LLMScan.online

AI Model Security Scan Report

B

Score 81/100

Report summary

GGUF v3 model | architecture: gemma4 | size: 7.5B | quantization: Q6_K | 720 tensors | 3 warnings / 5 informational

Model

Filename	`gemma-4-E4B-it-SDFT_Heretic_RP.i1-IQ3_M.gguf`
Format	`gguf`
Architecture	`gemma4`
Size	`4.4 GB`
SHA-256	`434ed4a3938649780c23486849e2d62138d681e3628aa9396a1c83b0d5821f49`
Scan mode	`static-deep`
Scanned at	`2026-09-15T05:25:04+00:00Z`
Report ID	`1636e2a0e30f430bb65e14f4e64b35fa`

Findings (8)

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	Statistical Anomaly	**134 weight tensor(s) are >20x the median size: per_layer_model_proj.weight, blk.0.ffn_down.weight, blk.0.ffn_gate.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=8, head_count_kv=2 — requires GQA-aware runtime **Recommendation:** Ensure your runtime (llama.cpp >= b1500, Ollama >= 0.1.30) supports GQA
INFO	Embedding Size	**Large embedding tensor 'per_layer_token_embd.weight': 2,312,110,080 bytes** vocab=262,144 x emb_dim=10,752 x 0.8203125 B/elem -> expected 2,312,110,080 B — size is consistent with architecture **Recommendation:** No action needed; large embeddings are expected for large-vocabulary models

INFO	Embedding Size	**Large embedding tensor 'token_embd.weight': 550,502,400 bytes** vocab=262,144 x emb_dim=2,560 x 0.8203125 B/elem -> expected 550,502,400 B — size is consistent with architecture **Recommendation:** No action needed; large embeddings are expected for large-vocabulary models
INFO	Quantization	**Model is quantized: IQ3_S, Q4_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)
INFO	Quantization	**Declared file_type (IQ3_S) does not match effective tensor dtypes (Q6_K)** The GGUF general.file_type field declares 'IQ3_S' but the actual tensors in this file are 'Q6_K'. This is common for multi-file model packages where the projector/adapter GGUF inherits the main model's file_type metadata. The effective dtype shown in the report reflects actual tensor storage. **Recommendation:** No action required if this is a companion file (mmproj, adapter). For standalone models, this may indicate incorrect GGUF export.

Runtime compatibility

RUNTIME	SUPPORTED	NOTES
llama.cpp	No	GGUF v3 container: supported | quant Q6_K: supported | 128k context: requires sufficient VRAM/RAM and --ctx-size flag — architecture 'gemma4' is uncommon; check runtime release notes for explicit support
Ollama	No	GGUF via llama.cpp backend: container and quantization supported — architecture 'gemma4' is uncommon; check runtime release notes for explicit support
LM Studio	No	GGUF container and quantization supported; GPU/CPU inference — architecture 'gemma4' 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 05:48 UTC by LLMScan.online (engine v2.0.0, report schema v1.0). Full interactive report: https://llmscan.online/report/1636e2a0e30f430bb65e14f4e64b35fa — Static analysis only; not a guarantee of safety. See llmscan.online/docs/how-it-works for methodology.