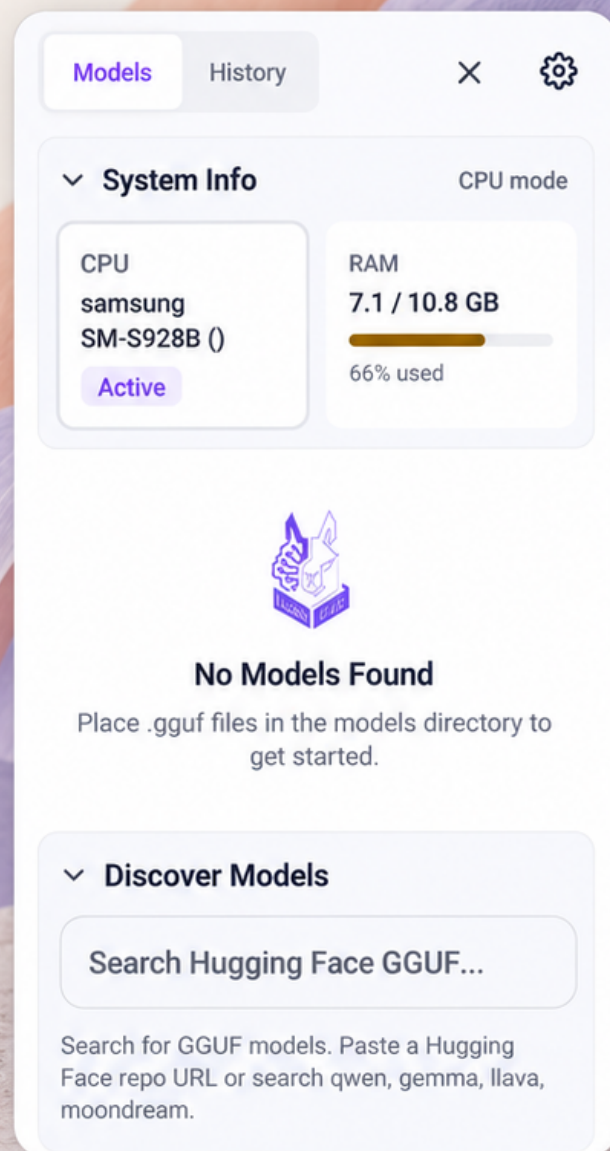


LlamaBox

Offline AI chat for Android

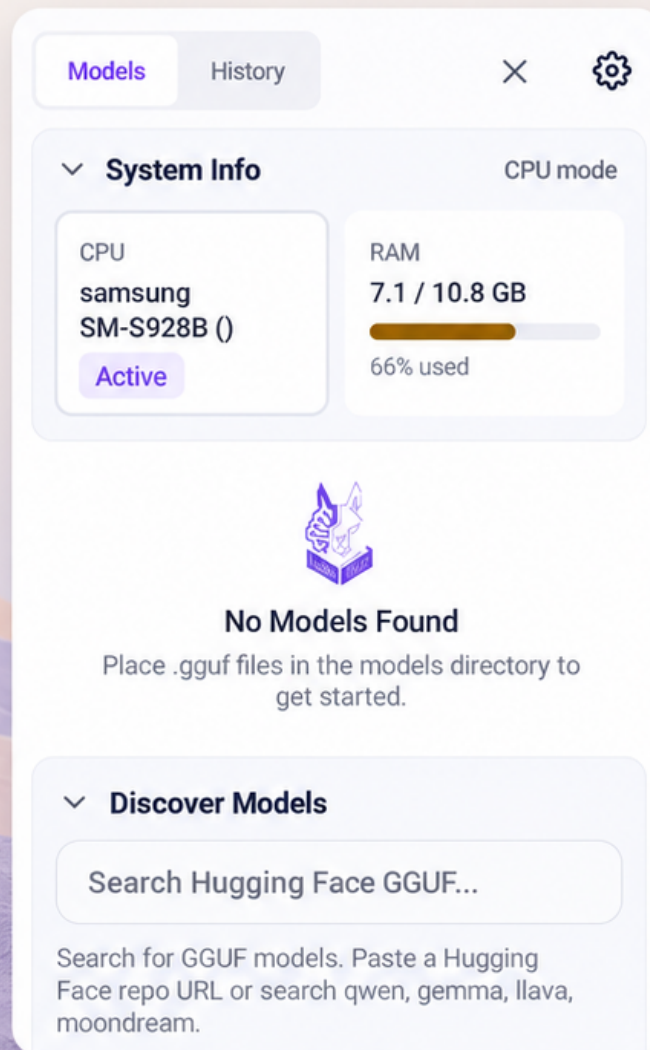
Check storage. Then check **memory**.

A model download uses storage.
Running it also needs working memory,
so check both before choosing.



See what your phone **is using.**

The sidebar shows device details and current RAM usage.
Check it before choosing a model to load.



LlamaBox

Offline AI chat for Android

Start small. Learn from the load.

Try a smaller compatible model first.
If loading fails, revisit compatibility,
available memory and the model's
requirements.

Available on Google Play



Models History 

Discover Models

Search for GGUF models. Paste a Hugging Face repo URL or search qwen, gemma, llava, moondream.

Curated starters

Qwen2.5 0.5B Instruct
Q4_K_M 0.5B ~491 MB
Fast, compact multilingual model. Great for everyday tasks on low-end devices.
[Download \(~491 MB\)](#)

SmolLM2 360M Instruct
Q8_0 0.36B ~386 MB
Ultra-small model trained for instruction following. Best quality for its size.
[Download \(~386 MB\)](#)

SmolLM2 1.7B Instruct
Q4_K_M 1.7B ~1060 MB
Fast mid-size model with strong reasoning. Best speed/quality trade-off f...