

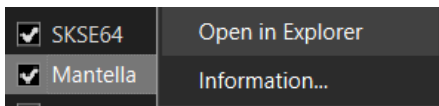
# Skyrim Installation

## Quick Start

1. Download and install the [required mods](#) using your mod manager.
2. Open the folder where your mod manager installed Mantella:

### ▼ Mod Organizer 2 Users

You can find your Mantella mod folder by right clicking the Mantella mod in the Mod Organizer 2 UI and selecting "Open in Explorer":



### ▼ Vortex Users

You can find your Mantella mod folder by going to `\Data\Mantella` from your Skyrim game folder.

Eg `C:\Games\Steam\steamapps\common\Skyrim Special Edition\Data\Mantella`.

3. Create a free account with [OpenRouter](#). Go to the "Keys" tab in the top right corner and generate a new key. Save the key value to the `GPT_SECRET_KEY.txt` file in your Mantella folder found in step 2.
4. Launch Skyrim. Mantella will open a window automatically when Skyrim is launched and will say "Waiting for player to select an NPC..." when it is ready (this might take a few minutes when launched for the first time). When you start the game, the Mantella spell will be added to your inventory. Cast this spell on an NPC to start a conversation.

#### Note

Mantella's default text-to-speech service, Piper, only supports English. For text-to-speech services that support other languages, see [here](#).

## How To Talk To Any NPC In Skyrim And Skyrim VR Using M...



### ► Upgrading from a previous Mantella version

If you get stuck anywhere in the installation process, please see [Issues Q&A](#) or reach out on [Discord](#).

# Requirements

## Skyrim

### Warning

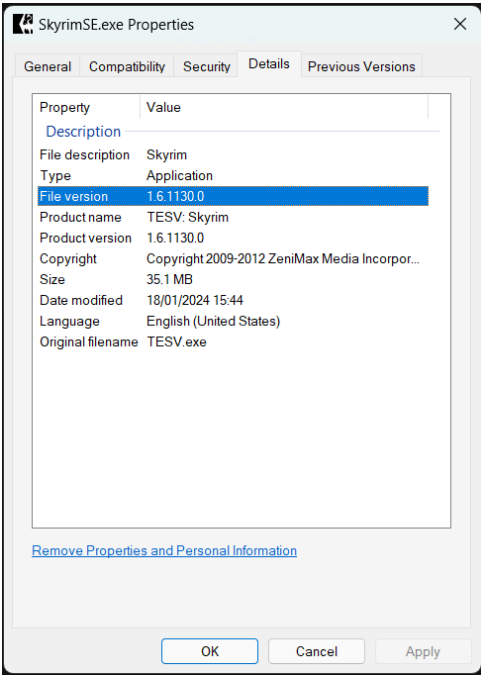
As Mantella accesses and writes to files within your Skyrim folder, it is unlikely to work correctly **if you have Skyrim stored in Program Files / (x86)**. Please ensure that you have Skyrim stored outside of this folder (eg `C:\Games\Steam`).

Steam does not allow to create a new Steam Game Library on the same disk. You can either move the whole Steam client outside [as described on this Steam Support page](#) or use [LostDragonist/steam-library-setup-tool](#) to allow multiple Steam Game Libraries on one disk.

# Required Skyrim Mods

Warning

Always ensure you are downloading the right version of each mod for your version of Skyrim. **This is the #1 reason for installation problems.** You can check your Skyrim version by right-clicking its exe file in your Skyrim folder and going to Properties -> Details -> File version. VR users can just download the VR version of each mod if available, or SE if not.



Please follow the installation instructions on each of the linked pages:

Mod	Notes
<a href="#">SKSE</a>	Once installed by following the included readme.txt, run SKSE instead of the Skyrim exe. Note that there is a separate VR version of SKSE.
<a href="#">VR Address Library for SKSEVR</a> or <a href="#">Address Library for SKSE</a> <a href="#">Plugins</a>	
<a href="#">FonixData File (Mod Manager Install)</a>	For lip sync generation.
<a href="#">UIExtensions</a>	If using text input instead of mic.
<a href="#">SkyUI</a>	To access Mantella's MCM.
<a href="#">Microsoft Visual C++ Redistributable 2015-2022</a>	
<a href="#">Mantella</a>	

## Optional Skyrim Mods

These mods aren't strictly necessary for Mantella to work, but they do greatly improve the experience.

Mod	Notes
<a href="#">No NPC Greetings</a>	Recommended so that Mantella voicelines are not interrupted by vanilla voicelines.
<a href="#">World Encounter Hostility Fix - Performance Version</a>	Stops certain NPCs from turning hostile when you cast the Mantella spell on them. This mod requires the <a href="#">Unofficial Skyrim Special Edition Patch (USSEP)</a> . Mantella needs to be loaded after USSEP in your load order.

## Compatibility

- Some users have reported that Skyrim crashes when Mantella is used with **Fuz Ro D'oh**. A possible fix is to disable and re-enable Fuz Ro D'oh.
- Mantella requires **Windows 10 / 11** (it is yet unconfirmed whether it works on Windows 7).
- Mantella needs to be loaded after the Unofficial Skyrim Special Edition Patch (**USSEP**) mod in your load order.

## Usage

### Mantella UI

The Mantella UI should open in your browser when the Mantella window starts, but if it does not, it can be accessed here: [http://localhost:4999/ui/?\\_\\_theme=dark](http://localhost:4999/ui/?__theme=dark).

#### Note

In order to access the Mantella UI, the Mantella window needs to be running.

### MCM

The Mantella MCM can be accessed in-game ( `SYSTEM -> MOD CONFIGURATION -> Mantella` ). Many options can be tweaked here, such as NPC actions and radiant conversations.



## Conversations

The Mantella spell & power should be added to your inventory under the Illusion category in the Magic menu once you install the mod. Conversations can be started by selecting an NPC with the spell, power, hotkey (enabled in the MCM menu), or via the in-game dialogue menu.

You can end a conversation by casting the Mantella End Conversation spell, or by simply saying / typing "goodbye" (recommended). NPCs will respond with "Safe travels" when conversations are ended via the latter method.

Once a conversation has started, you can add more NPCs to the conversation between each of your responses to start a group conversation:

```
Cast Mantella on NPC1 -> Say hi to NPC1 -> Cast Mantella on NPC2 -> Say hi to NPC1 & NPC2
etc
```

Radiant conversations can also be enabled in the MCM menu. These are conversations that are randomly started between idle NPCs. You can join the conversation by casting Mantella on either NPC.

## Text Input

Text input can be enabled by disabling mic input in Mantella's MCM menu.

## Restarting the Mantella window

If for whatever reason the Mantella window crashes or gets stuck, you can restart it by going to your `Mantella MCM -> Advanced -> Restart Mantella.exe`.

## Actions

By default, actions are disabled in the Mantella MCM. You can enable these in the MCM's `General` tab. If you're struggling to get an NPC to perform an action on their own, you can force these actions to occur by simply saying / typing the action you want to occur (eg "follow", "inventory", etc).

## Saved Data

When a conversation ends, a summary of the conversation is saved to a local text file for each NPC in the conversation. These summaries are then loaded the next time an NPC is spoken with. You can view / edit these summaries in your `My Games/Mantella/data/Skyrim/conversations/YourPlayerName1/` folder.

## Services

While the [quick start guide](#) is the fastest way to get set up, Mantella can be heavily tweaked to your liking and can connect to a variety of different services.

## Language Models (LLMs)

LLMs power the creation of responses by NPCs. There are a number of different LLMs to choose from, ranging from free local models to large externally hosted models.

### Note

Some smaller models may struggle to handle long term conversations and memory summarising.

If you have already followed the [quick start guide](#) you will have an OpenRouter account set up with Gemma 2 9B, a free and easy to use LLM. You can select from a variety of other LLMs hosted on OpenRouter (both free and paid) via the `Model` option in the `Large Language Model` tab of the [Mantella UI](#).

If you would prefer to use a different service, or host your own LLM locally, see the options below:

## API Models

### ▼ OpenAI

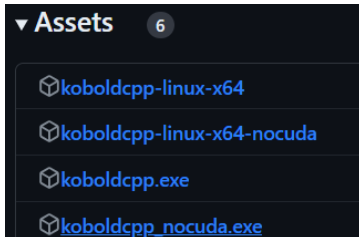
Copy your OpenAI secret API key (see [here](#) if you need help finding it (you will need to set up an account if you haven't already)) and paste into your `GPT_SECRET_KEY.txt` file (see the [quick start guide](#) to learn where `GPT_SECRET_KEY.txt` is stored). Note that you will need to set up your payment details for the API to work.

Once completed, in the `LLM Service` dropdown of the `Large Language Model` tab of the [Mantella UI](#) select "OpenAI". Update the list of model options in the `Model` setting and select an OpenAI model.

## Local Models

### ▼ koboldcpp

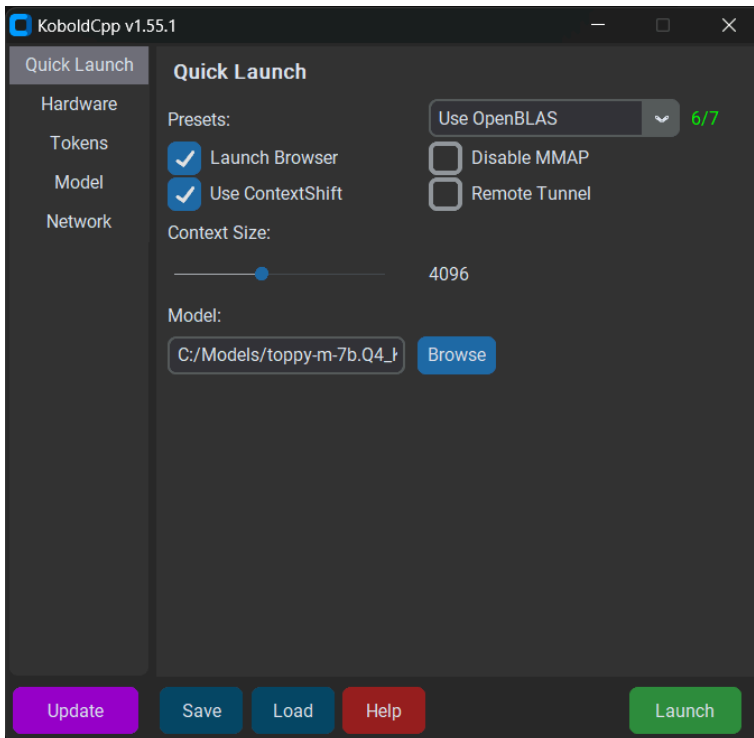
1. Install koboldcpp's latest release from [here](#). If you want to run koboldcpp on your CPU or otherwise do not have an NVIDIA GPU, download `koboldcpp_nocuda.exe` under "Assets". If you have an NVIDIA GPU with CUDA support, download `koboldcpp.exe` under "Assets".



2. Download a local model, such as `toppy-m-7b.Q4_K_S.gguf` from [here](#).



3. Run koboldcpp.exe. When presented with the launch window, drag the "Context Size" slider to 4096. Click the "Browse" button next to the "Model" field and select the model you downloaded. Click "Launch" in the bottom right corner.





**Optional**

Under the "Presets" drop down at the top, choose either Use CLBlast, or Use CuBlas (if using Cuda). You will then see a field for GPU Layers. If you want to use CPU only leave it at 0. If you want to use your GPU, you can experiment with how many "layers" to offload to your GPU based on your system.

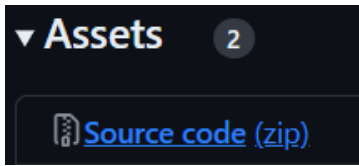
4. In the `LLM Service` dropdown of the `Large Language Model` tab of the [Mantella UI](#) select "KoboldCpp".

**Note**

Make sure koboldcpp is running when Mantella is running!

## ▼ text-generation-webui

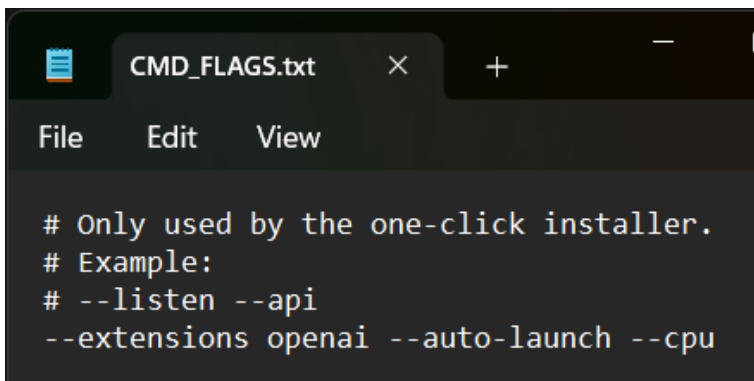
1. Install the latest text-generation-webui .zip from [here](#).



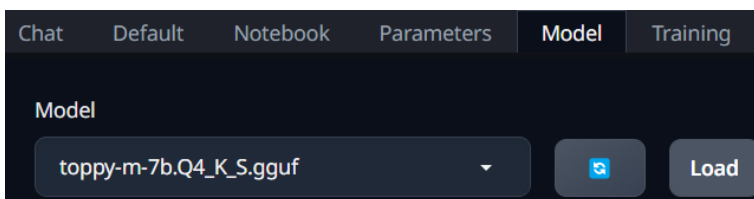
2. Place a local model into the `text-generation-webui\models` folder (to get started, you can download `topyy-m-7b.Q4_K_S.gguf` from [here](#)).



3. Paste the text “--extensions openai --auto-launch” (as well as “--cpu” for CPU users) into the installed folder’s CMD\_FLAGS.txt file.



4. Start text-generation-webui and wait for the UI to open in your web browser. Navigate to the “Model” tab, select your model from the drop-down list, and click “Load”.



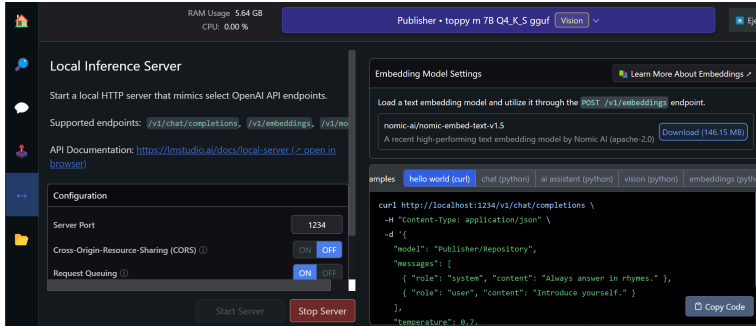
5. In the `LLM Service` dropdown of the `Large Language Model` tab of the [Mantella UI](#) select “textgenwebui”.

### Note

Make sure text-generation-webui is running when Mantella is running!

## ▼ LM Studio

1. Download and install LM Studio from [here](#).
2. Open LM Studio and download an LLM model.
3. Go to the Local Server tab, and in the drop-down at the top of the page select your model. If a Windows Firewall message pops up click Allow. Click the Start Server button.



4. In the `LLM Service` text box of the `Large Language Model` tab of the [Mantella UI](#), set the value to `http://localhost:1234/v1/`. This is the URL that LM Studio is running your model on. If everything works correctly you should see a message display in Mantella.exe saying "Running Mantella with local language model" when you start it.

### Note

Make sure LM Studio is running when Mantella is running!

## Other LLM Services

Mantella has the ability to support other language model services, although these services do need to support outputs in the OpenAI format (like text-generation-webui does via the "--extensions openai" option above).

## Vision

LLM vision can be enabled in the `vision` tab of the [Mantella UI](#). If the selected model in the `Large Language Model` tab has vision capabilities, a screenshot of your game window will be passed to the model every time you respond to an NPC.

Vision capabilities can also be run locally via koboldcpp. To get started, download the required [mmproj](#) file for your local LLM (as an example, [Toppo](#) requires `llama-7b-mmproj-v1.5-Q4_0.gguf`). In the koboldcpp launch window, under the `Model Files` tab, set the path to your mmproj file via the `LLaVa mmproj` setting. When koboldcpp is launched, and vision is enabled in the [Mantella UI](#), your local LLM will support vision.

## Text-to-Speech

By default, Mantella comes packaged with a fast and local TTS service called Piper (note: Piper only supports English). Mantella can also be run with xVASynth (local) or XTTS (local / online) to generate NPC voicelines. See the comparison below for more details:

Feature	Piper	xVASynth	XTTS
Fastest	✓ *		
Easiest setup	✓		
Highest quality			✓
Greatest emotional range		✓	
Support for modded NPCs		✓	✓
Easy to create custom voice models			✓
Local support	✓	✓	✓
API support			✓
Can run on a second PC			✓

\*XTTS may generate faster responses when running externally.

### ▼ xVASynth

1. Download xVASynth via [Steam](#) (preferred) or [Nexus](#). Do not store xVASynth in your Skyrim folder.
2. Download the Skyrim voice models. You can either download all models via a torrent, via the xVASynth UI if you have Nexus Premium, or manually via the Nexus Mods page:

#### ► xVASynth Model Installation Options

3. (Optional) Download the xVASynth DeepMoji Plugin [here](#). It boosts the emotionality of xVASynth voice models to make them sound less robotic (only available in English).
4. Select xVASynth in the Text-to-Speech tab of the [Mantella UI](#) and set the path to your xVASynth folder.

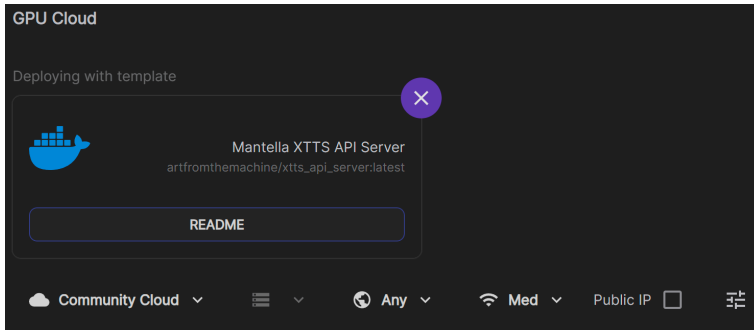
### ▼ XTTS

## ▼ Local

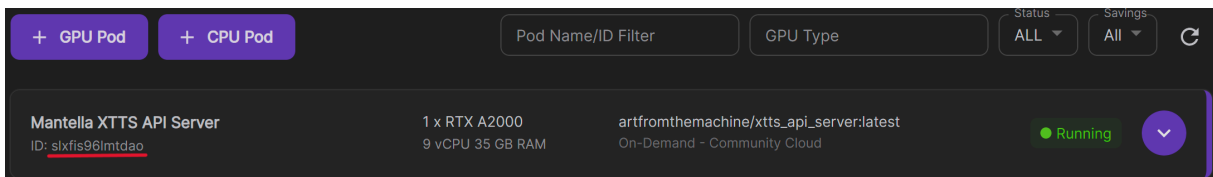
1. Download MrHaurrus's XTTS API server from [here](#) and unzip it.
2. Download the [voice models \(latents\) folder](#) for your desired language. Extract this folder into the same folder as `xtts-api-server-mantella.exe` above. In other words, you should have a folder called `latent_speaker_folder` in your XTTS folder.
3. Select XTTS in the Text-to-Speech tab of the [Mantella UI](#) and set the path to your XTTS folder.
4. (Optional) If you are using an NVIDIA GPU, the `XTTS DeepSpeed` setting can improve response times by 2-4x if you also have `XTTS Device` set to "cuda".

### ▼ External (from \$0.14/hr)

1. Make a RunPod account (<https://www.runpod.io/>) and add some credits (\$10 minimum).
2. Follow [this link](#) to open the Mantella XTTS Pod. Choose a GPU to run the Pod with (if in doubt, choose the cheapest). Note that cheaper GPUs are available if you change the "Secure Cloud" setting to "Community Cloud".



3. On the following two pages just click "Continue" and "Deploy" (you don't need to change any settings here).
4. Once the Pod is running, select XTTS in the Text-to-Speech tab of the [Mantella UI](#) and set `XTTS URL` to `https://{pod_id}-8020.proxy.runpod.net/`, with {pod-id} being the ID of the running Pod. Eg `https://a1b2c3d4qwerty-8020.proxy.runpod.net/`.



5. It takes a few minutes for the Pod to start up, you can check the progress in Logs -> Container Logs from the Pod's dropdown menu. It will say "Uvicorn running" when it is ready.
6. When you are finished using Mantella, make sure to delete the Pod so that you are no longer charged! It only takes a few clicks to set up a new Pod so don't feel guilty about deleting it.

## Whisper

### ▼ Whisper via your CPU is handled automatically. Open this section only if you like tinkering

guillaumekln's Faster-Whisper version of Whisper is used as Speech-To-Text engine by Mantella. The engine is **already part of the executable** and will download a chosen model automatically when launched. Uses a single CPU core by default when listening to the set

default Windows microphone. Alternatively text input can be enabled by setting `microphone_enabled = 0` within MantellaSoftware/config.ini.

It is reasonably fast even in CPU mode with the base model. Optionally, to use GPU/CUDA mode, some extra files are required, see [Faster Whisper documentation](#). Note that cuBLAS may already be part of the CUDA Toolkit, so you may only require the  `cudnn_###_infer64_8.dll`  files to be beside the Mantella executable. Afterwards enable `process_device = cuda` under `[Microphone]` in MantellaSoftware/config.ini.

---

Copyright © 2024, Art from the Machine  
Made with [Sphinx](#) and @pradyunsg's [Furo](#)