Technical Report — AI-Powered GitHub Analysis (Updated Benchmarks)

Author: umeshrai01
Date: 16 February 2026

Model Benchmarking:
Notes: Attempts were made to benchmark local Ollama models via CLI.
Some runs produced outputs; timing metrics were not fully available due to
local Ollama server issues (bind/address or service availability).

Model: llama3.2
Sample output (truncated):

```
MLX: Failed to load symbol: mlx_metal_device_info [?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25
```

Model: llama3
Sample output (truncated):

```
MLX: Failed to load symbol: mlx_metal_device_info [?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25
```

Timing:
- Wall-clock timings: Not available for all runs due to server connection issues.
- Recommendation: run the following commands locally to capture exact timings:

```
  /usr/bin/time -p ollama run llama3.2 "<PROMPT>" > out_llama3.2.txt 2>&1
  /usr/bin/time -p ollama run phi-3-mini "<PROMPT>" > out_phi3.txt 2>&1
```

Appendix — Full Captured Outputs (cleaned)

```
--- llama3.2 output ---
MLX: Failed to load symbol: mlx_metal_device_info
[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?

total duration:       6.441595792s
load duration:        4.788081375s
prompt eval count:    48 token(s)
prompt eval duration: 221.611166ms
prompt eval rate:     216.60 tokens/s
eval count:           55 token(s)
eval duration:        1.400723663s
eval rate:            39.27 tokens/s
[?25l[?25h

--- llama3 output ---
MLX: Failed to load symbol: mlx_metal_device_info
[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?2026l[?2026h[?25l [?25h[?

total duration:       15.834643625s
load duration:        11.795091416s
prompt eval count:    33 token(s)
prompt eval duration: 990.861625ms
prompt eval rate:     33.30 tokens/s
eval count:           44 token(s)
eval duration:        2.862285707s
eval rate:            15.37 tokens/s
[?25l[?25h
```