

App: Local Whisper Speech-To-Text (stt_whisper2)



- [Requirements](#)
 - [Installation](#)
 - [Supplying alternate models](#)
 - [Scaling](#)
 - [App store](#)
 - [Repository](#)
 - [Known Limitations](#)
-

The *stt_whisper2* app is one of the apps that provide Speech-To-Text functionality in Nextcloud and act as a media transcription backend for the [Nextcloud Assistant app](#), the *talk* app and [other apps making use of the core Speech-To-Text API](#). The *stt_whisper2* app specifically runs only open source models and does so entirely on-premises. Nextcloud can provide customer support upon request, please talk to your account manager for the possibilities.

This app supports input and output in languages other than English if the underlying model supports the language.

This app uses [faster-whisper](#) under the hood. Output quality will differ depending on which model you use, we recommend the following models:

- “
- OpenAI Whisper large v3 turbo (multilingual)
 - OpenAI Whisper medium.en (English only)

Whisper large v3 supports about ~100 languages and shows outstanding performance in ~10 of them. For more details see the [OpenAI Whisper paper](#)

Requirements

- Minimal Nextcloud version: 28
- This app is built as an External App and thus depends on AppAPI v2.3.0
- Nextcloud AIO is supported
- We currently support NVIDIA GPUs and x86_64 CPUs
- CUDA $\geq$ v12.2 on your host system
- GPU Sizing

“ ◦ A NVIDIA GPU with at least 4GB VRAM

- CPU Sizing

“ ◦ The more cores you have and the more powerful the CPU the better, we recommend 10-20 cores
◦ The app will hog all cores by default, so it is usually better to run it on a separate machine
◦ 4GB for the app

Installation

0. Make sure the [Nextcloud Assistant app](#) is installed
1. [Install AppAPI and setup a Deploy Demon](#)
2. Install the *stt_whisper2* "Local Speech-To-Text" ExApp via the "Apps" page in the Nextcloud web admin user interface

Supplying alternate models

This app allows supplying alternate models in the `/nc_app_stt_whisper2_data` directory of the docker container. You can use any [*faster-whisper* model by Systran on hugging face](#) in the following way:

1. git cloning the respective repository
2. Copying the folder with the git repository to `/nc_app_stt_whisper2_data` inside the docker container.

3. Restarting the Whisper ExApp
4. Selecting the respective model in the Nextcloud AI admin settings

Scaling

It is currently not possible to scale this app, we are working on this. Based on our calculations an instance has a rough capacity of 4h of transcription throughput per minute (measured with 8 CPU threads on an Intel(R) Xeon(R) Gold 6226R). It is unclear how close to real-world usage this number is, so we do appreciate real-world feedback on this.

App store

You can also find this app in our app store, where you can write a review:

https://apps.nextcloud.com/apps/stt_whisper2

Repository

You can find the app's code repository on GitHub where you can report bugs and contribute fixes and features: https://github.com/nextcloud/stt_whisper2

Nextcloud customers should file bugs directly with our customer support.

Known Limitations

- We currently do not support live transcription
- We currently only support languages supported by the underlying Whisper models
- The whisper models perform unevenly across languages, and may show lower accuracy on low-resource and/or low-discoverability languages or languages where there was less training data available. The models also exhibit disparate performance on different accents and dialects of particular languages, which may include higher word error rate across speakers of different genders, races, ages, or other demographic criteria.
- Language models are likely to generate false information and should thus only be used in situations that are not critical. It's recommended to only use AI at the beginning of a creation process and not at the end, so that outputs of AI serve as a draft for example and not as final product. Always check the output of language models before using it.
- Make sure to test the language model you are using it for whether it meets the use-case's quality requirements

- Language models notoriously have a high energy consumption, if you want to reduce load on your server you can choose smaller models or quantized models in exchange for lower accuracy
 - Customer support is available upon request, however we can't solve false or problematic output, most performance issues, or other problems caused by the underlying model. Support is thus limited only to bugs directly caused by the implementation of the app (connectors, API, front-end, AppAPI)
-

👤 #3

📧 cocobook-sync 👤	2024-03-18 18:45:19 UTC
📧 cocobook-sync 👤	2026-04-26 19:39:19 UTC