LLM

- https://collabnix.com/best-ollama-models-in-2025-complete-performance-comparison/

For Production Deployment:

- Primary Choice: DeepSeek-R1 32B for reasoning-heavy applications
- Coding Tasks: Qwen2.5-Coder 7B for optimal balance of capability and efficiency
- General Purpose: Llama 3.3 70B for maximum versatility
- Edge Computing: Phi-4 14B for resource-constrained environments

Optimization Strategies:

- Always enable **Flash Attention** and KV-cache quantization
- Use **Q4_K_M** quantization for production deployments
- Implement caching for repeated queries
- Monitor GPU memory usage and implement automatic model swapping
- Use load balancing for high-throughput applications

Hardware	Llama 3.3 8B (tokens/sec)	Llama 3.3 70B (tokens/sec)	Llama 3.2
RTX 4090	89.2	12.1	
RTX 3090	67.4	8.3	
A100 40GB	156.7	45.2	
M3 Max 128GB	34.8	4.2	
Strix Halo 128GB ollama		5.1	85.02
Strix Halo 128GB llama.cpp			90
RTX 3060			131.76

ROCM

model	capabilities	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
llama3.2	completion tools	“3.2B”	131072	“Q4*K*M”	52.78	1957.30
qwen-strixhalo	completion tools	“30.5B”	262144	“Q4*K*M”	53.54	1056.37
qwen3-coder	completion tools	“30.5B”	262144	“Q4*K*M”	52.10	776.55
qwen3:30b-a3b	completion tools thinking	“30.5B”	262144	“Q4*K*M”	50.19	803.06
gpt-oss:20b	completion tools thinking	“20.9B”	131072	“MXFP4”	45.37	519.90
glm-4.7-flash	completion tools thinking	“29.9B”	202752	“Q4*K*M”	41.54	470.09

model	capabilities	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
qwen3:8b	completion tools thinking	“8.2B”	40960	“Q4*K*M”	32.68	890.98
qwen3-coder-next	completion tools	“79.7B”	262144	“Q4*K*M”	33.06	380.21
qwen2.5-coder:14b-instruct-q4*K*M	completion tools insert	“14.8B”	32768	“Q4*K*M”	17.25	527.74
gemma4:latest	completion vision audio tools thinking	“8.0B”	131072	“Q4*K*M”	50.15	1704.89
gemma4:e2b	completion vision audio tools thinking	“5.1B”	131072	“Q4_K_M”	83.07	2799.72

NVIDIA GeForce RTX 3060

model	capabilities	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
gemma4:e2b	completion vision audio tools thinking	“5.1B”	131072	“Q4_K_M”	102.44	4202.89

VULKAN

model	capabilities	size	context	quantization	eval rate [token/s]	prompt eval rate [token/s]
qwen3-coder	completion tools	“30.5B”	262144	“Q4*K*M”	54.03	805.43
llama3.2	completion tools	“3.2B”	131072	“Q4_K_M”	52.54	1838.82
gpt-oss:20b	completion tools thinking	“20.9B”	131072	“MXFP4”	43.36	475.60

ollama model

```
FROM qwen3-coder

# STRIX HALO AGENTIC TUNING
PARAMETER num_ctx 128000
PARAMETER num_batch 1024
PARAMETER num_predict 4096

SYSTEM """
You are a Strix Halo Optimized Coding Agent.
Always use asynchronous patterns and favor memory-efficient algorithms.
"""
```