# Local LLM Benchmark

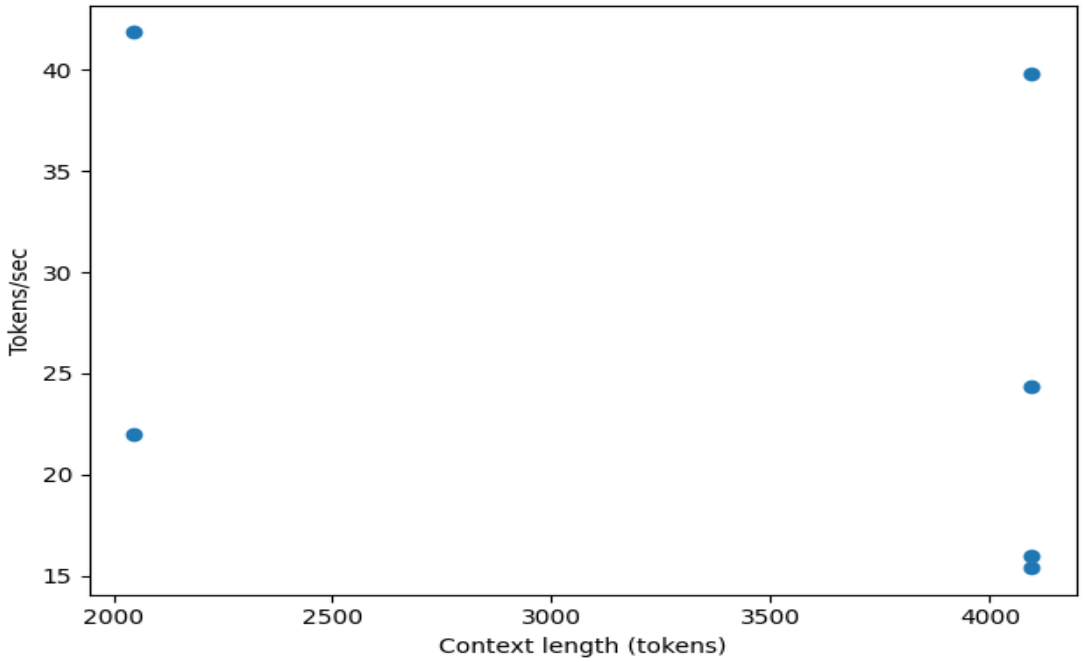
Apple Silicon · LM Studio

Throughput · Latency · Context Scaling

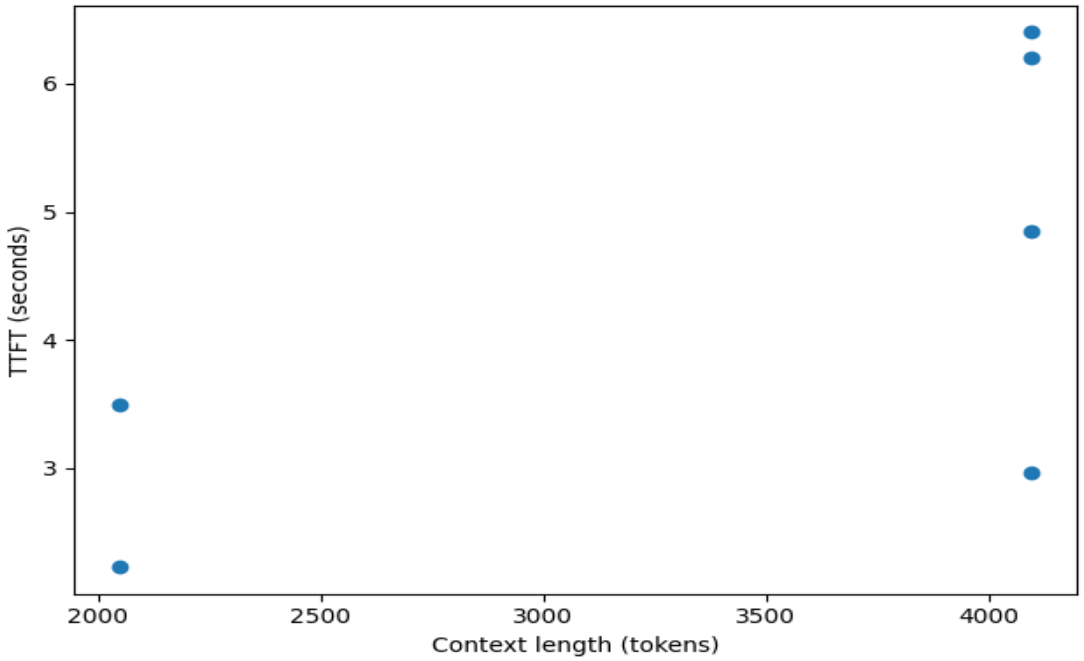
## Benchmark Results (Canonical Dataset)

| Model | Quant | Context | Tokens/sec | TTFT (s) | Output Tokens |
|---|---|---|---|---|---|
| DeepSeek-Coder-V2-Lite | Q8_0 | 4096 | 15.4 | 6.4 | 596 |
| DeepSeek-Coder-V2-Lite | Q4_K_M | 4096 | 16.0 | 6.2 | 597 |
| DeepSeek-Coder-V2-Lite | Q4_K_M | 2048 | 22.0 | 3.5 | 596 |
| Qwen3-4B-2507 Kimi-K2 Thinking | Q8 | 4096 | 24.37 | 4.85 | 2295 |
| Qwen3-VI-4B | Q4 | 4096 | 39.78 | 2.97 | 651 |
| Qwen3-VL-4B | Q4 | 2048 | 41.84 | 2.23 | 651 |

## Tokens/sec vs Context Length




## TTFT vs Context Length


TTFT (seconds)
6
5
4
3
2000
2500
3000
3500
4000
Context length (tokens)

## Practical Recommendations

**Fast interactive coding**
Qwen3-VL-4B @ 2048

• Lowest TTFT
• Highest throughput
• Normal-length outputs

**Long-context coding & refactors**
DeepSeek-Coder-V2-Lite Q8_0 @ 4096

• Slower, but stable
• Predictable coder-style outputs

**Code review & explanation**
Qwen3-4B Kimi-K2 Thinking @ 4096

# Source & Interactive Resources

GitHub Repository: https://github.com/emilyg888/local-llm-benchmarks
Interactive Charts: https://emilyg888.github.io/local-llm-benchmarks/

*All benchmarks are hardware-dependent and intended for comparative evaluation only.*