Inferenz Benchmarks

2026-04-18 vLLM mit cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit

vLLM optionen:

```
--model cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit
--tensor-parallel-size 2
--max-model-len 65536
--gpu-memory-utilization 0.85
--enable-prefix-caching
--reasoning-parser qwen3
--enable-auto-tool-choice
--tool-call-parser qwen3_coder
--max-num-seqs 32
--speculative-config '{"method":"qwen3_next_mtp","num_speculative_tokens":2}'
```

Benchmark mit `uvx llama-benchy --base-url http://cogito.buero.ping.de:8000/v1 --depth 2000 32768 63000`

model	test	t/s	peak t/s	ttfr (ms)	est_ppt (ms)	e2e_ttft (ms)
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d2000	5463.38 ± 111.87		748.82 ± 14.93	741.48 ± 14.93	748.93 ± 14.93
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d2000	103.13 ± 22.06	112.49 ± 24.41			
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d32768	5178.25 ± 25.55		6731.33 ± 33.06	6724.00 ± 33.06	6731.41 ± 33.05
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d32768	25.65 ± 1.43	27.93 ± 1.52			
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	pp2048 @ d63000	4534.72 ± 42.10		14353.15 ± 133.93	14345.82 ± 133.93	14353.26 ± 133.94
cyankiwi/Qwen3.6-35B-A3B-AWQ-4bit	tg32 @ d63000	12.85 ± 3.50	14.45 ± 3.21			

Plan: P2P einschalten, da geht noch mehr...

Revision #3
Created 2026-04-18 16:38:47 UTC by Sven Neuhaus
Updated 2026-04-19 10:07:45 UTC by Sven Neuhaus