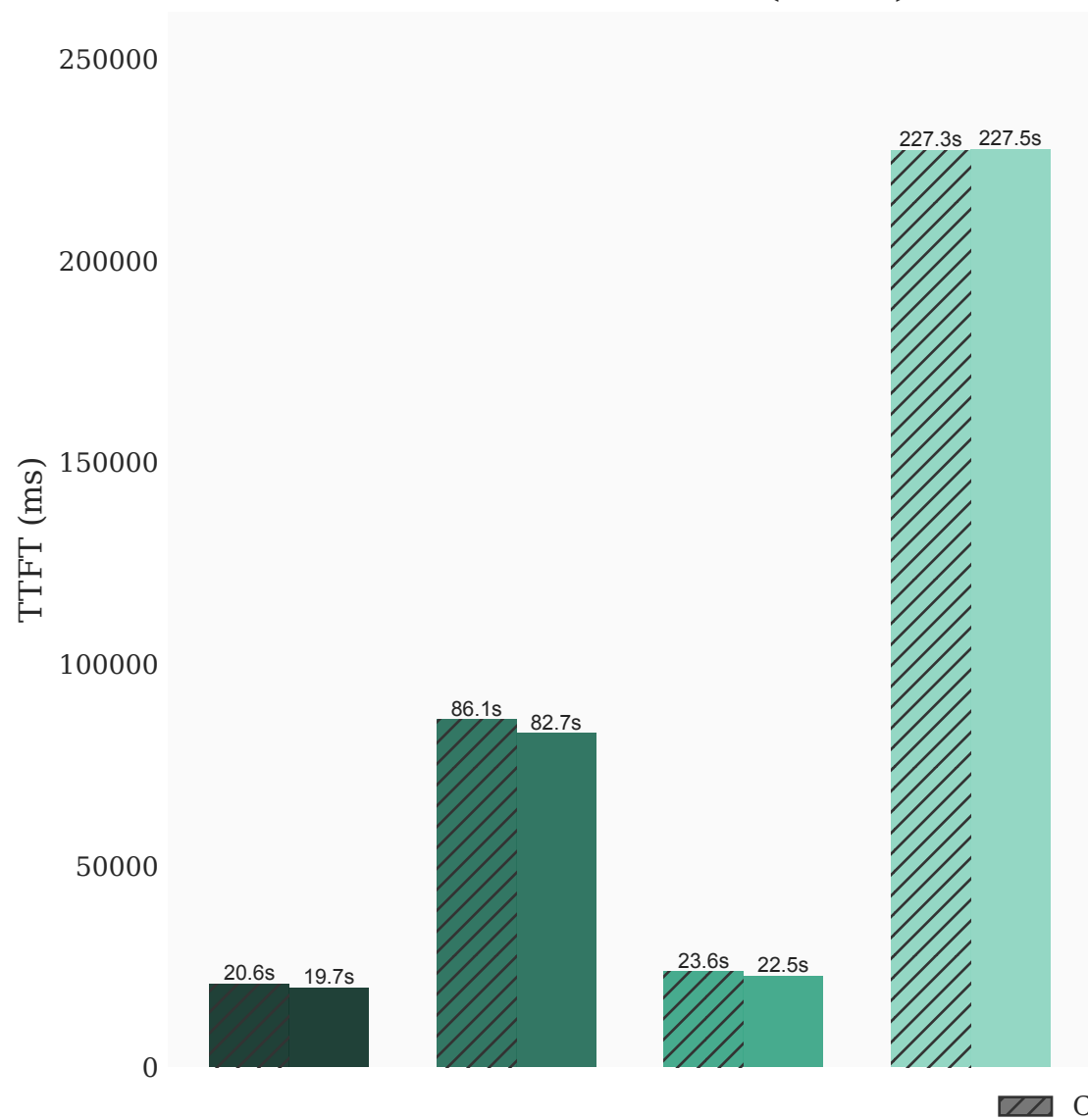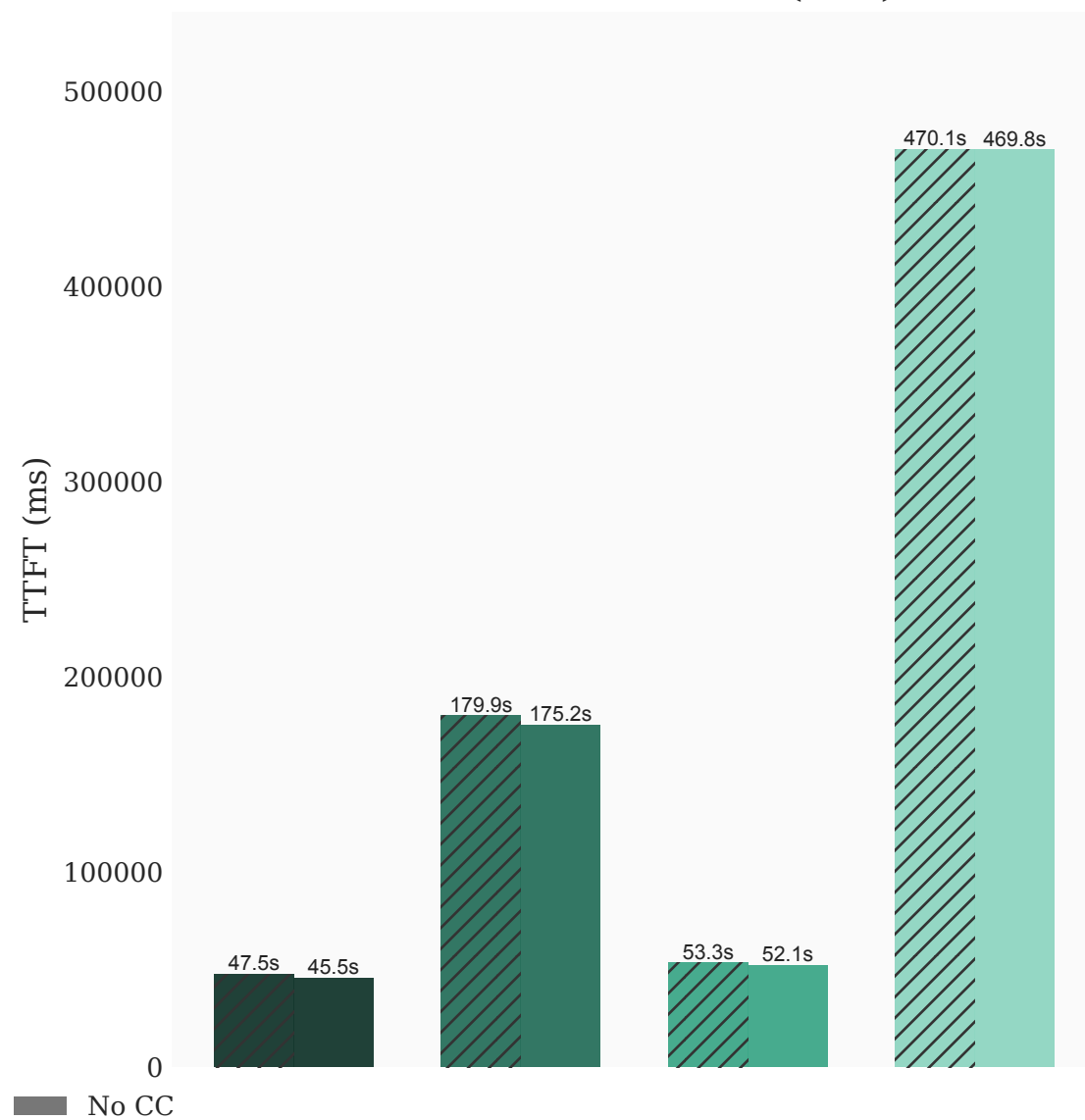
CC   No CC

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

Random (4000 ⇒ 1000) (100 Concurrent Requests)

Time to First Token (Mean)

Time to First Token (P99)

Time to First Token (Mean):
- LLama 3.1 8B: 20.6s (CC), 19.7s (No CC)
- Mistral 3.1 24B: 86.1s (CC), 82.7s (No CC)
- GPT OSS 120B: 23.6s (CC), 22.5s (No CC)
- LLama 3.3 70B Int4: 227.3s (CC), 227.5s (No CC)

Time to First Token (P99):
- LLama 3.1 8B: 47.5s (CC), 45.5s (No CC)
- Mistral 3.1 24B: 179.9s (CC), 175.2s (No CC)
- GPT OSS 120B: 53.3s (CC), 52.1s (No CC)
- LLama 3.3 70B Int4: 470.1s (CC), 469.8s (No CC)

Legend: CC, No CC

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

Random (4000 ⇒ 1000) (50 Concurrent Requests)

Time to First Token (Mean)

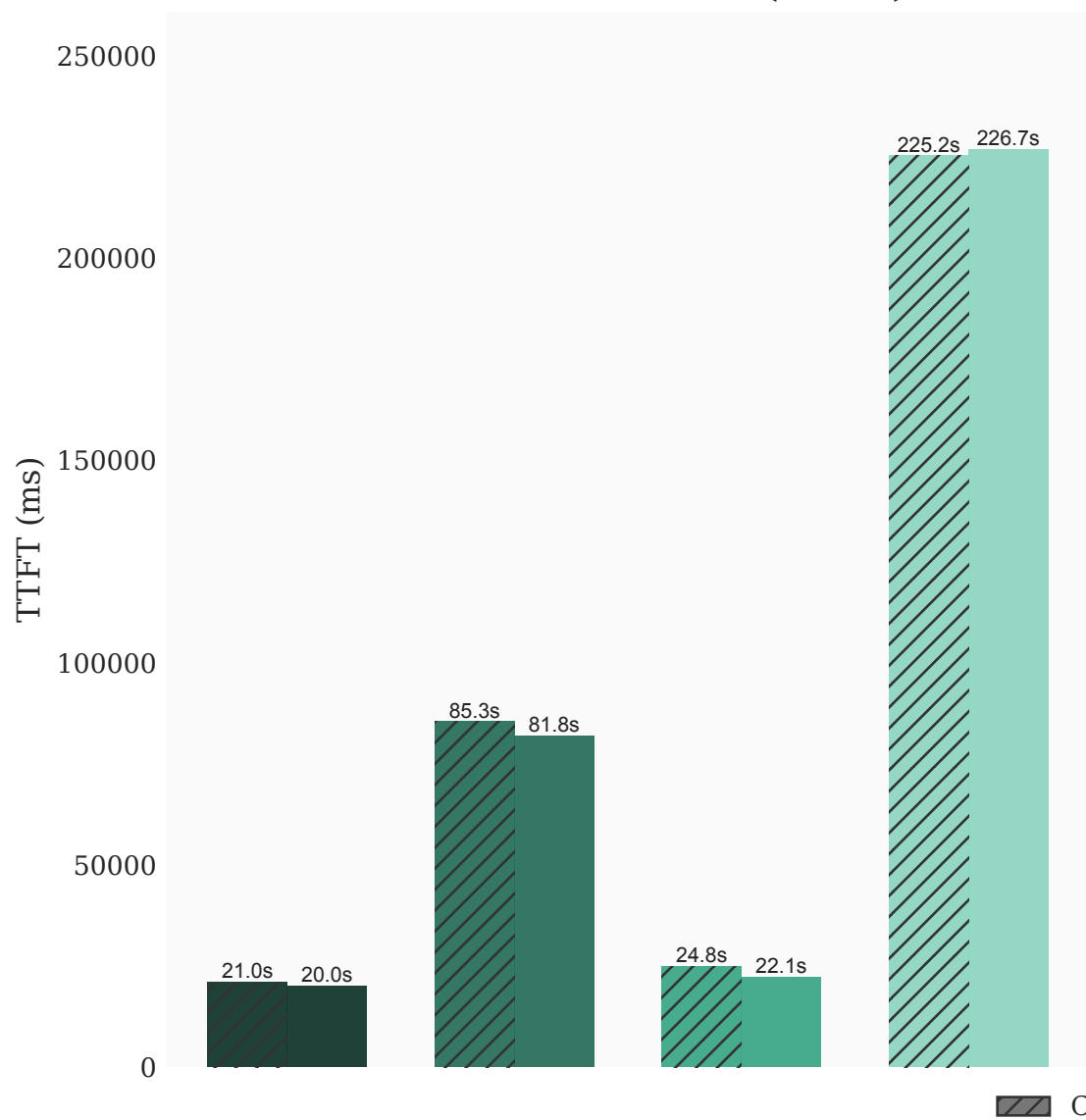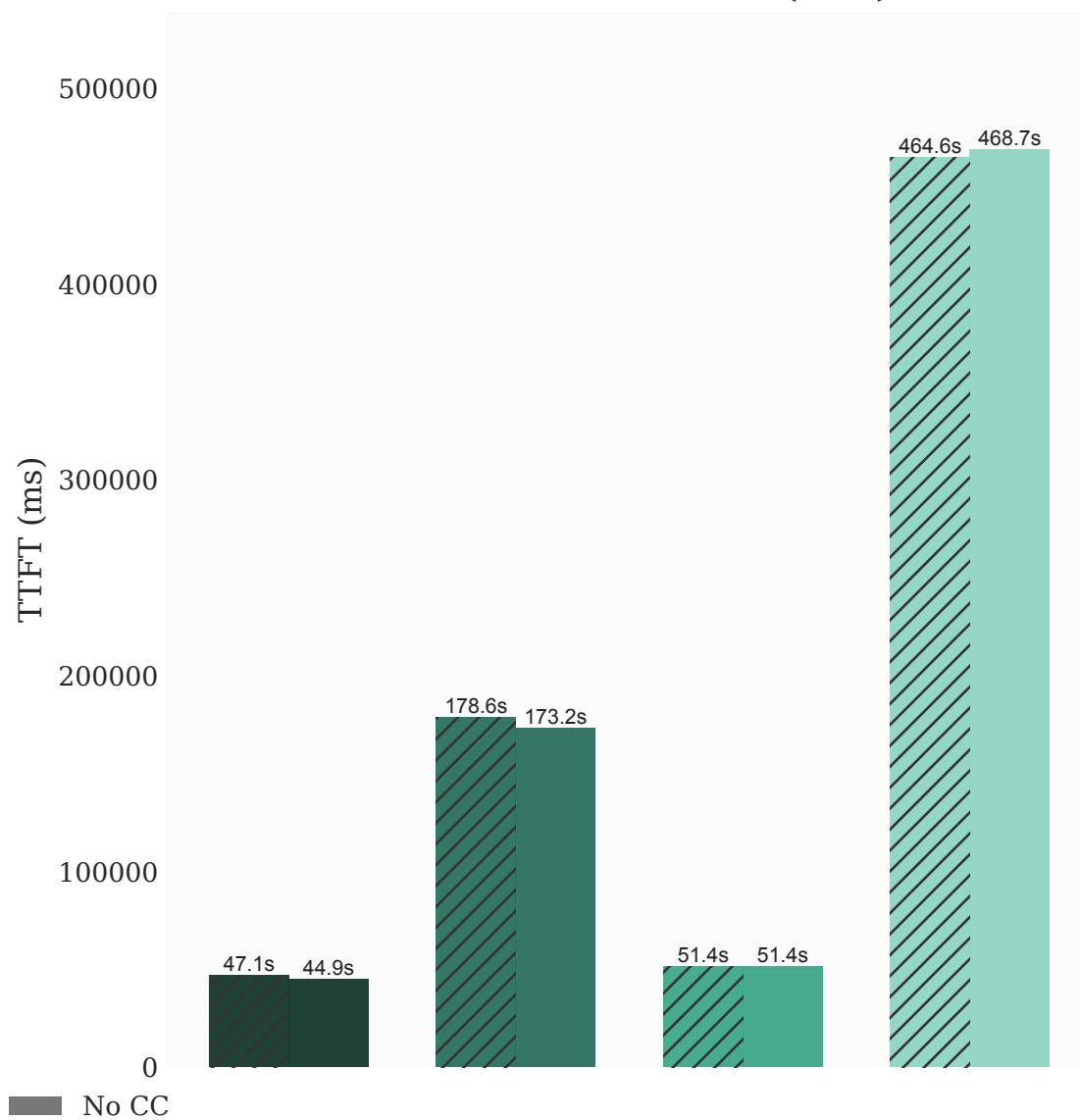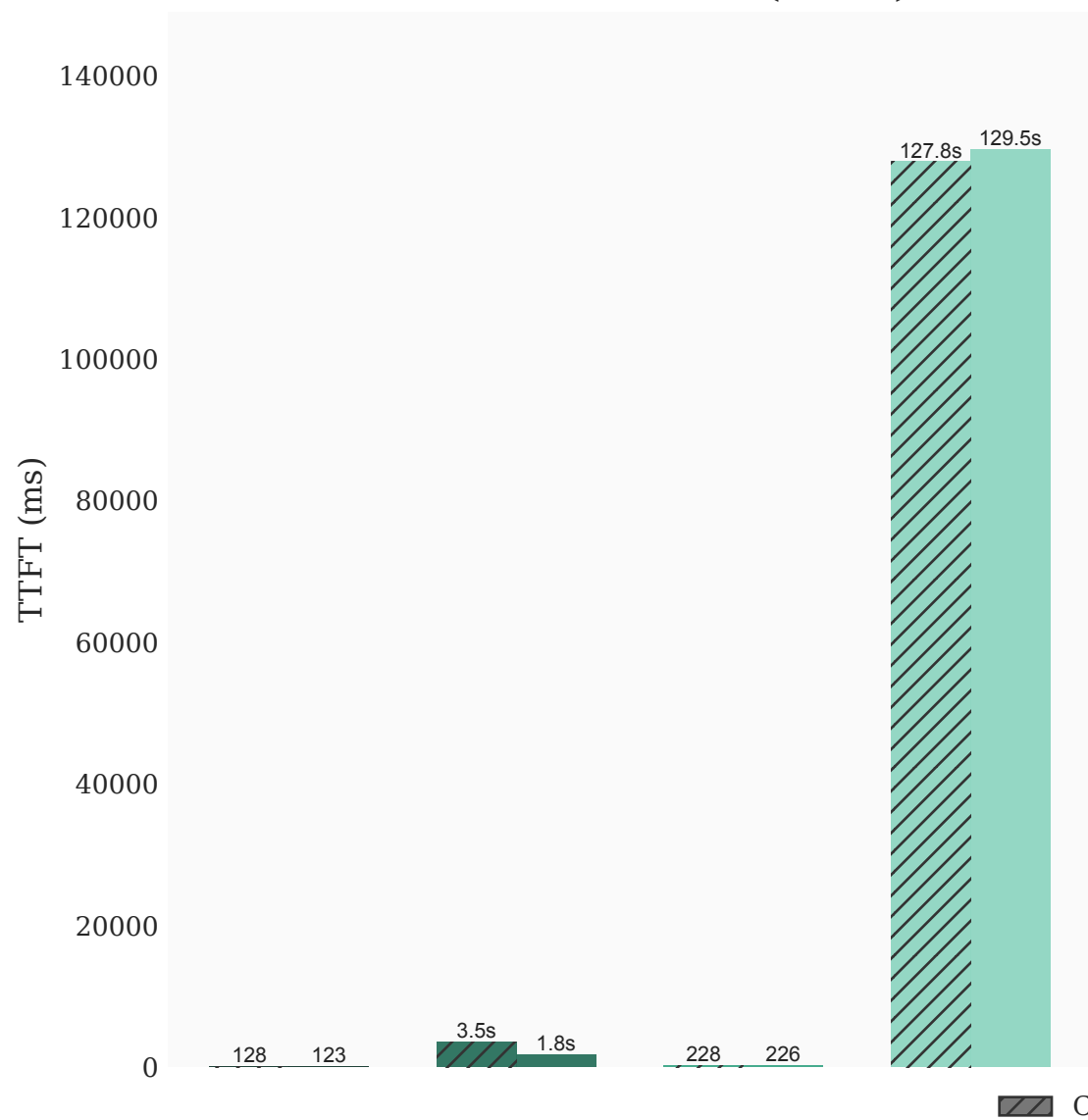TTFT (ms)

21.0s 20.0s — LLama 3.1 8B
85.3s 81.8s — Mistral 3.1 24B
24.8s 22.1s — GPT OSS 120B
225.2s 226.7s — LLama 3.3 70B Int4

Time to First Token (P99)

TTFT (ms)

47.1s 44.9s
178.6s 173.2s
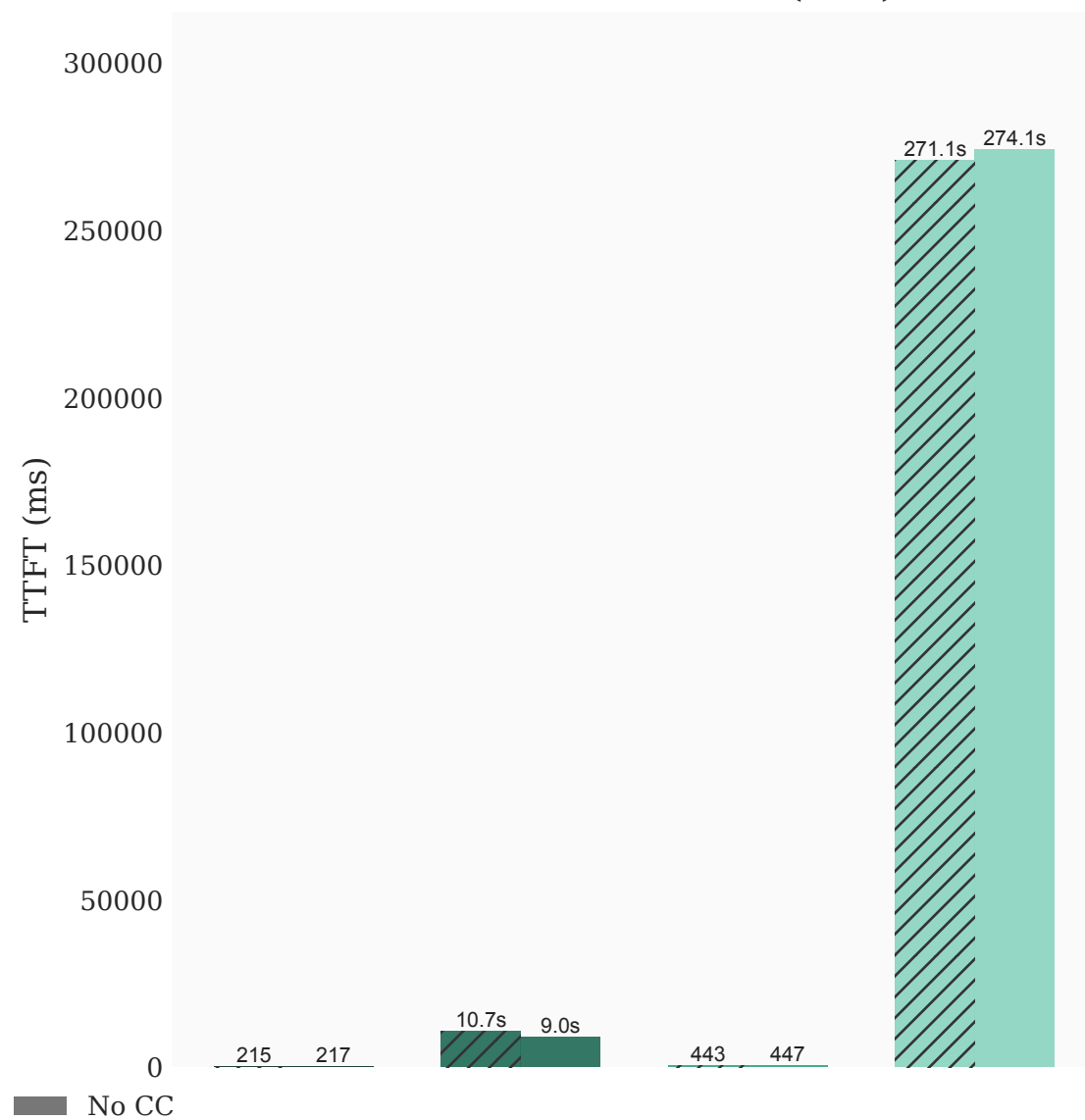51.4s 51.4s
464.6s 468.7s

Legend: ▨ CC ▧ No CC

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

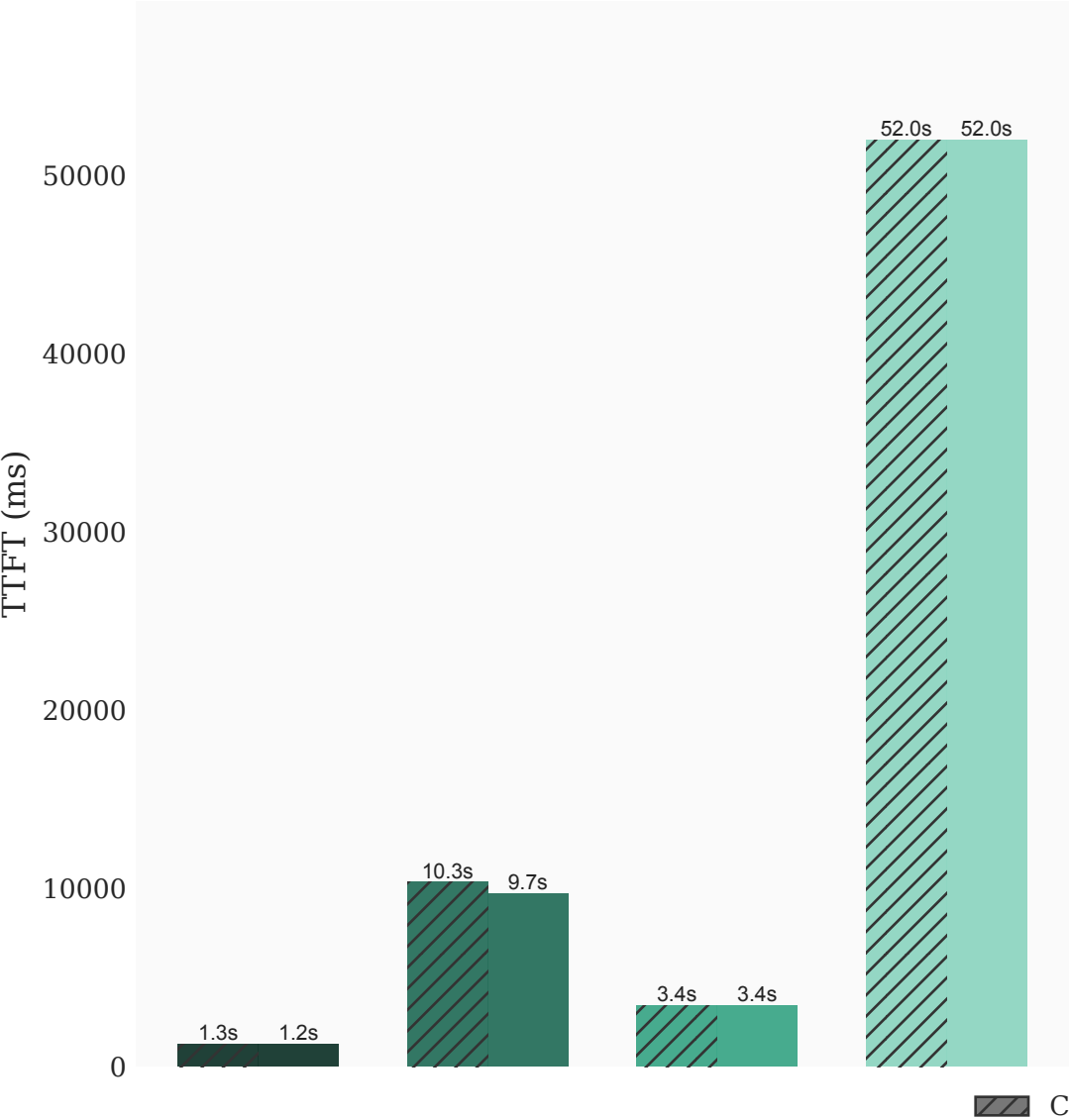# Random (4000 ⇒ 1000) (1 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

- 128
- 123
- 3.5s
- 1.8s
- 228
- 226
- 127.8s
- 129.5s

Time to First Token (P99)

TTFT (ms)

- 215
- 217
- 10.7s
- 9.0s
- 443
- 447
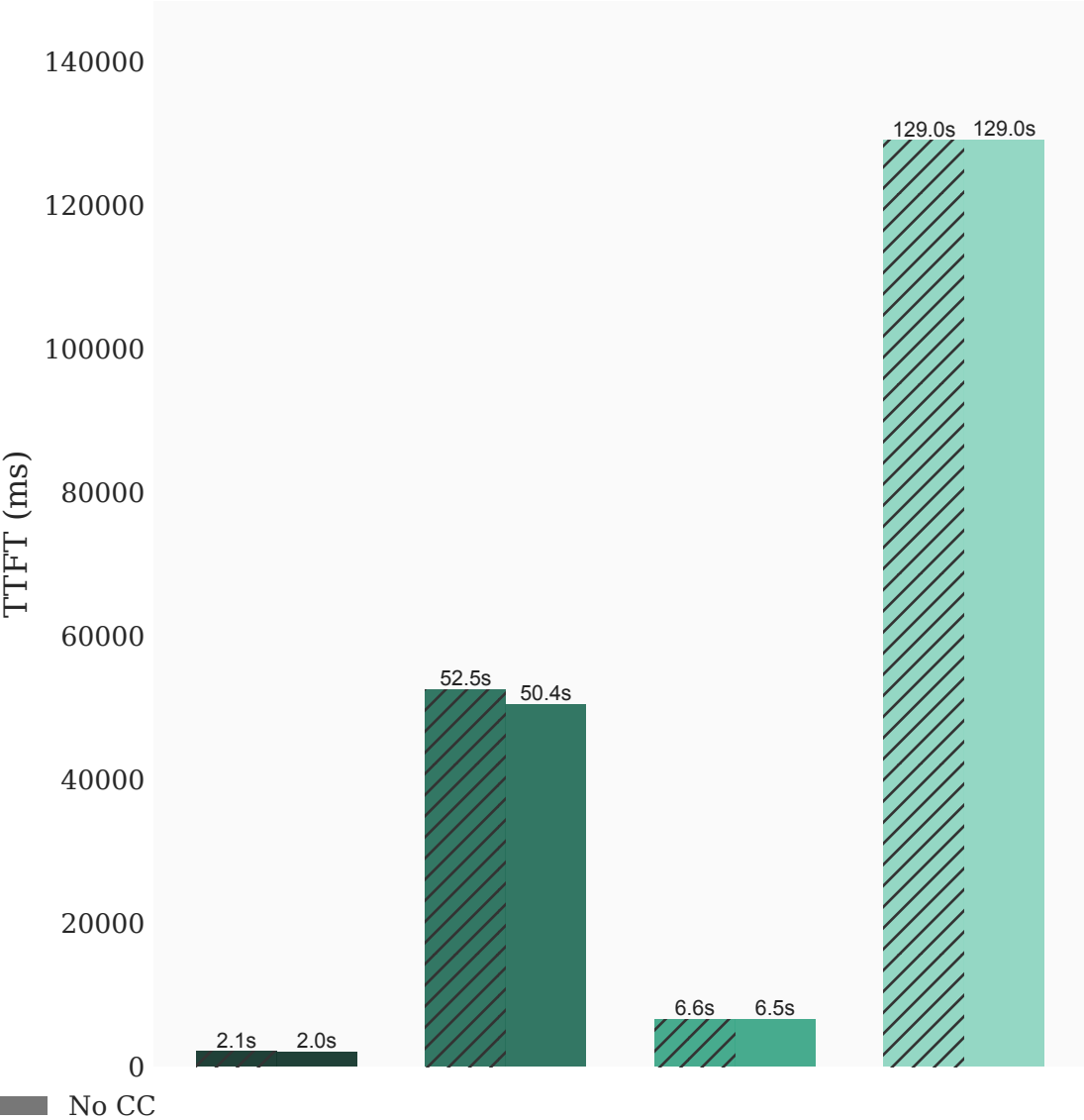- 271.1s
- 274.1s

⟋⟋ CC   No CC

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

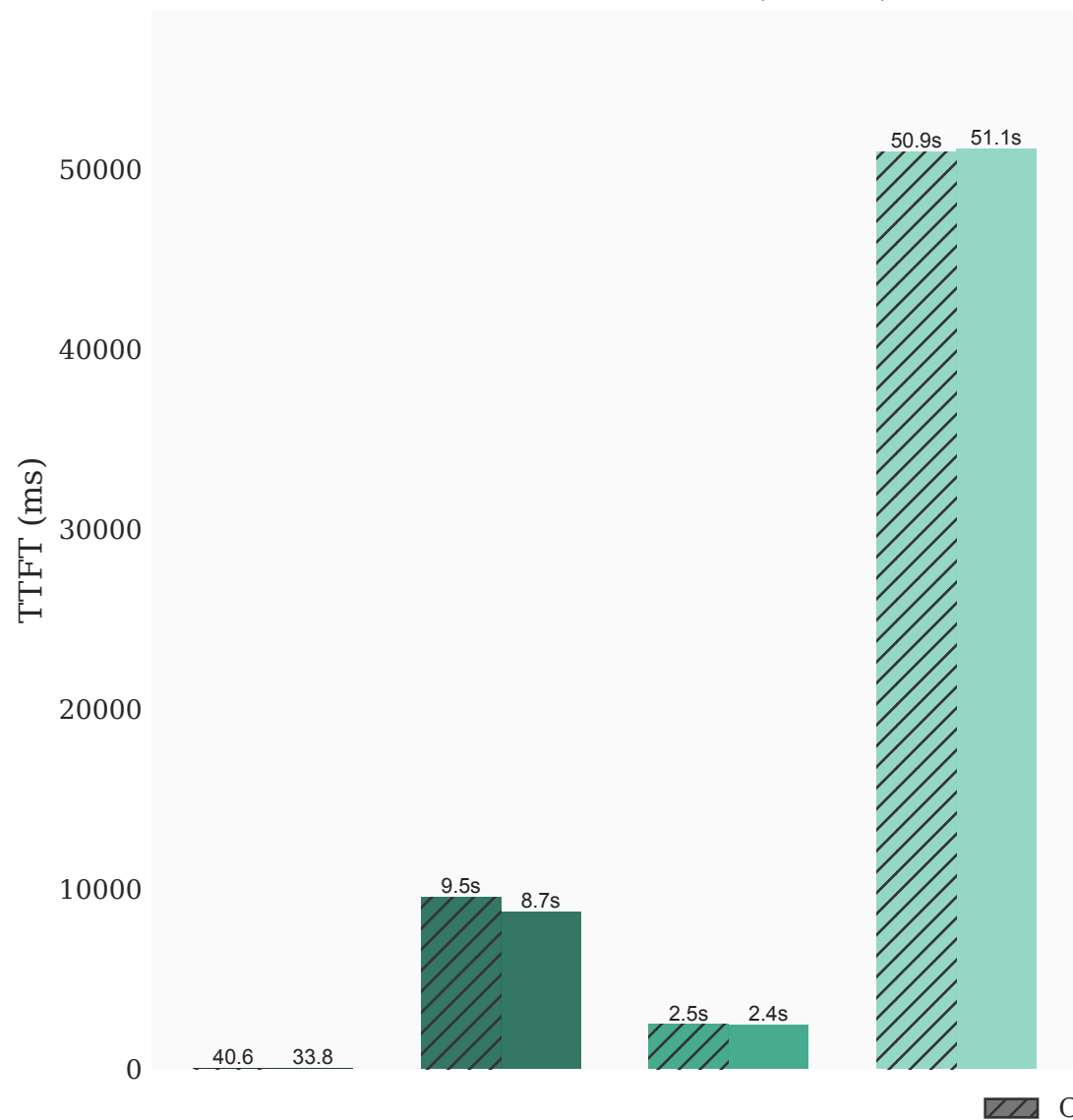# Random (1000 ⇒ 1000) (100 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

52.0s 52.0s

10.3s 9.7s

3.4s 3.4s

1.3s 1.2s

Time to First Token (P99)

TTFT (ms)

129.0s 129.0s

52.5s 50.4s

6.6s 6.5s

2.1s 2.0s

CC   No CC

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

Random (1000 ⇒ 1000) (50 Concurrent Requests)

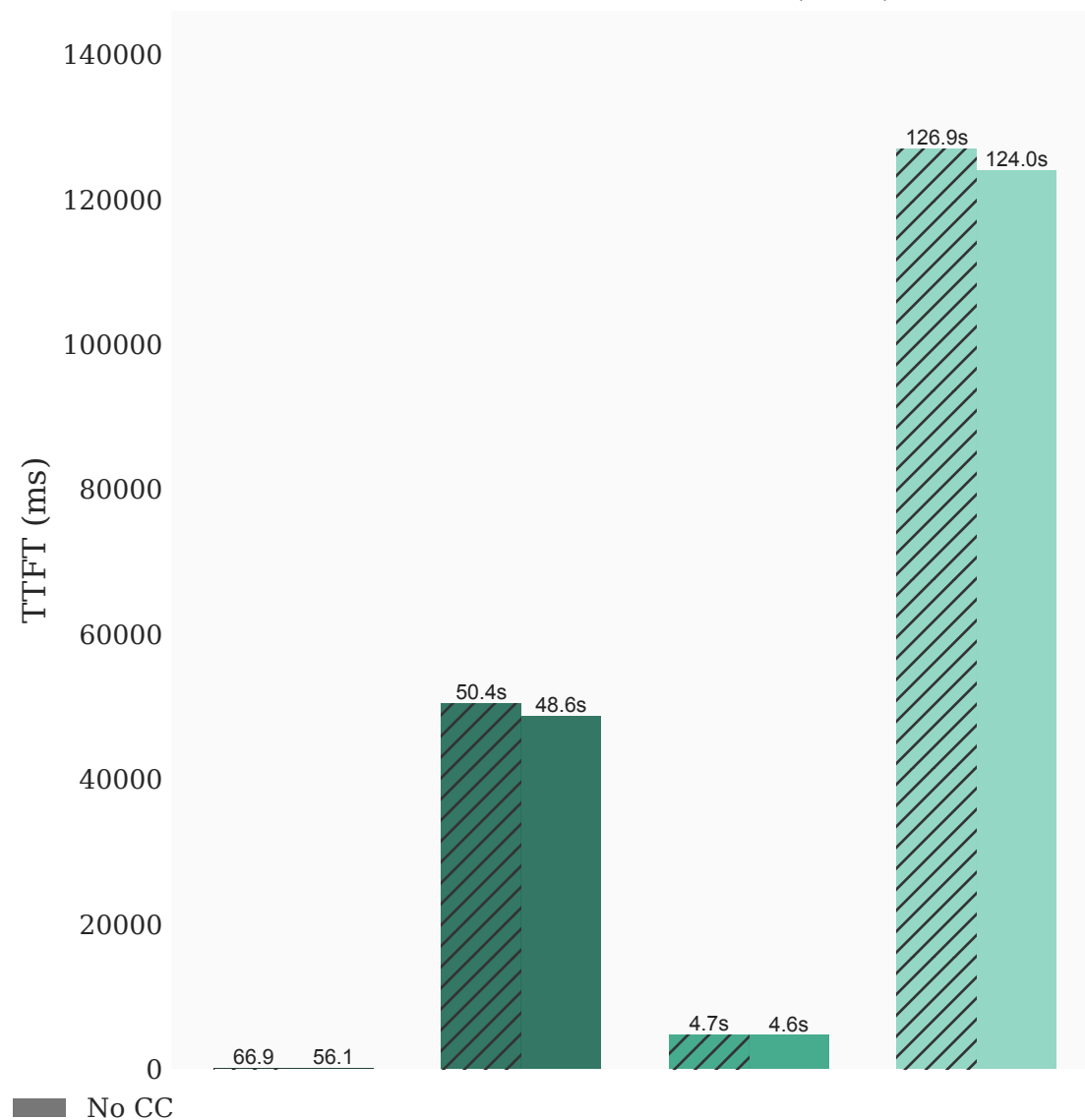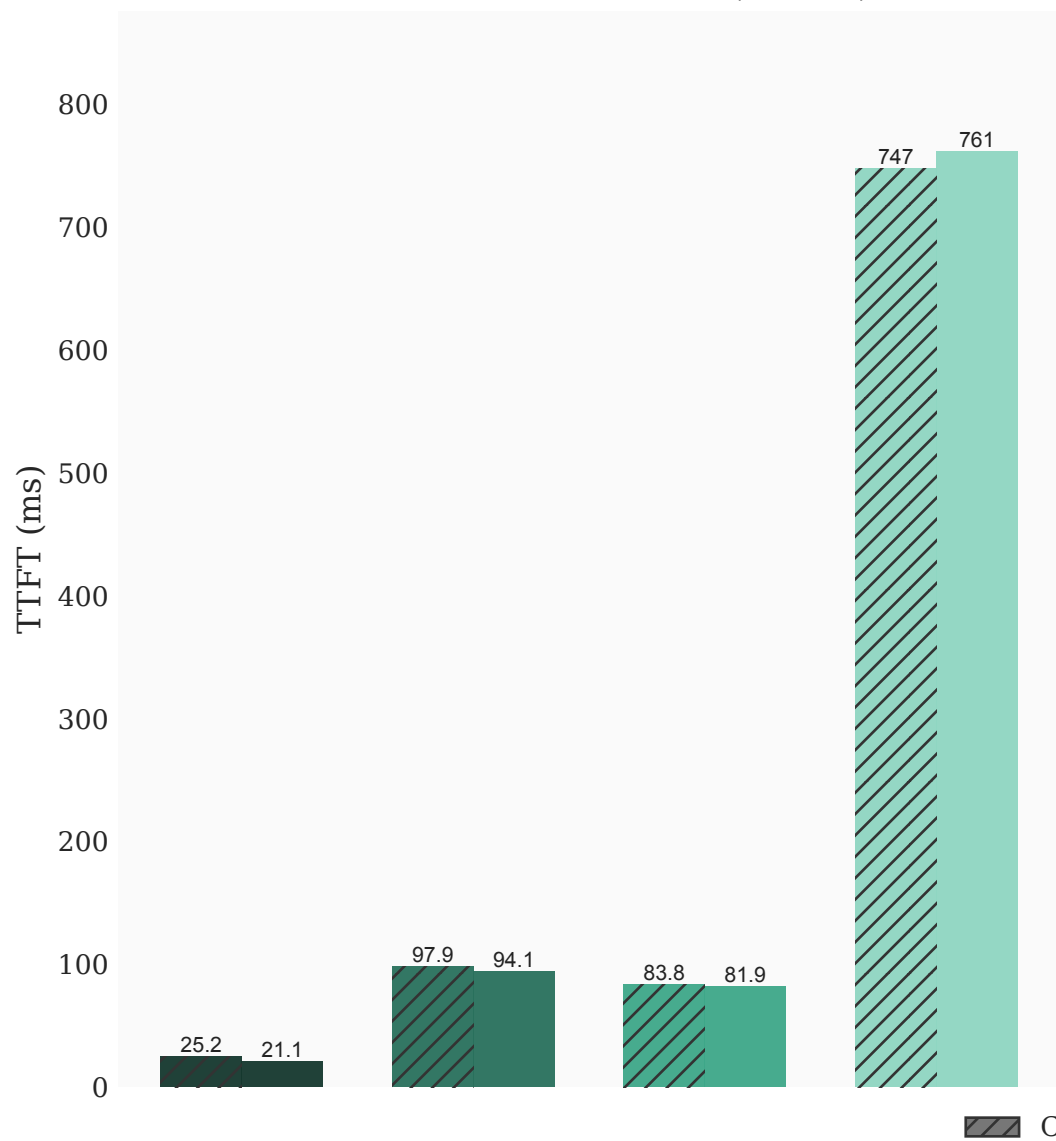## Time to First Token (Mean)

Time to First Token (P99)

Mean values:
- 40.6 / 33.8
- 9.5s / 8.7s
- 2.5s / 2.4s
- 50.9s / 51.1s

P99 values:
- 66.9 / 56.1
- 50.4s / 48.6s
- 4.7s / 4.6s
- 126.9s / 124.0s

Legend: ▨ CC   ▬ No CC

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

Random (1000 ⇒ 1000) (1 Concurrent Requests)

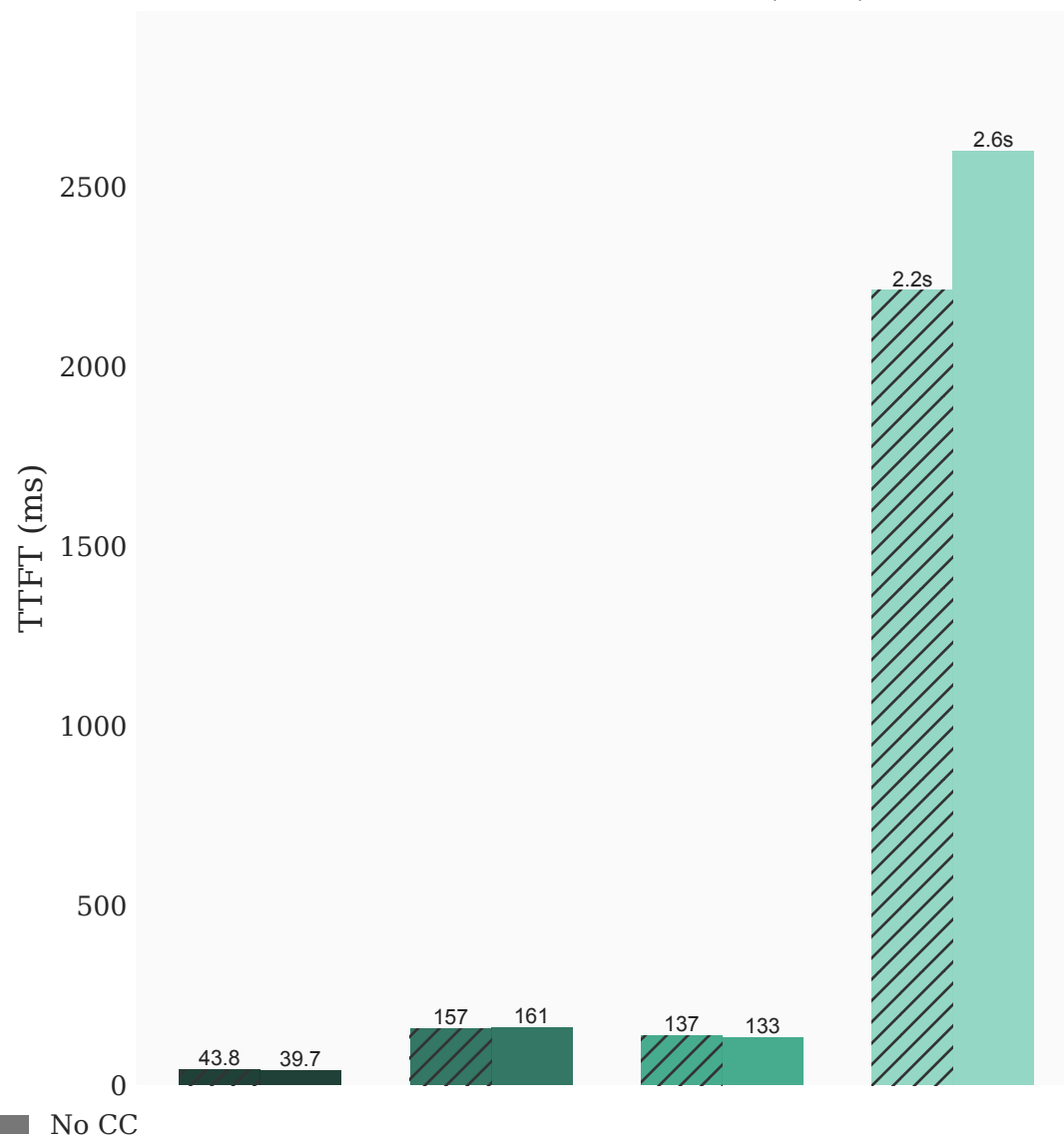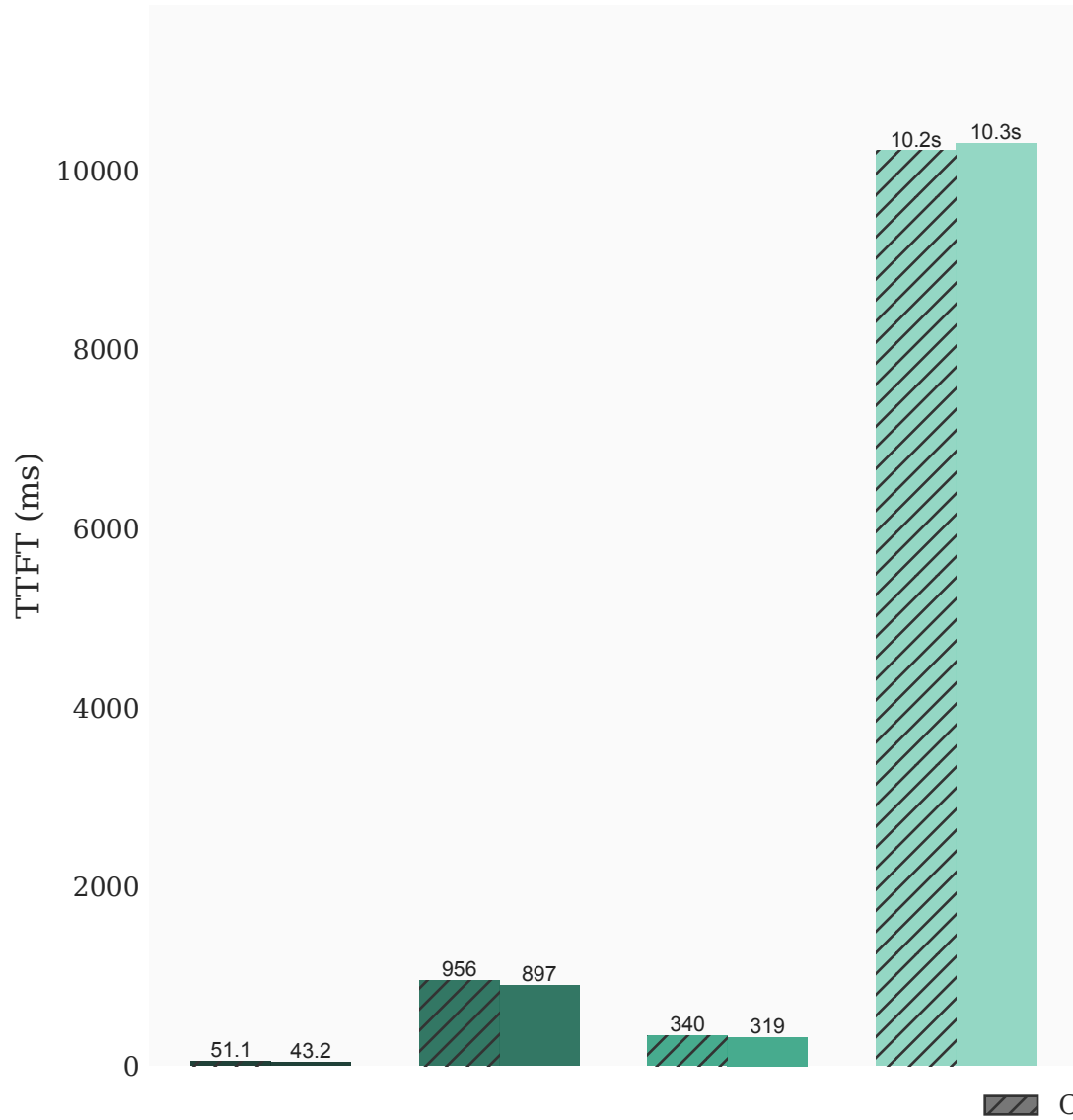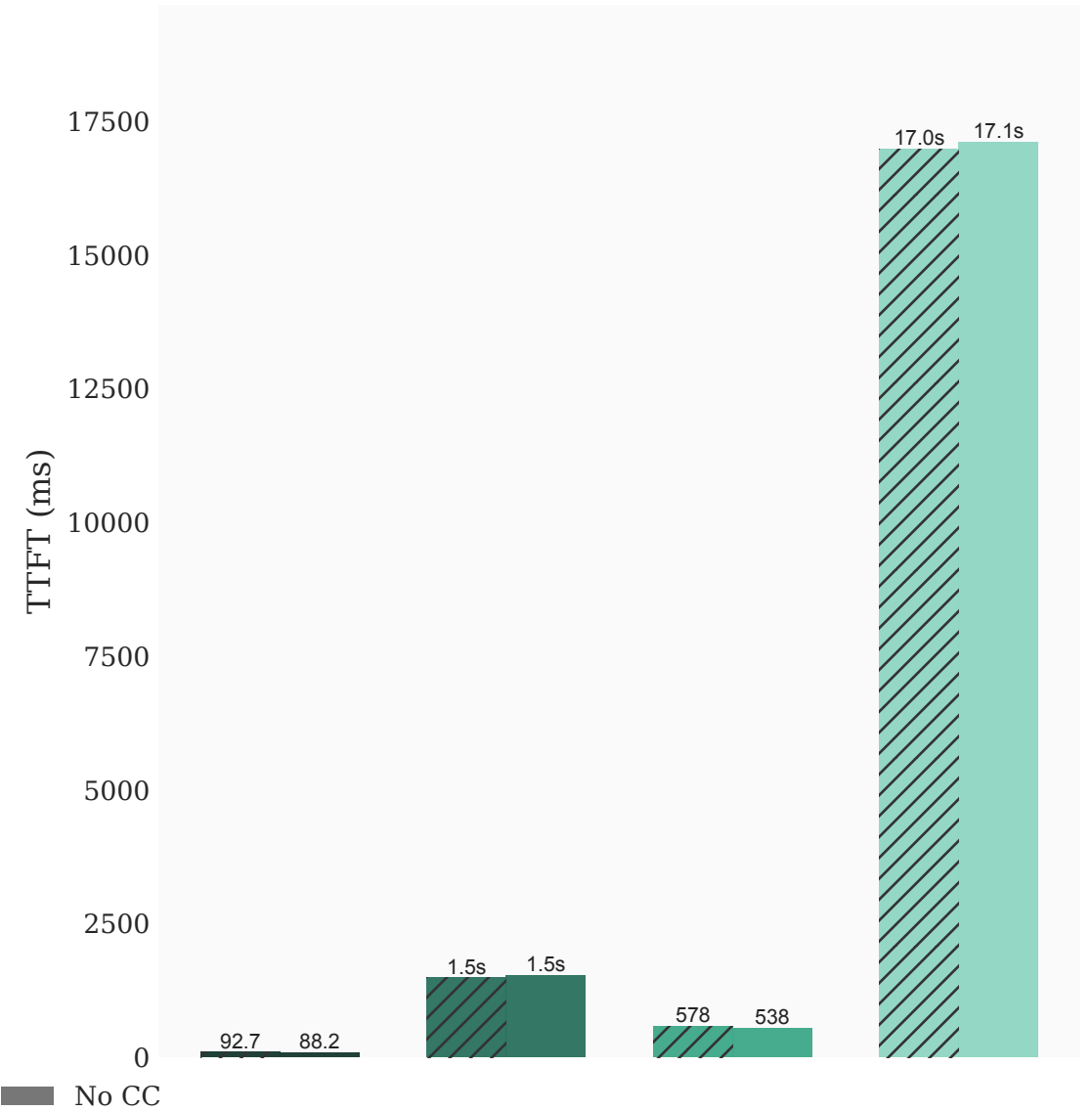## Time to First Token (Mean)

Time to First Token (P99)

CC	No CC

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

ShareGPT (100 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

10000

8000

6000

4000

2000

0

51.1 43.2   956 897   340 319   10.2s 10.3s

Time to First Token (P99)

TTFT (ms)

17500

15000

12500

10000

7500

5000

2500

0

92.7 88.2   1.5s 1.5s   578 538   17.0s 17.1s

CC   No CC

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

ShareGPT (50 Concurrent Requests)

Time to First Token (Mean)

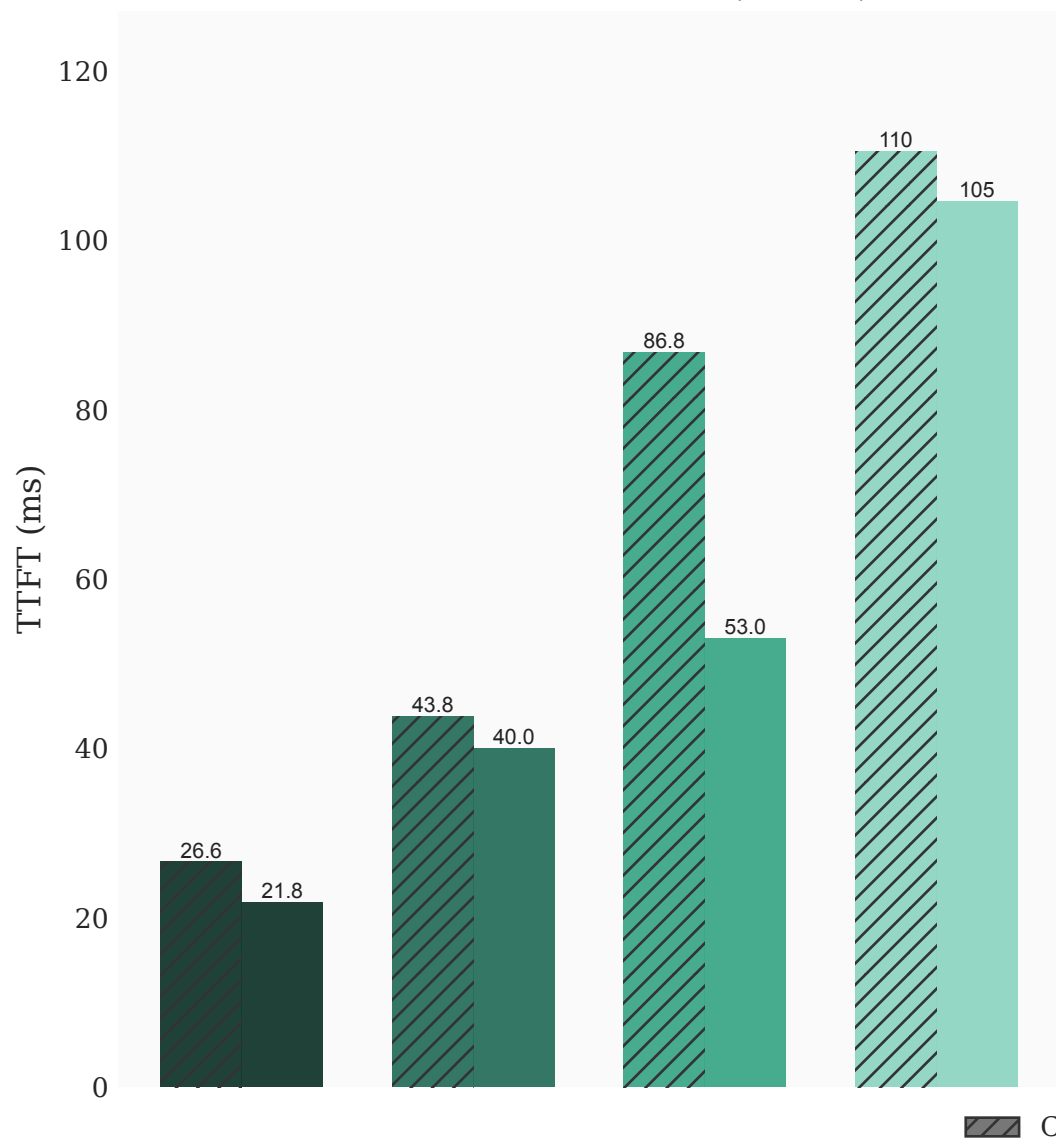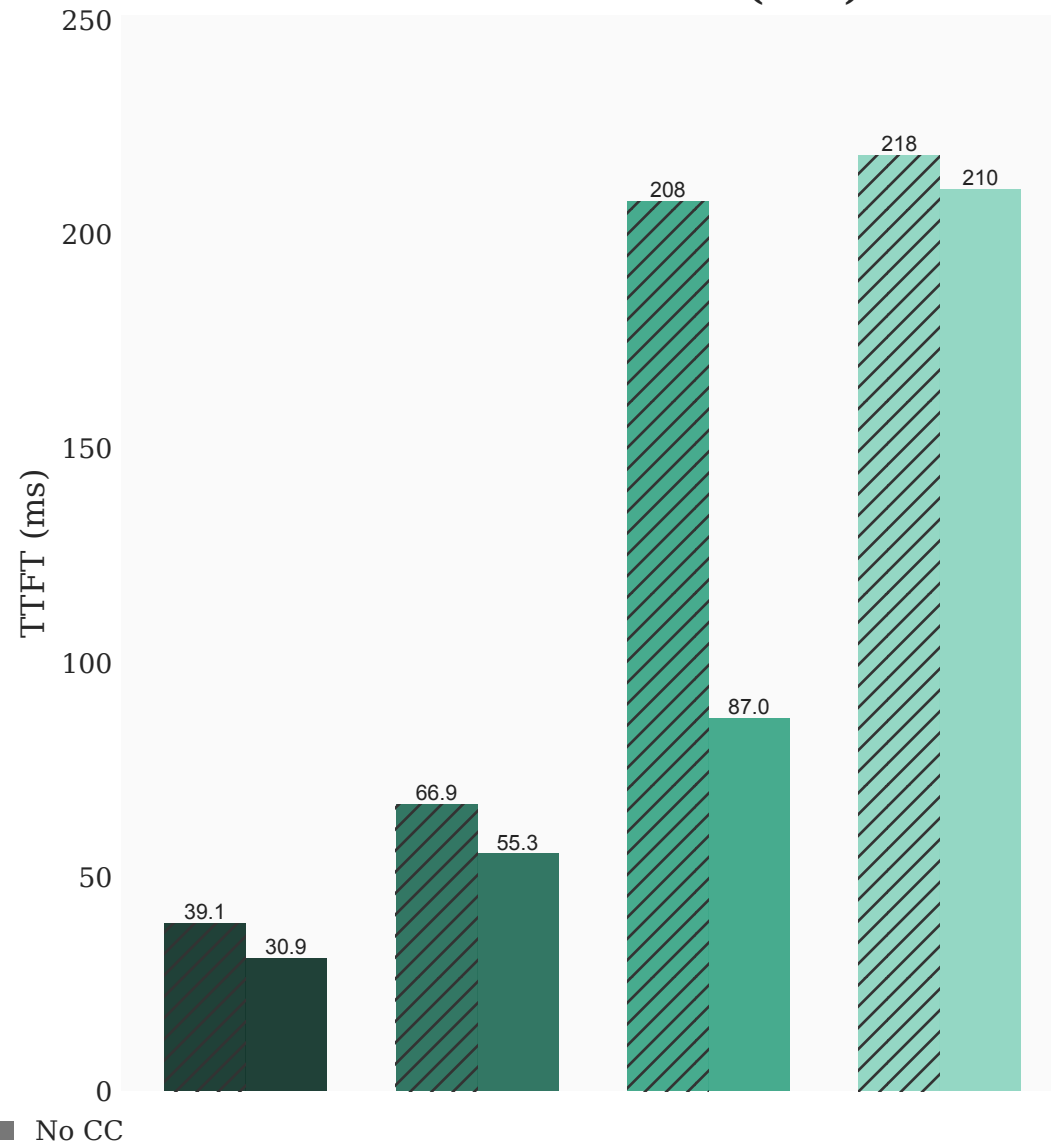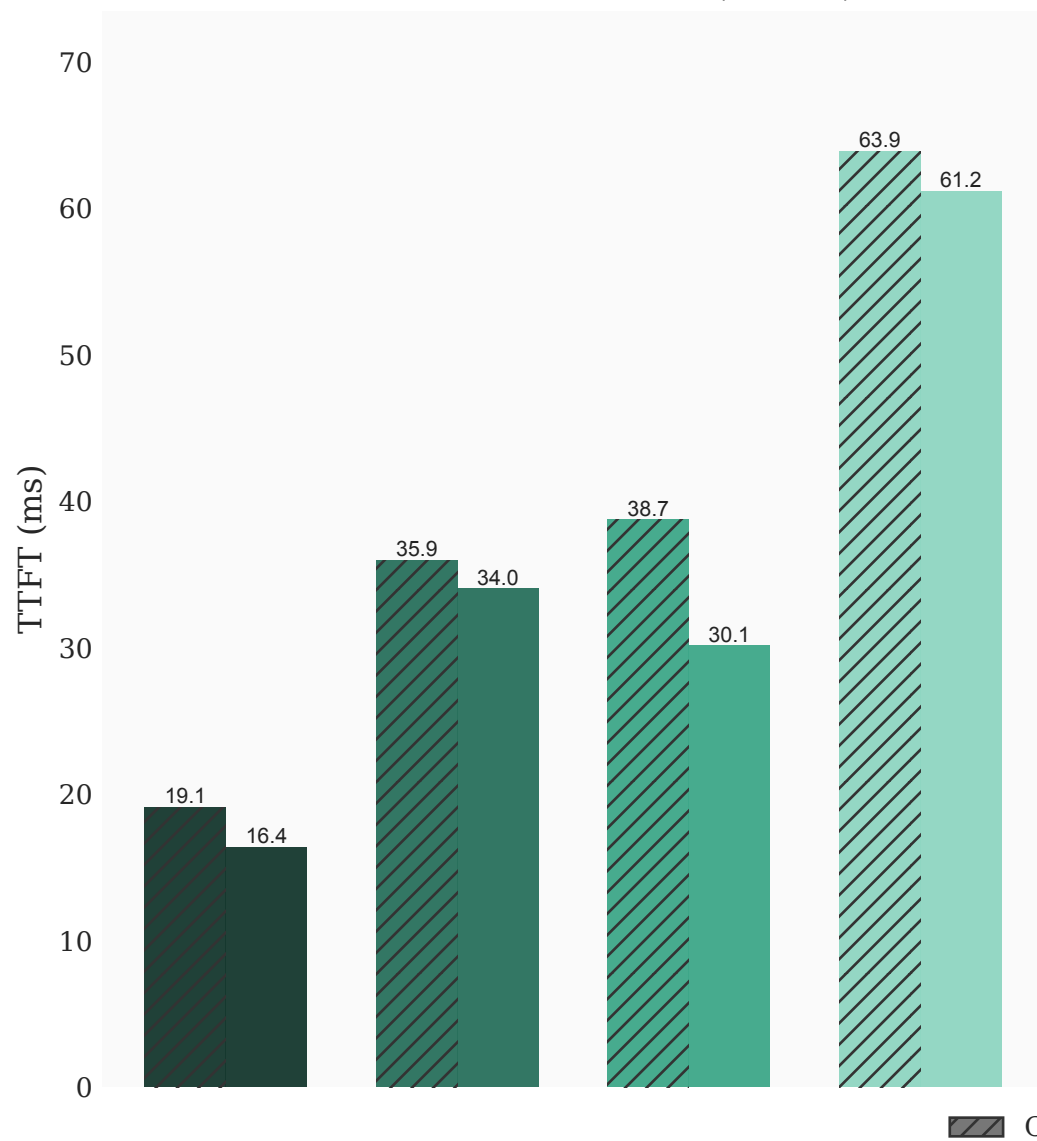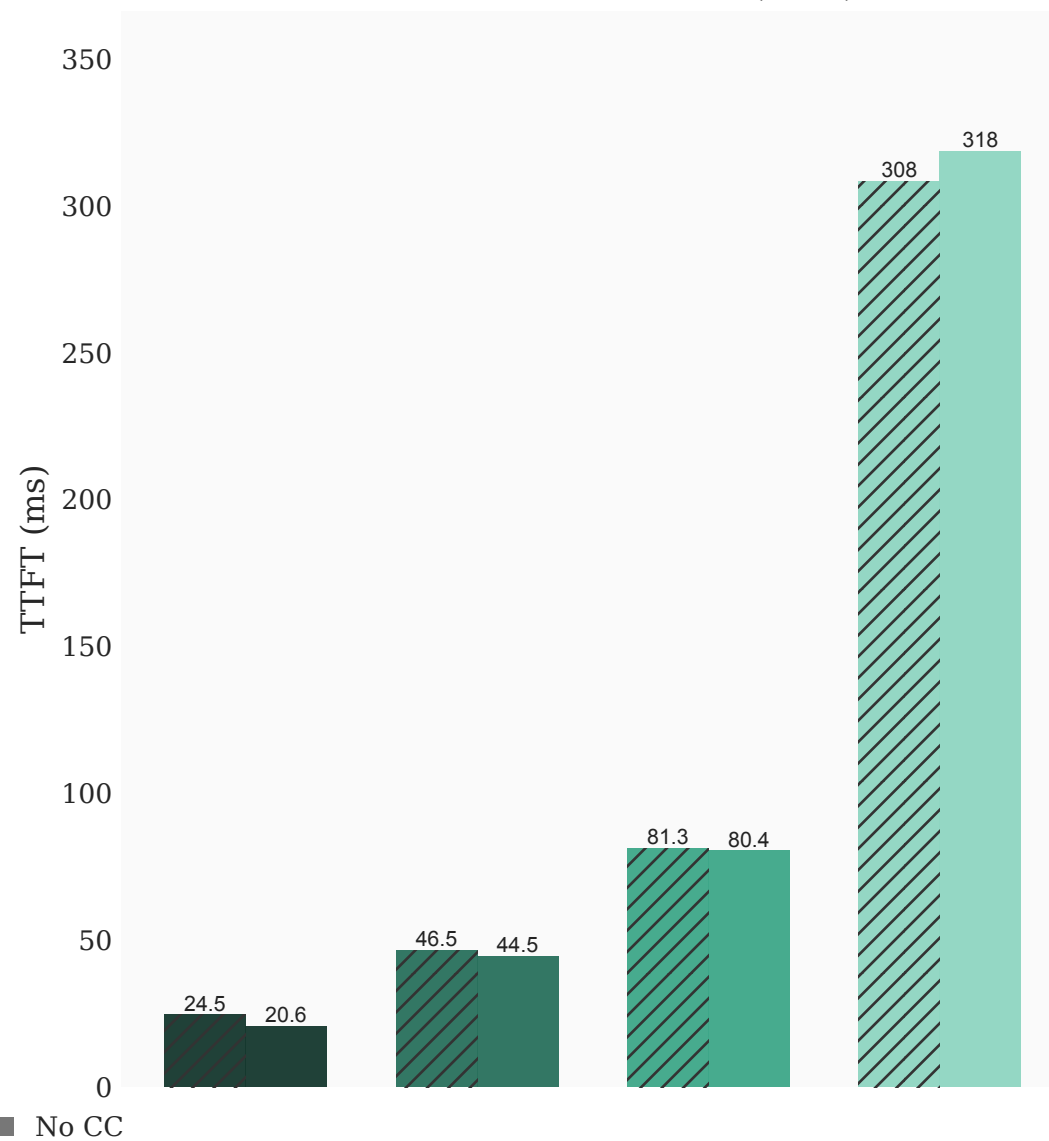## Time to First Token (P99)

Legend: CC (hatched), No CC (solid)

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

Time to First Token (Mean) values:
- LLama 3.1 8B: CC 26.6, No CC 21.8
- Mistral 3.1 24B: CC 43.8, No CC 40.0
- GPT OSS 120B: CC 86.8, No CC 53.0
- LLama 3.3 70B Int4: CC 110, No CC 105

Time to First Token (P99) values:
- LLama 3.1 8B: CC 39.1, No CC 30.9
- Mistral 3.1 24B: CC 66.9, No CC 55.3
- GPT OSS 120B: CC 208, No CC 87.0
- LLama 3.3 70B Int4: CC 218, No CC 210

ShareGPT (1 Concurrent Requests)

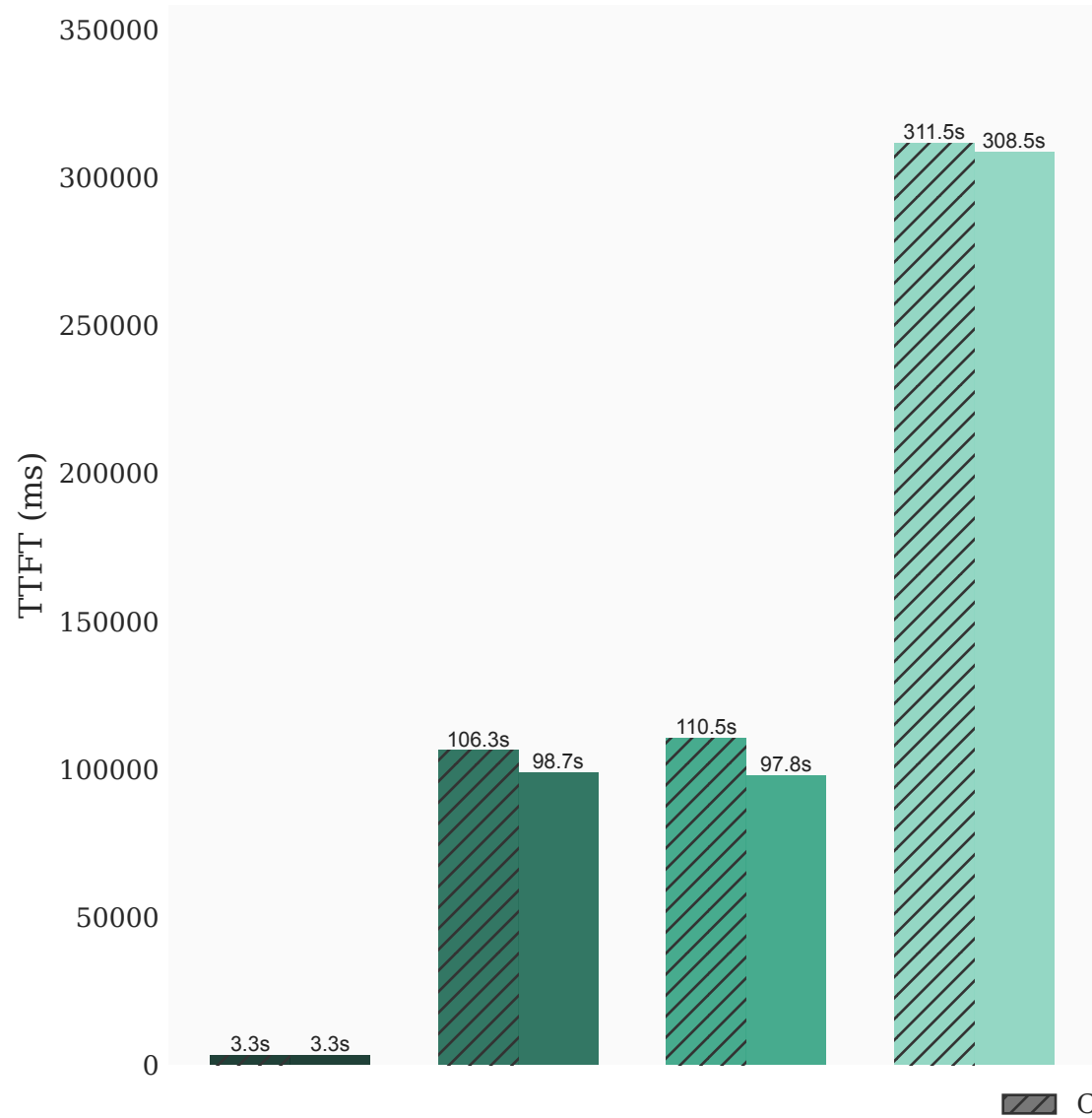## Time to First Token (Mean)
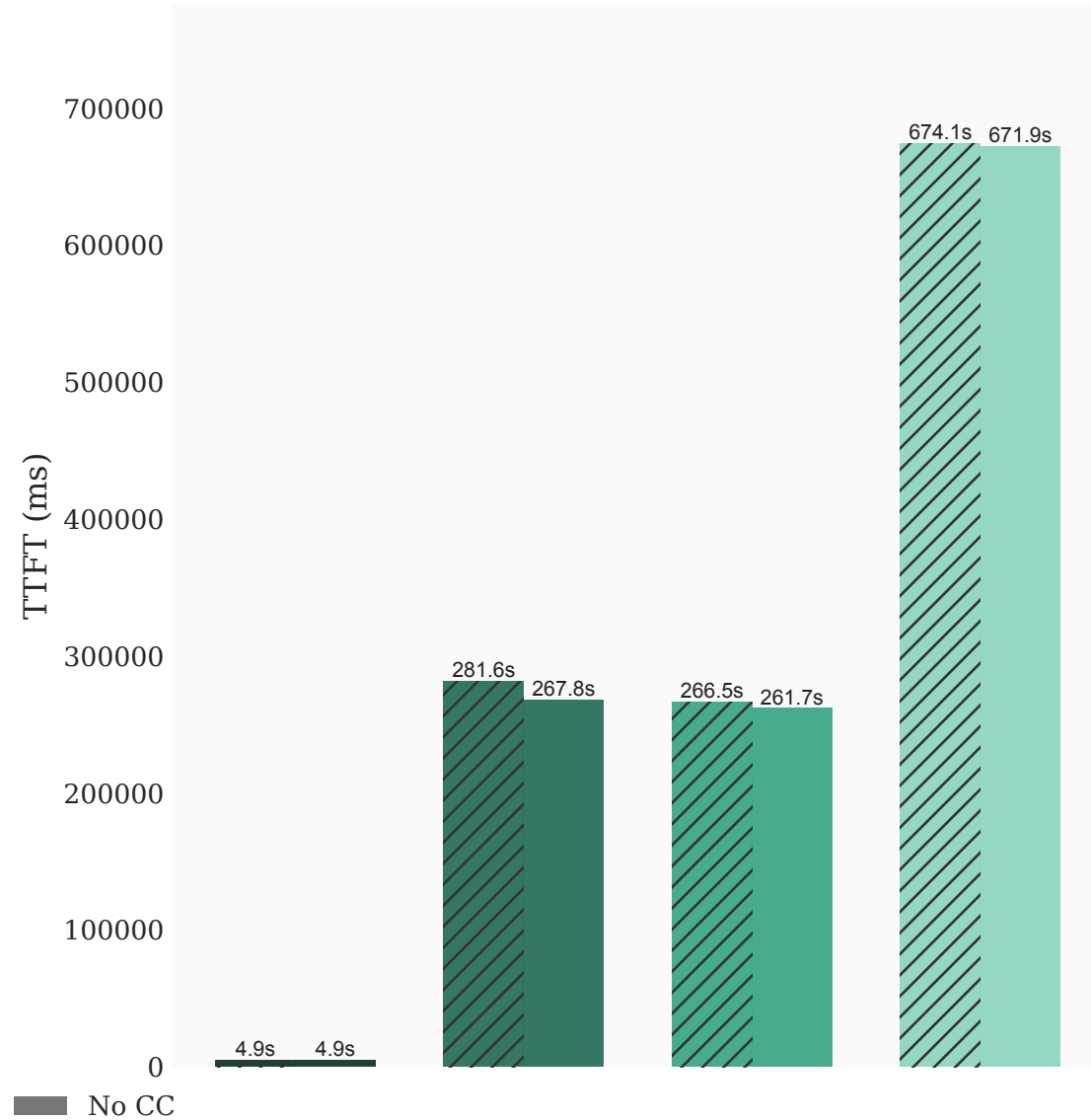

Time to First Token (P99)

Legend: CC, No CC

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

Edit 10K Characters (100 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

- 3.3s / 3.3s
- 106.3s / 98.7s
- 110.5s / 97.8s
- 311.5s / 308.5s

Time to First Token (P99)

TTFT (ms)

- 4.9s / 4.9s
- 281.6s / 267.8s
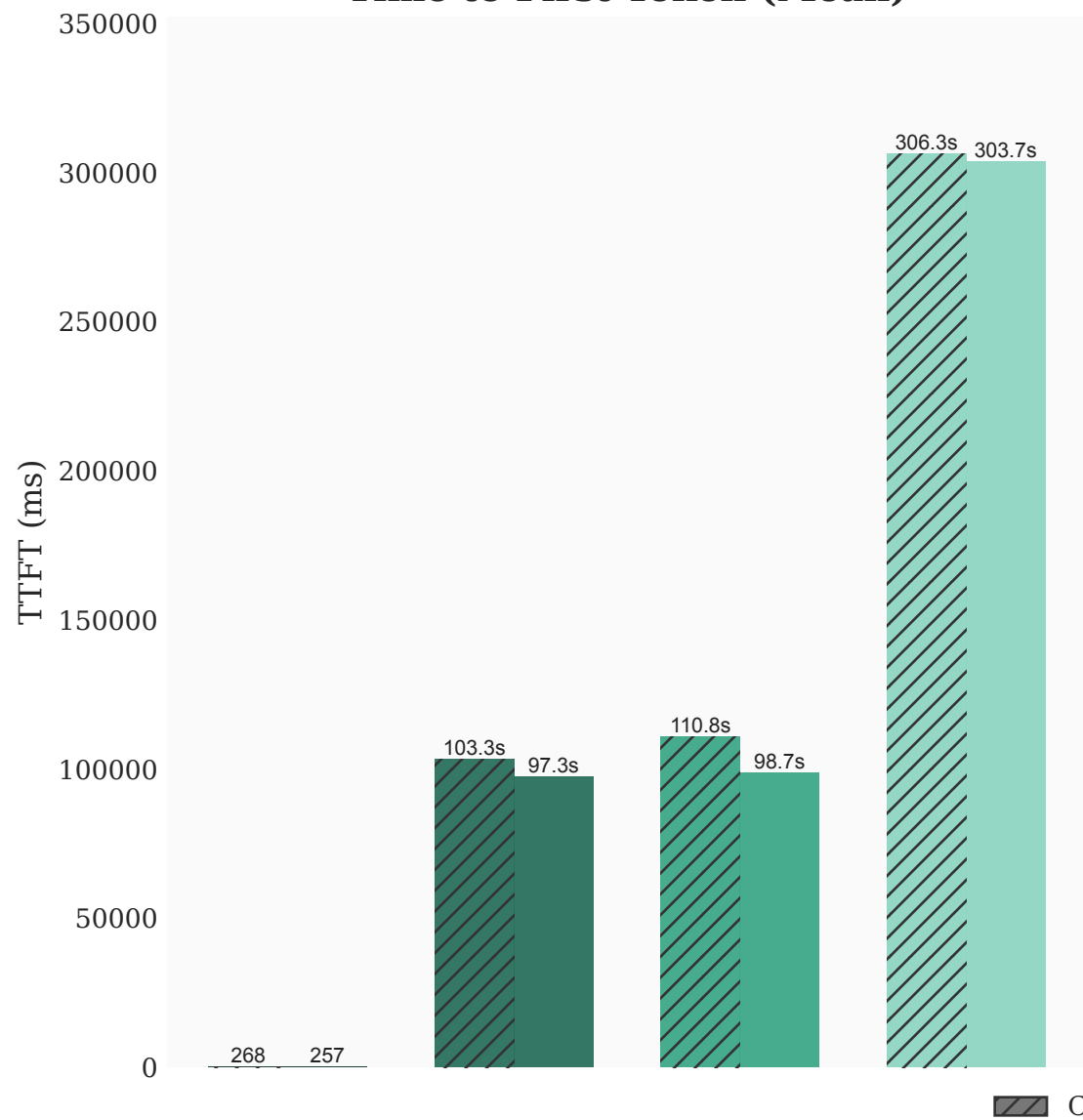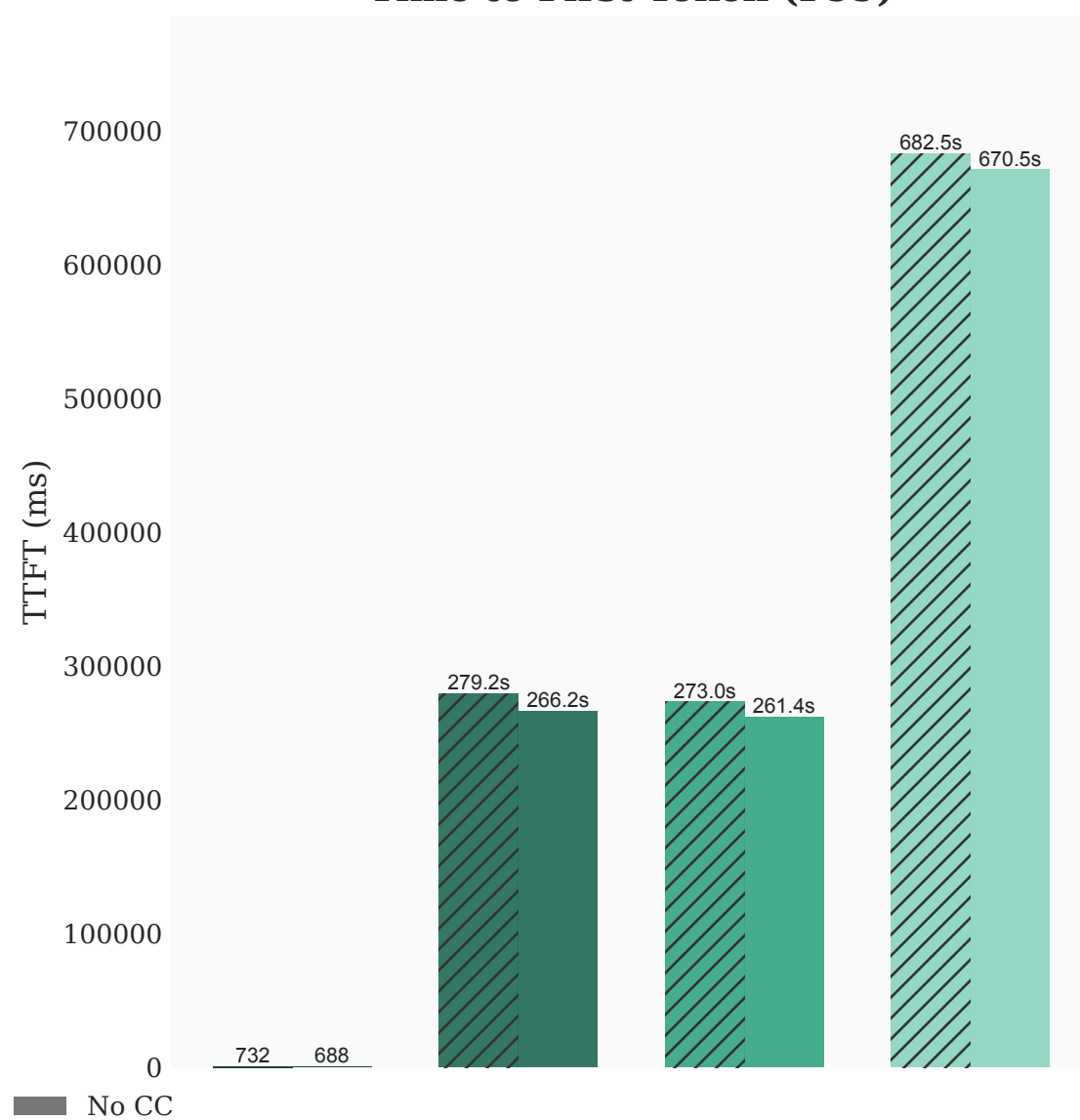- 266.5s / 261.7s
- 674.1s / 671.9s

Legend: CC | No CC

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

Edit 10K Characters (50 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

268 257
103.3s 97.3s
110.8s 98.7s
306.3s 303.7s

Time to First Token (P99)

TTFT (ms)

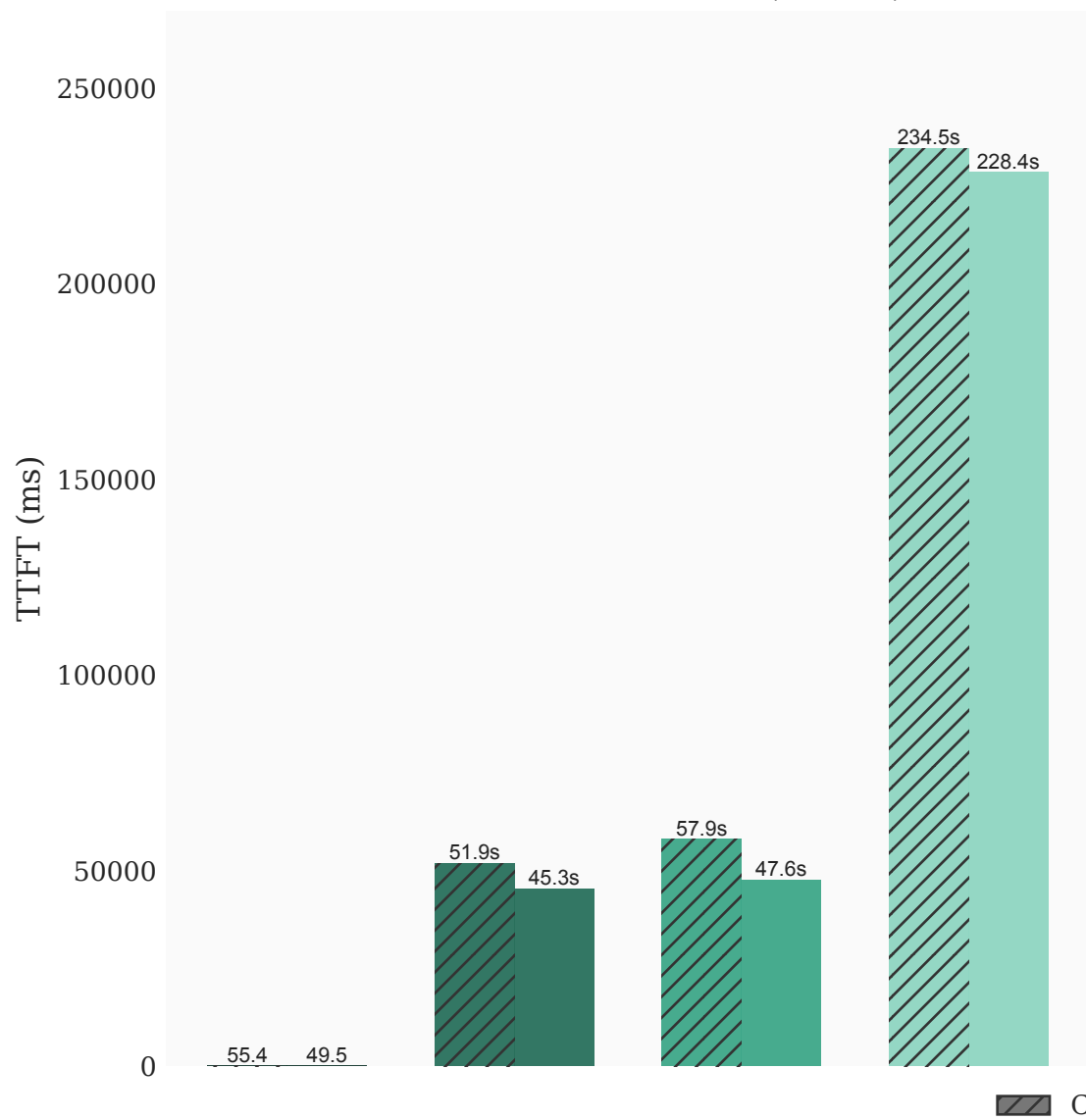732 688
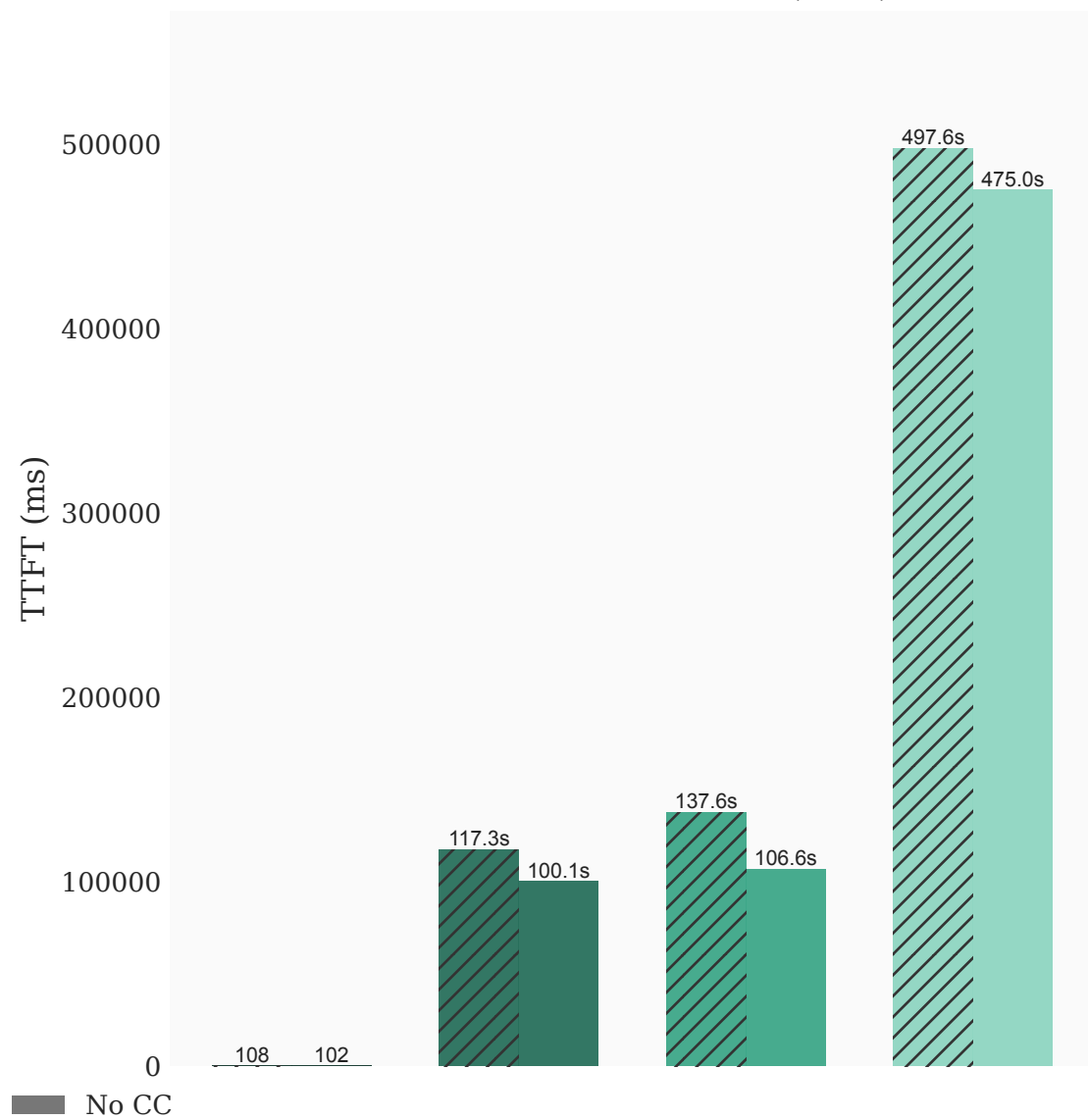279.2s 266.2s
273.0s 261.4s
682.5s 670.5s

CC   No CC

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

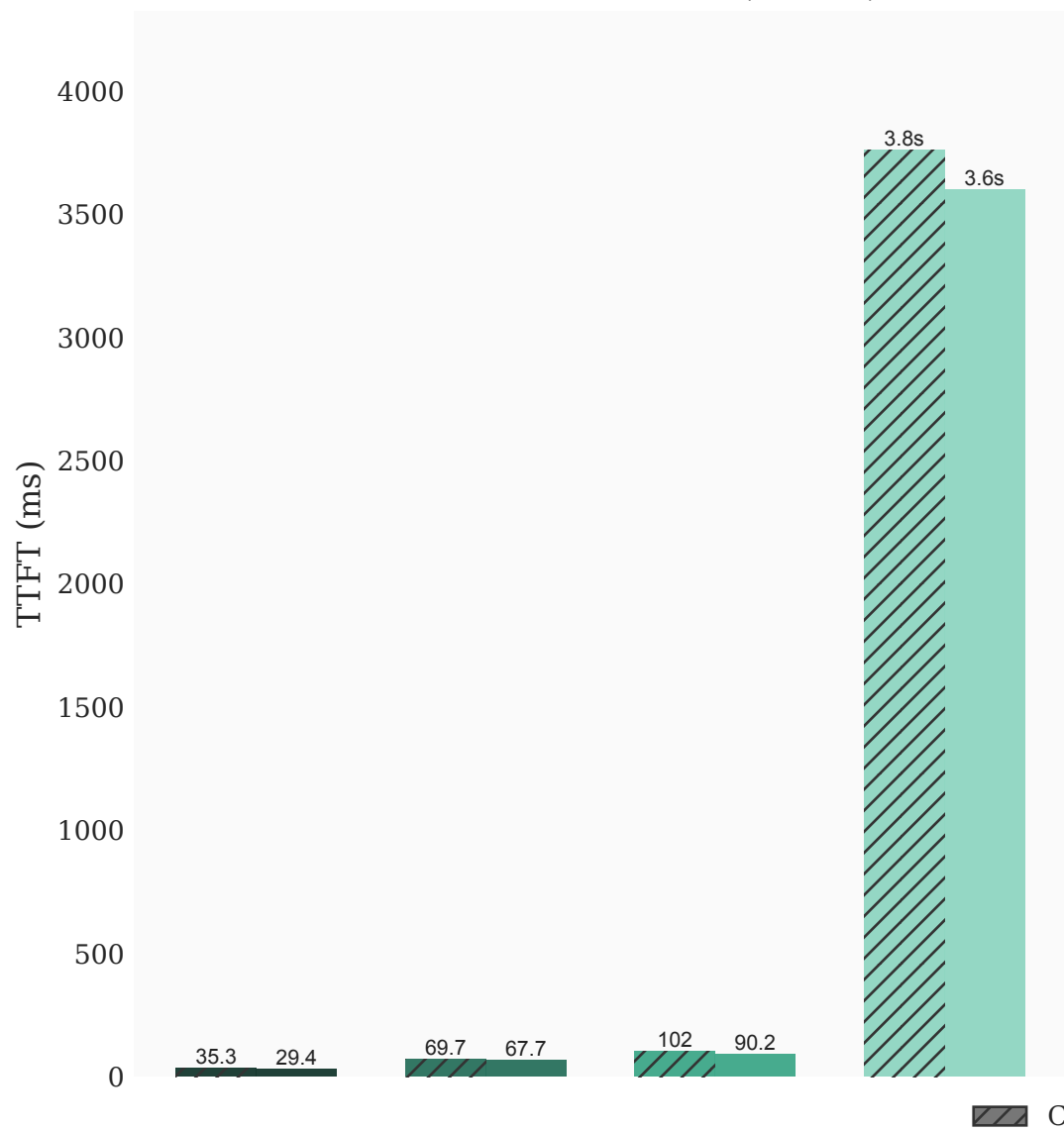# Edit 10K Characters (1 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

- 55.4
- 49.5
- 51.9s
- 45.3s
- 57.9s
- 47.6s
- 234.5s
- 228.4s

Time to First Token (P99)

TTFT (ms)

- 108
- 102
- 117.3s
- 100.1s
- 137.6s
- 106.6s
- 497.6s
- 475.0s

CC   No CC
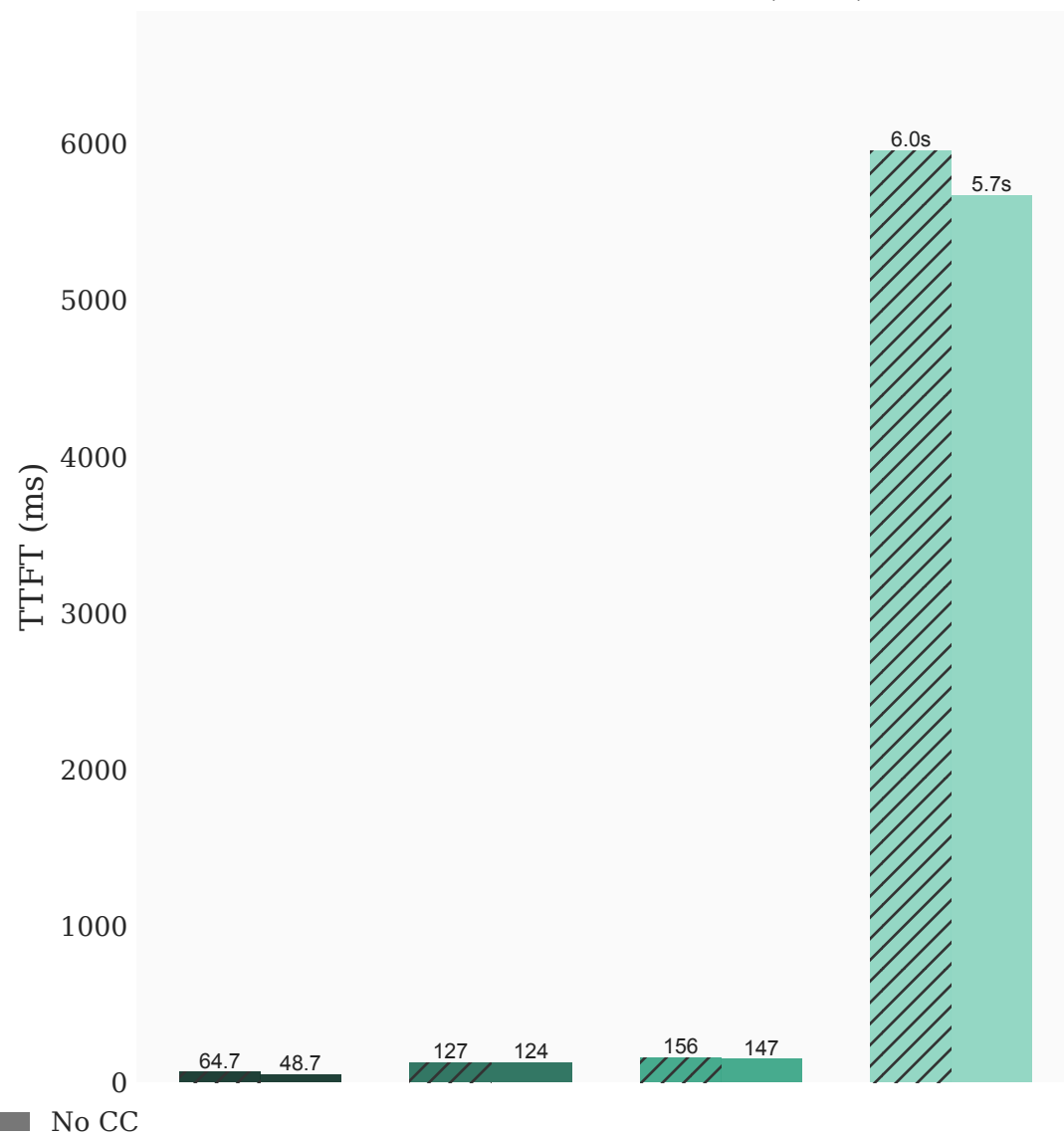
LLama 3.1 8B   Mistral 3.1 24B   GPT OSS 120B   LLama 3.3 70B Int4

Numina Math (100 Concurrent Requests)

Time to First Token (Mean)

TTFT (ms)

- LLama 3.1 8B: 35.3 (CC), 29.4 (No CC)
- Mistral 3.1 24B: 69.7 (CC), 67.7 (No CC)
- GPT OSS 120B: 102 (CC), 90.2 (No CC)
- LLama 3.3 70B Int4: 3.8s (CC), 3.6s (No CC)

Time to First Token (P99)

TTFT (ms)

- LLama 3.1 8B: 64.7 (CC), 48.7 (No CC)
- Mistral 3.1 24B: 127 (CC), 124 (No CC)
- GPT OSS 120B: 156 (CC), 147 (No CC)
- LLama 3.3 70B Int4: 6.0s (CC), 5.7s (No CC)

Legend: CC, No CC

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

Numina Math (50 Concurrent Requests)

Time to First Token (Mean)

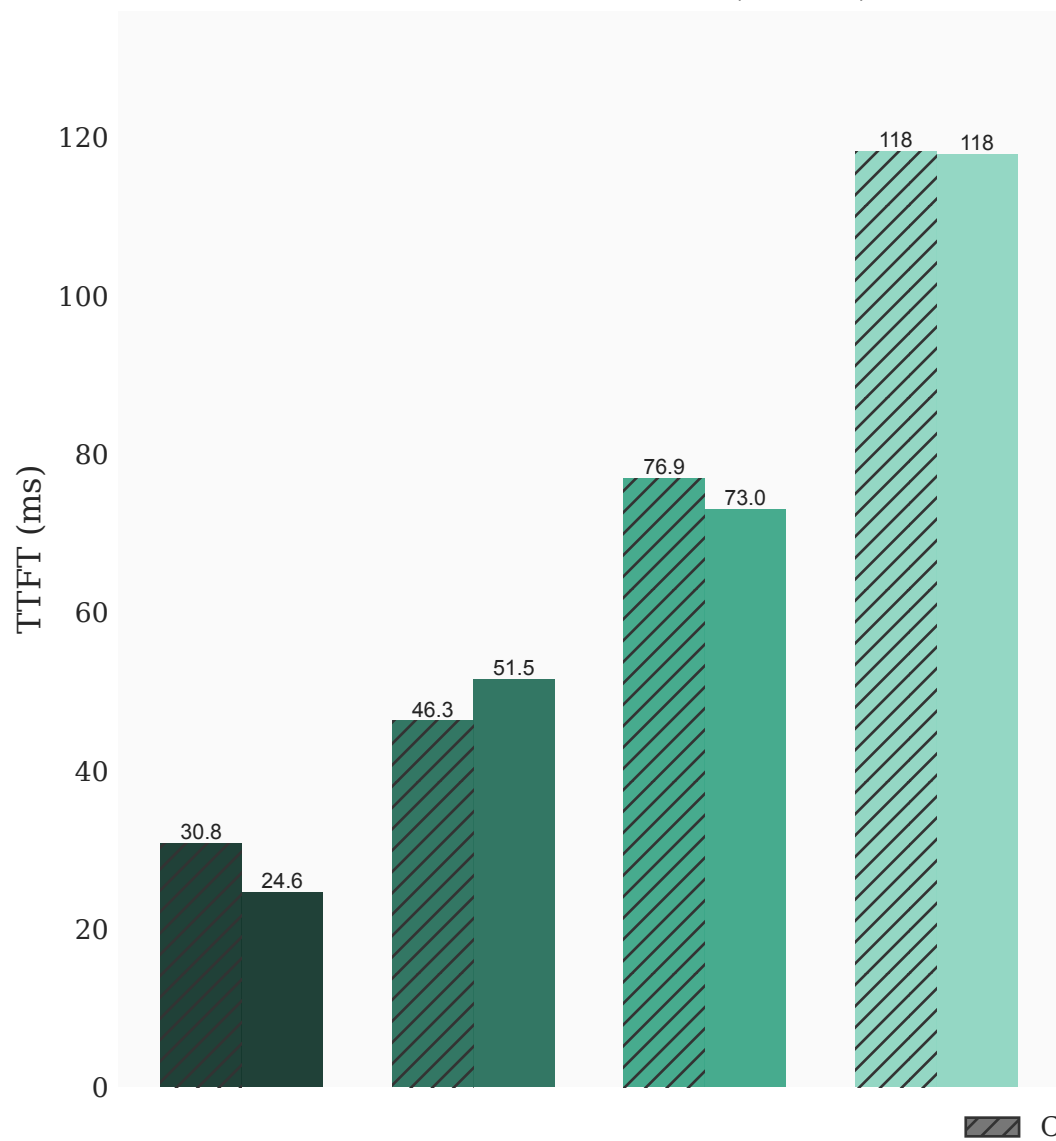TTFT (ms)

Values: 30.8, 24.6, 46.3, 51.5, 76.9, 73.0, 118, 118

Time to First Token (P99)

TTFT (ms)

Values: 46.6, 36.7, 65.1, 97.1, 124, 115, 235, 235

Legend: CC, No CC

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

Numina Math (1 Concurrent Requests)

Time to First Token (Mean)
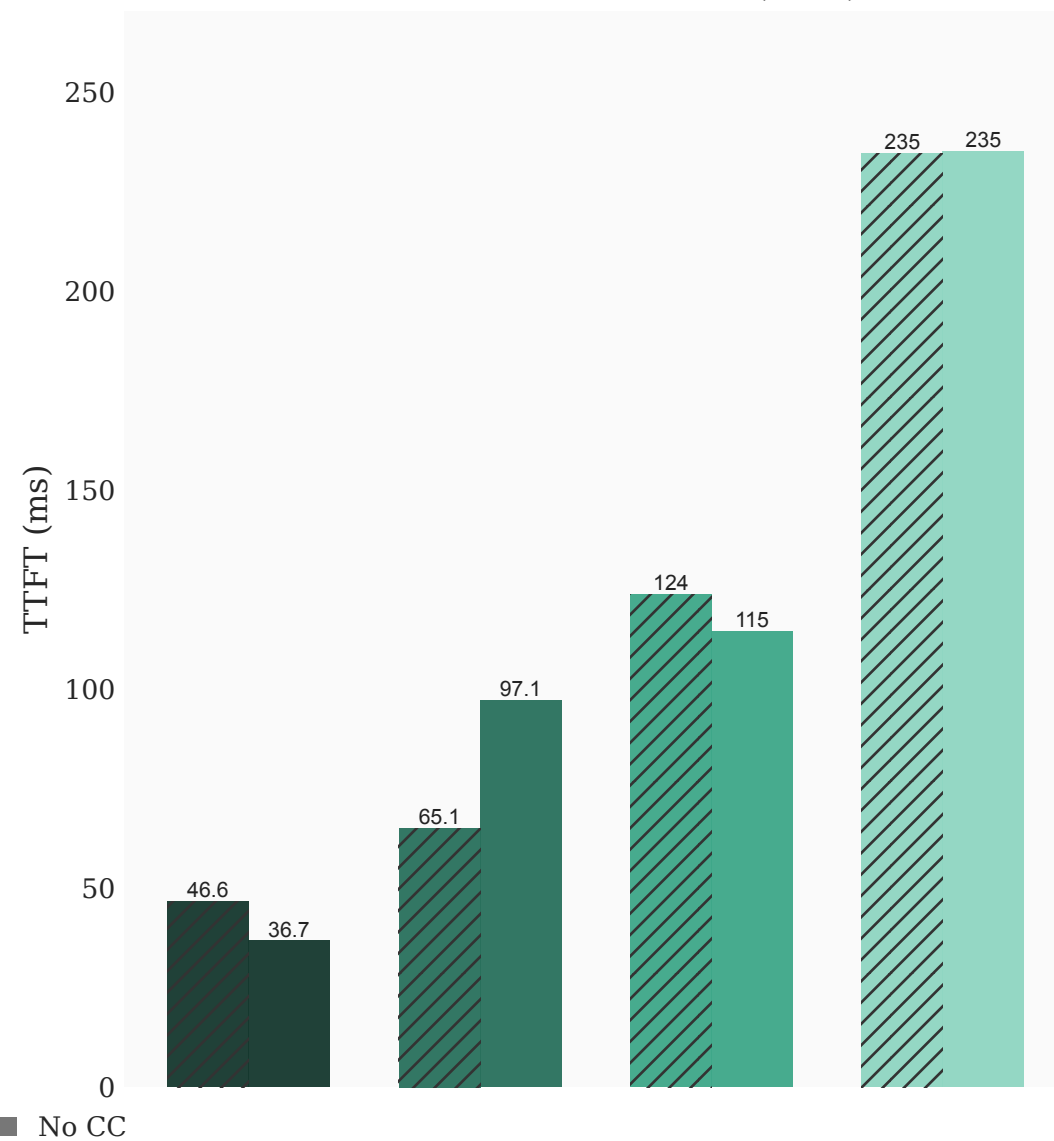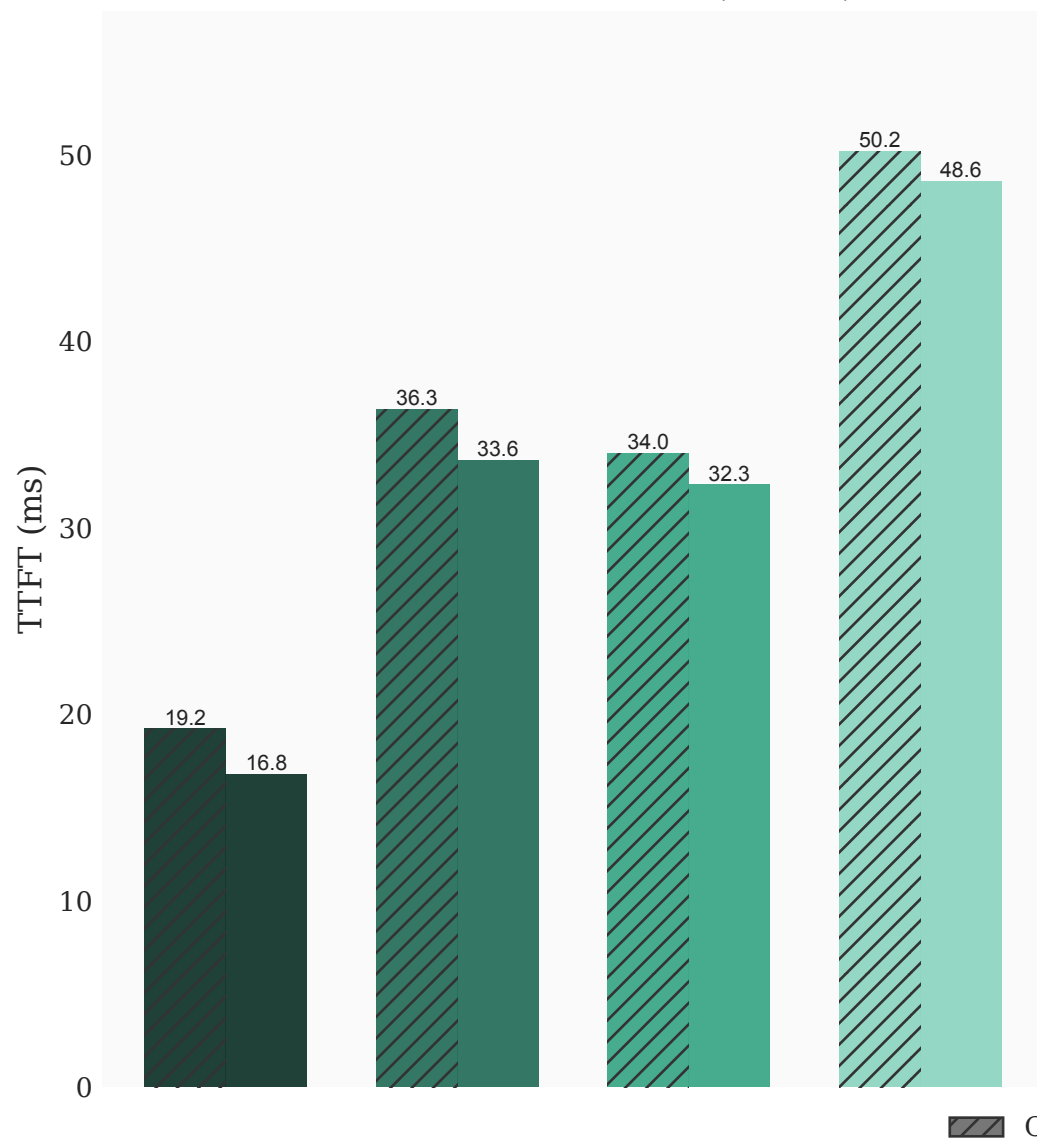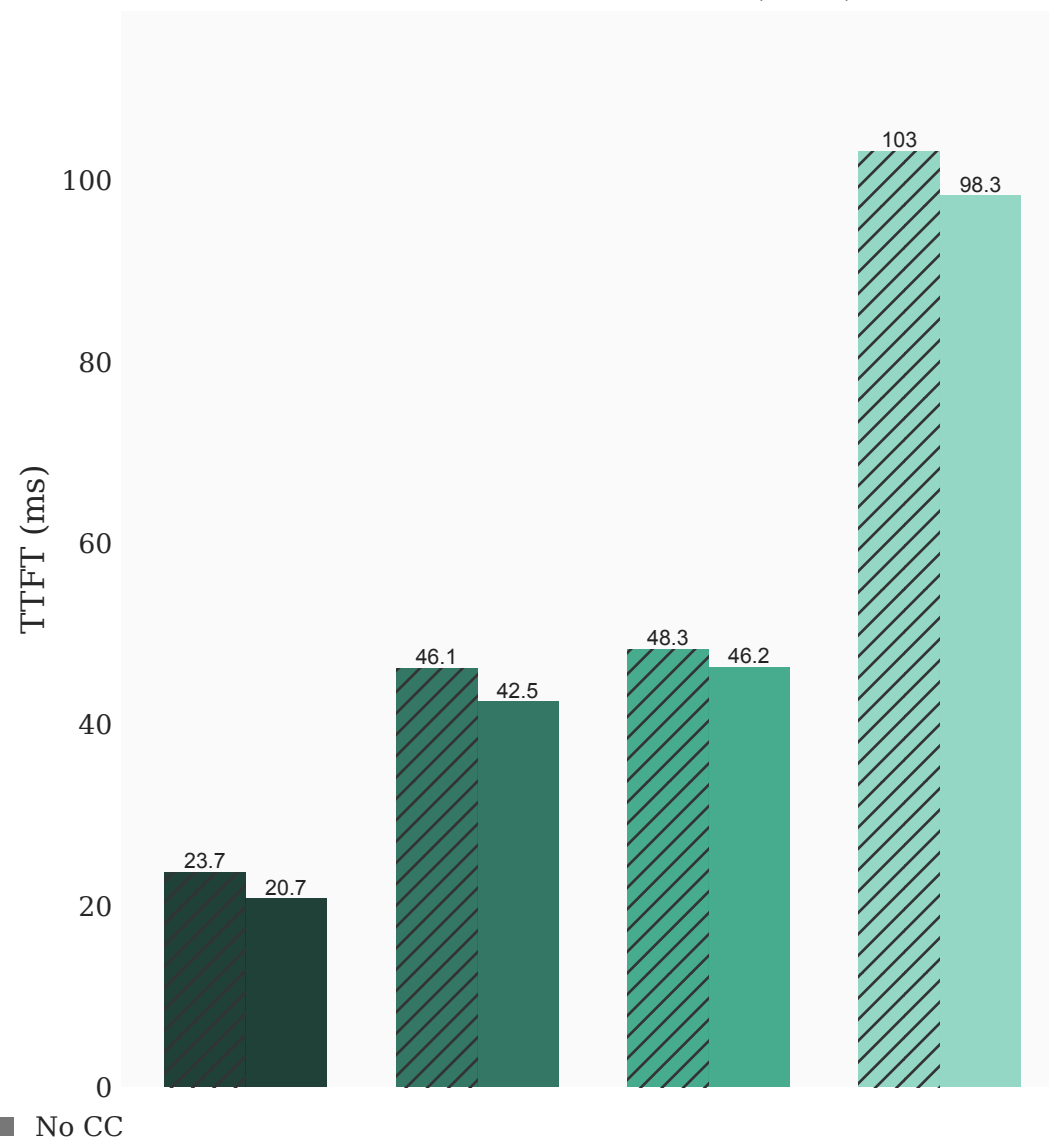

Time to First Token (P99)

Legend: CC, No CC

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