TTFT: ContextCache vs Full Prefill (20 tools)

787 ms

114 ms

7x faster

Legend:
- KV Cache Load
- Suffix Prefill

Y-axis: Time to First Token (ms)

X-axis categories:
- Full Prefill (every request)
- ContextCache (cache hit)