

Report summary

GGUF v3 model | architecture: llama | quantization: F32 | 9 tensors | 3 warnings

Model

Filename	vocab-size-mismatch.gguf
Format	gguf
Architecture	llama
Size	5.2 KB
SHA-256	9e7eff45b328d1d03a42993cdec05ec80af0da2d3e8adb6bbd0b145c0a964505
Scan mode	static-deep
Scanned at	2026-08-10T20:38:48+00:00Z
Report ID	ae5bddae7b5141b6b71427d7b86a015c

Findings (3)

SEVERITY	CATEGORY	DETAIL
MEDIUM	Statistical Anomaly	Embedding tensor 'token_embd.weight' inferred vocab=32 != KV vocab_size=999 Mismatch between embedding dimensions and KV metadata vocab_size may indicate a tampered, corrupted, or mismatched tokenizer/model pair Recommendation: Verify against the original model repository
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

Runtime compatibility

RUNTIME	SUPPORTED	NOTES
llama.cpp	Yes	GGUF v3 container: supported quant F32: supported
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 2026-08-21 06:52 UTC by LLMScan.online (engine v2.0.0, report schema v1.0). Full interactive report: <https://llmscan.online/report/ae5bddae7b5141b6b71427d7b86a015c> — Static analysis only; not a guarantee of safety. See llmscan.online/docs/how-it-works for methodology.