LLMScan.online
AI Model Security Scan Report

B
Score 83/100

Report summary

GGUF v3 model | architecture: phi3 | size: mini | quantization: Q4_K | 196 tensors | 4 warnings / 3 informational

Model

Filename	`phi-4-mini-instruct.Q4_K_M.gguf`
Format	`gguf`
Architecture	`phi3`
Size	`2.3 GB`
SHA-256	`88c00229914083cd112853aab84ed51b87bdf6b9ce42f532d8c85c7c63b1730a`
Scan mode	`static-deep`
Scanned at	`2026-10-07T08:42:45+00:00Z`
Report ID	`078a5c740d4f494fb4233a4e7eaa432b`

Findings (7)

SEVERITY	CATEGORY	DETAIL
MEDIUM	Metadata Security	**External URLs embedded in 1 KV metadata field(s)** Fields: ['general.repo_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	**No embedded provenance metadata** None of general.author, general.source.url, general.base_model.0, or general.source.huggingface.repository are embedded in the GGUF file. Without any origin signal the model cannot be traced to a trusted source or compared against a reference checkpoint for tampering detection. For files uploaded directly, provenance is especially critical because the scanner cannot infer repository origin from the file alone. **Recommendation:** Add at least one of: general.author, general.source.huggingface.repository, or general.source.url
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
INFO	Architecture	**Grouped Query Attention (GQA) detected** head_count=24, head_count_kv=8 — requires GQA-aware runtime **Recommendation:** Ensure your runtime (llama.cpp >= b1500, Ollama >= 0.1.30) supports GQA
INFO	Embedding Size	**Large embedding tensor 'token_embd.weight': 504,161,280 bytes** vocab=200,064 x emb_dim=3,072 x 0.8203125 B/elem -> expected 504,161,280 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	Yes	GGUF v3 container: supported | quant Q4_K: supported | 128k context: requires sufficient VRAM/RAM and --ctx-size flag
Ollama	Yes	GGUF via llama.cpp backend: container and quantization supported
LM Studio	Yes	GGUF container and quantization supported; GPU/CPU inference
vLLM	No	GGUF not supported; requires safetensors format
HuggingFace Transformers	No	GGUF not supported; requires safetensors or PyTorch checkpoint

Generated 2026-11-06 03:37 UTC by LLMScan.online (engine v2.0.0, report schema v1.0). Full interactive report: https://llmscan.online/report/078a5c740d4f494fb4233a4e7eaa432b — Static analysis only; not a guarantee of safety. See llmscan.online/docs/how-it-works for methodology.