Package: llamacppR (via r-universe)

May 11, 2026

Type Package

Title Ellmer-Native llama.cpp Chats for R

Version 0.1.2

Description Provides an ellmer-style chat interface backed by native llama.cpp inference. The package vendors llama.cpp, exposes a chat_llamacpp() constructor for local GGUF models, supports token streaming, basic tool-calling loops, and helpers for downloading a curated default model.

License MIT + file LICENSE

Encoding UTF-8

LazyData false

SystemRequirements CMake, C++17

Imports cli, coro, curl, ellmer (>= 0.4.0), jsonlite, Rcpp, R6

LinkingTo Rcpp

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

RoxygenNote 7.3.3

Config/pak/sysreqs cmake libssl-dev

Repository https://arkraieski.r-universe.dev

Date/Publication 2026-04-11 05:43:13 UTC

RemoteUrl https://github.com/arkraieski/llamacppR

RemoteRef HEAD

RemoteSha 6b7c11a864985875e97c0421d46f2ecda033551f

Contents

`chat_llamacpp` *Create an ellmer-style llama.cpp chat*

Description

Creates a local chat object backed by native `llama.cpp` inference while following the `ellmer` chat API style.

Usage

```
chat_llamacpp(
  system_prompt = NULL,
  model,
  seed = NULL,
  params = ellmer::params(),
  echo = c("none", "output", "all"),
  n_ctx = 2048L,
  n_batch = n_ctx,
  n_threads = 0L,
  n_gpu_layers = 0L
)
```

Arguments

`system_prompt`	Optional system prompt.
`model`	Path to a local GGUF model file.
`seed`	Optional seed forwarded to llama.cpp sampling.
`params`	An `ellmer::params()` list.
`echo`	Whether to echo generated output.
`n_ctx`	Context size.
`n_batch`	Batch size used for prompt evaluation.
`n_threads`	CPU threads used by llama.cpp.
`n_gpu_layers`	Number of layers to offload to GPU when supported.

llamacpp_default_model_path

Get the cache path for a curated default model

Description

Returns the local cache path used by llamacppR for one of the curated default GGUF model presets.

Usage

```
llamacpp_default_model_path(model = c("3b", "0.5b", "starcoder", "deepseek"))
```

Arguments

model	Which curated default model path to return.

llamacpp_download_default_model

Download a curated default GGUF model

Description

Downloads a curated GGUF model from Hugging Face and returns the local path.

Usage

```
llamacpp_download_default_model(
  model = c("3b", "0.5b", "starcoder", "deepseek"),
  path = NULL,
  force = FALSE
)
```

Arguments

model	Which curated default model to download.
path	Destination path for the downloaded model.
force	Whether to overwrite an existing file.

llamacpp_download_model *Download a curated model preset*

Description

Downloads one of the curated model presets shipped with llamacppR.

Usage

```
llamacpp_download_model(
  model = c("qwen_3b", "qwen_0_5b", "starcoder", "deepseek"),
  path = NULL,
  force = FALSE
)
```

Arguments

model	Preset id or alias.
path	Destination path for the downloaded model.
force	Whether to overwrite an existing file.

llamacpp_is_gguf *Check whether a file is GGUF*

Description

Validates the magic bytes at the start of a file to determine whether it looks like a GGUF model.

Usage

```
llamacpp_is_gguf(path)
```

Arguments

path	Path to inspect.

llamacpp_list_models *List local GGUF models in the llamacppR cache*

Description

Lists GGUF files found in the local llamacppR cache directory and marks whether they match one of the curated default model presets.

Usage

```
llamacpp_list_models(path = llamacpp_cache_dir(), recursive = TRUE)
```

Arguments

path	Directory to scan for GGUF files.
recursive	Whether to scan subdirectories recursively.

llamacpp_model_info *Inspect a GGUF model through llama.cpp*

Description

Loads a GGUF model through native llama.cpp bindings and returns basic metadata.

Usage

```
llamacpp_model_info(
  model,
  n_ctx = 2048L,
  n_batch = n_ctx,
  n_threads = 0L,
  n_gpu_layers = 0L
)
```

Arguments

model	Path to a GGUF file.
n_ctx	Context size used when opening the model.
n_batch	Batch size used when opening the model.
n_threads	Number of CPU threads.
n_gpu_layers	Number of GPU layers to offload when supported.

llamacpp_model_path *Get the cache path for a curated model preset*

Description

Returns the local cache path used by llamacppR for a curated model preset.

Usage

```
llamacpp_model_path(model = c("qwen_3b", "qwen_0_5b", "starcoder", "deepseek"))
```

Arguments

model	Preset id or alias.

llamacpp_model_presets *List curated llama.cpp model presets*

Description

Returns the curated model catalog shipped with llamacppR, including stable preset ids, aliases, filenames, approximate sizes, and short descriptions.

Usage

```
llamacpp_model_presets()
```

llamacpp_unload *Unload a llama.cpp session*

Description

Explicitly releases the native llama.cpp model and context associated with a chat or session object.

Usage

```
llamacpp_unload(x)
```

Arguments

x	A chat object created by `chat_llamacpp()` or a native session pointer.

Index