LLMScan.online

AI Model Security Scan Report

B

Score 75/100

Report summary

GGUF v3 model | architecture: qwen35 | size: 9B | quantization: Q4_K | 427 tensors | 5 warnings / 4 informational

Model

Filename	`Nyx-RP-9B-Instruct-2608-v1-OBLITERATED.i1-Q4_K_S.gguf`
Format	`gguf`
Architecture	`qwen35`
Size	`5.0 GB`
SHA-256	`60cccfa9459171215fdd6e2ebff6bf51e26724cd0575c803c552d3fdc71f2694`
Scan mode	`static-deep`
Scanned at	`2026-09-15T04:54:45+00:00Z`
Report ID	`617210fd6f224ef1b69ae2e907d30628`

Findings (9)

SEVERITY	CATEGORY	DETAIL
MEDIUM	Metadata Security	**Complex agentic chat template — context-dependent risk** Template uses tool_calls, tool responses, namespace() — the full pattern for agentic pipelines. These are legitimate HuggingFace constructs but represent the highest-complexity template class. Risk is context-dependent: negligible for static inference, meaningful in autonomous-agent deployments where user-controlled content can reach the template renderer. **Recommendation:** Review template before use in agentic pipelines. Use sandboxed Jinja2 rendering and compare against the upstream author's reference template.
MEDIUM	Metadata Security	**External URLs embedded in 3 KV metadata field(s)** Fields: ['general.base_model.0.repo_url', 'general.url', 'general.source.url'] **Recommendation:** URLs in model metadata may be used for tracking or exfiltration when metadata is rendered; do not auto-fetch
LOW	Supply Chain	**No license field in GGUF metadata (general.license)** Absence of license information makes it impossible to determine usage rights **Recommendation:** Add general.license to model metadata before distributing; use an SPDX identifier
LOW	Supply Chain	**Fine-tuned model without traceable base_model field** Fine-tunes should declare the base model via general.base_model.0 for supply-chain transparency **Recommendation:** Add general.base_model.0 with the HuggingFace repo of the base model
LOW	Statistical Anomaly	**1 weight tensor(s) are >20x the median size: output.weight** Extreme outlier tensors may contain embedded payloads or crafted weight matrices **Recommendation:** Inspect outlier tensor names and compare against reference architecture
INFO	Architecture	**Grouped Query Attention (GQA) detected** head_count=16, head_count_kv=4 — requires GQA-aware runtime

		Recommendation: Ensure your runtime (llama.cpp >= b1500, Ollama >= 0.1.30) supports GQA
INFO	Architecture	**Selective F32 tensors in hybrid/shortconv layers (24 tensors across 24 block(s))** F32 weight tensors in conv1d, ssm layers of a Q4_K-quantized model. Largest: 131,072 B. This is expected in hybrid SSM/shortconv architectures (LFM2, Jamba, Mamba): small conv kernels and state-space projections are kept at full precision because they are numerically sensitive and do not benefit from quantization. Legitimately-F32 tensors (norm/bias/rope/embd): 153. **Recommendation:** Cross-check against the upstream model checkpoint to confirm these tensors are present in the original. Isolated or large F32 tensors in attention/FFN layers would be more suspicious.
INFO	Embedding Size	**Large embedding tensor 'token_embd.weight': 572,129,280 bytes** vocab=248,320 x emb_dim=4,096 x 0.5625 B/elem -> expected 572,129,280 B — size is consistent with architecture **Recommendation:** No action needed; large embeddings are expected for large-vocabulary models
INFO	Quantization	**Model is quantized: Q4_K, Q5_K, Q6_K** Quantization reduces size and inference cost at the expense of some accuracy **Recommendation:** Verify the quantization level is appropriate for your use case (Q4_K_M is a common balanced choice)

Runtime compatibility

RUNTIME	SUPPORTED	NOTES
llama.cpp	No	GGUF v3 container: supported | quant Q4_K: supported | 256k context: requires sufficient VRAM/RAM and --ctx-size flag — architecture 'qwen35' is uncommon; check runtime release notes for explicit support
Ollama	No	GGUF via llama.cpp backend: container and quantization supported — architecture 'qwen35' is uncommon; check runtime release notes for explicit support
LM Studio	No	GGUF container and quantization supported; GPU/CPU inference — architecture 'qwen35' is uncommon; check runtime release notes for explicit support
vLLM	No	GGUF not supported; requires safetensors format
HuggingFace Transformers	No	GGUF not supported; requires safetensors or PyTorch checkpoint

Generated 2026-09-21 05:55 UTC by LLMScan.online (engine v2.0.0, report schema v1.0). Full interactive report: https://llmscan.online/report/617210fd6f224ef1b69ae2e907d30628 — Static analysis only; not a guarantee of safety. See llmscan.online/docs/how-it-works for methodology.