

# LOCAL LLM SPEEDRUN GUIDE




*Kobold% glitchless*

Updated April 14, 2024

Welcome to the local LLM speedrun guide! This guide will help you set up a LLM (large language model) on your computer as quickly and easily as possible, on both Windows and Linux. This guide assumes you have 8GB of RAM or more.































## 1. Download koboldcpp

▼ Assets 6

 koboldcpp-linux-x64 	376 MB	4 days ago
 koboldcpp-linux-x64-nocuda	55.1 MB	4 days ago
 koboldcpp.exe 	299 MB	4 days ago
 koboldcpp_nocuda.exe	37.6 MB	4 days ago
 Source code (zip)		3 days ago
 Source code (tar.gz)		3 days ago

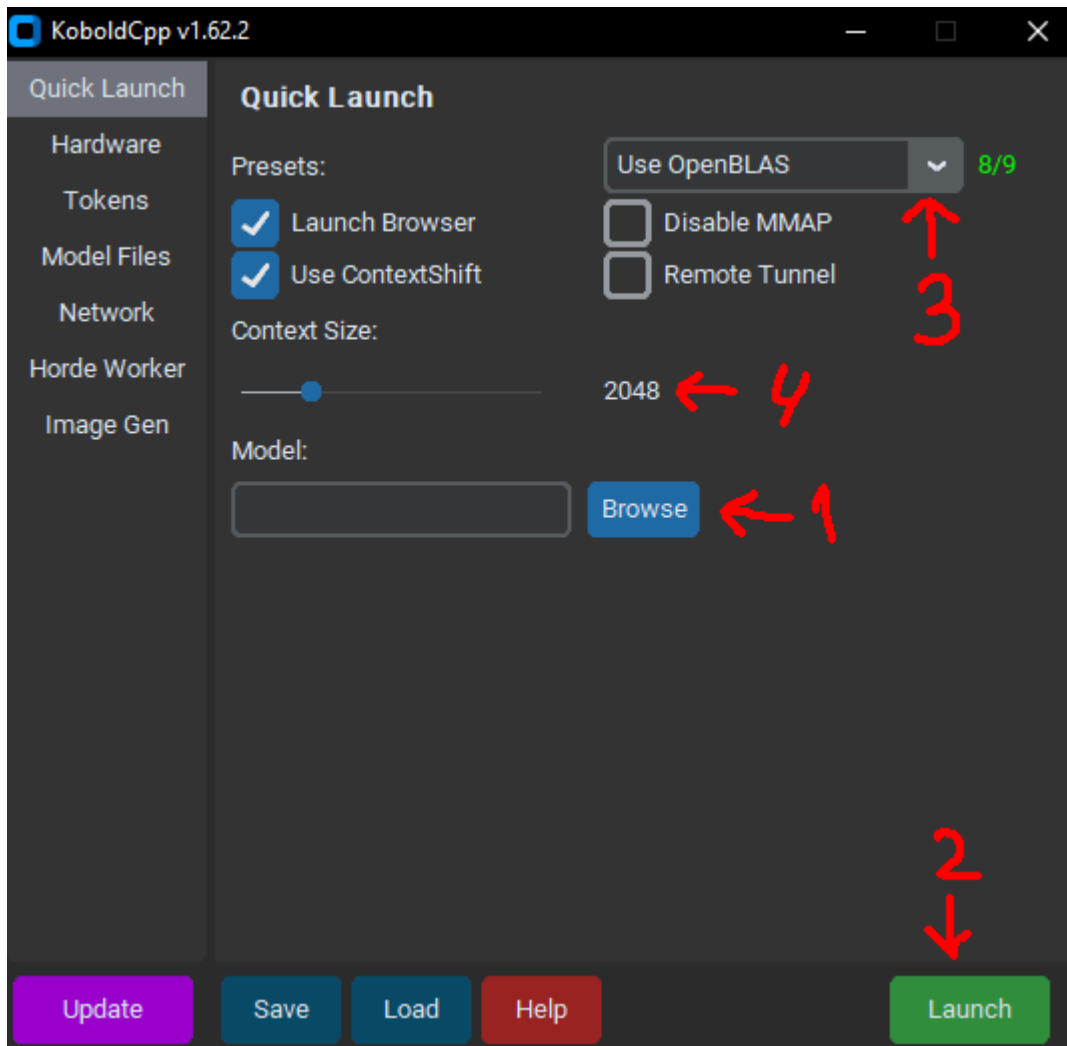
Go here <https://github.com/LostRuins/koboldcpp/releases> and download the latest version for your system. This is the LLM opener that you will be using. It can open models in GGUF format for you.

## 2. Get the model

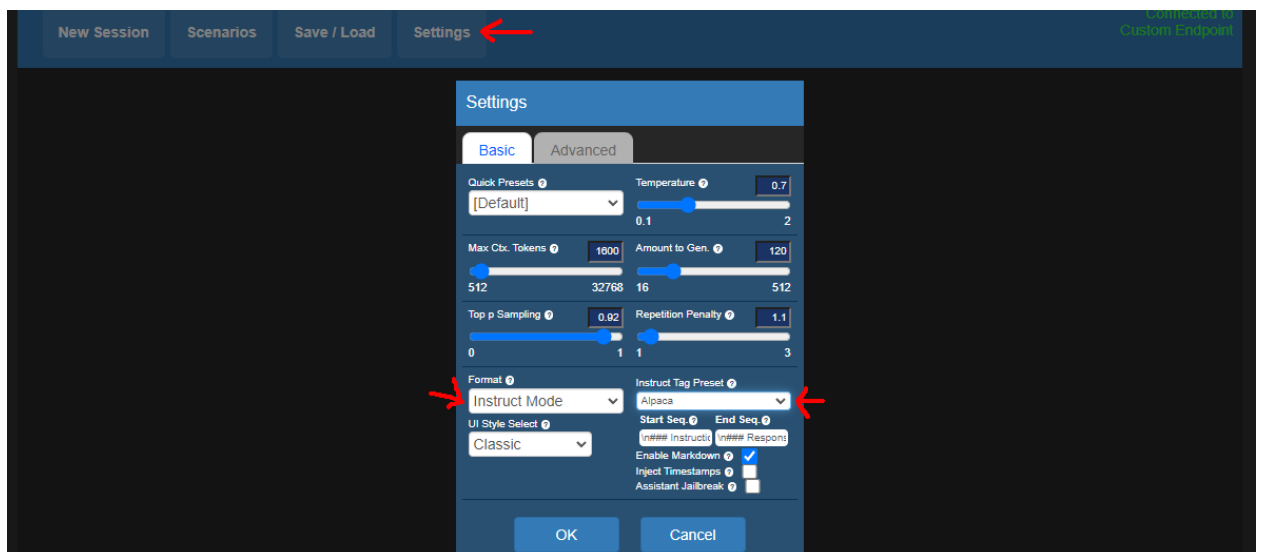
 Sao10K Update README.md 72faab7 <span>VERIFIED</span>			about 1 month ago
 .gitattributes	1.94 kB		Upload folder using huggingface_hub about 2 months ago
 Fimbulvetr-11B-v2-Test-14.q4_K_M.gguf 	6.46 GB 	 	Upload folder using huggingface_hub 2 months ago
 Fimbulvetr-11B-v2-Test-14.q5_K_M.gguf 	7.6 GB 		Upload folder using huggingface_hub 2 months ago
 Fimbulvetr-11B-v2-Test-14.q6_K_M.gguf 	8.81 GB 		Upload folder using huggingface_hub 2 months ago
 Fimbulvetr-11B-v2-Test-14.q8_0.gguf 	11.4 GB 		Upload folder using huggingface_hub 2 months ago
 Fimbulvetr-11B-v2.q3_K_S.gguf 	4.66 GB 		Upload folder using huggingface_hub about 2 months ago
 Fimbulvetr-11B-v2.q4_K_S.gguf 	6.12 GB 		Upload folder using huggingface_hub about 2 months ago
 README.md	1.74 kB		Update README.md about 1 month ago

Go here <https://huggingface.co/Sao10K/Fimbulvetr-11B-v2-GGUF/tree/main> and take q4\_K\_M gguf (first one). This is your model. It may not be the best, but it's not the worst.

### 3. Loading the model

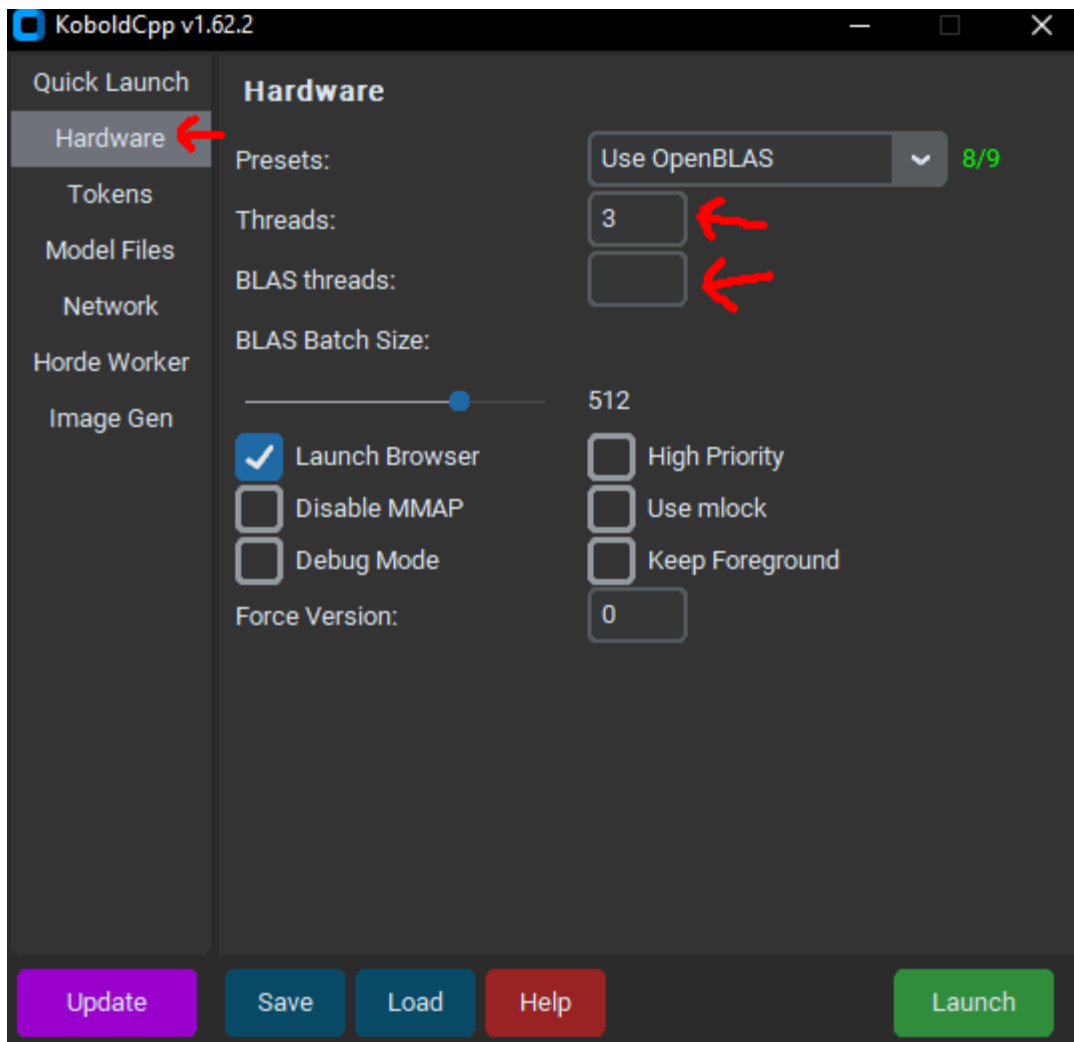


Now try opening koboldcpp and loading your model. Click on "Browse" (1) and pick your model file. Next, click "Launch" (2). Did it work?



Open your browser at localhost:5001 and try talking to your model using Instruct Mode with Alpaca format. Does it talk to you? Great! You're halfway there.

Now you can optimize. Open kobold launcher and change "Context Size" (4) to 4096. Also change "Max Ctx. Tokens" in browser UI. This will allow your model to remember more text. If you have NVIDIA change (3) to "cuBLAS", if you have AMD or Intel change to "Vulcan" or "CLBlast". This will make context processing faster.

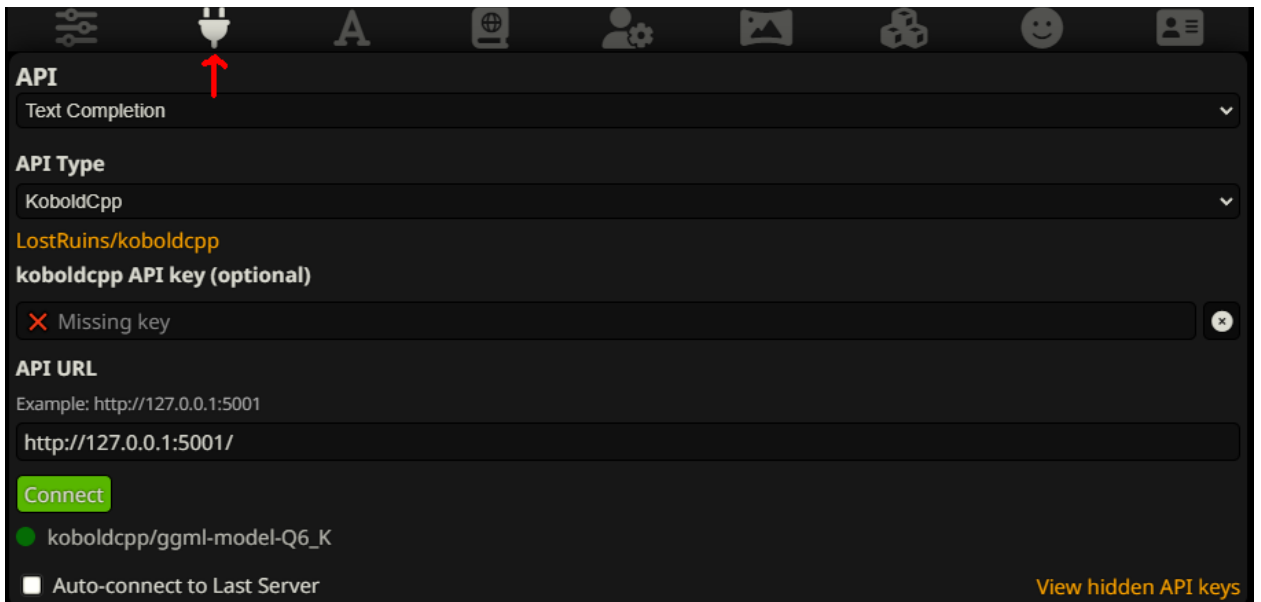


Go to "Hardware" in kobold launcher, set "Threads" to number of your physical cpu cores and "BLAS threads" to cores+1.

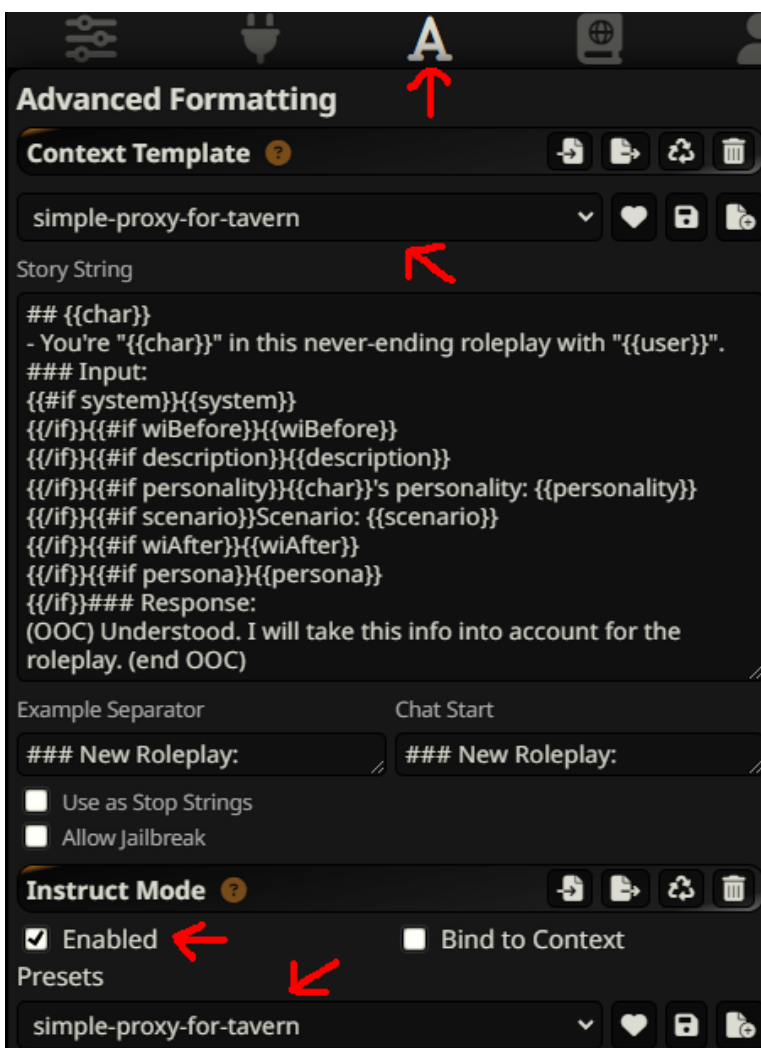
#### 4. (Optional) Setting up SillyTavern

Kobold webui is okay for simple textgen, but if you want something fancier you will need to install SillyTavern.

Just follow the steps here: <https://github.com/SillyTavern/SillyTavern?tab=readme-ov-file#-installation>



Open your SillyTavern, pick “Text Completion” and “KoboldCpp” under API with your localhost url. Now try talking to one of the included characters. Does it work? Great! Next go to presets and pick “simple-proxy-for-tavern”, this will make outputs a bit better. Also don’t forget to change max context like you did in koboldcpp.



## 5. (Optional) Getting better models

So, you have set up your model, but it feels dumb and you want something better? Find a benchmark that you like and pick a model based on that (Note: different models have different templates and context lengths, you will need to change templates in koboldcpp or sillytavern to the right one).

Usually you want to pick the biggest model that you can possibly load. As a rule of thumb, you will need (model file size + 15%) RAM to fully load the model with context.

Here are some benchmarks:

- [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)
  - Oldest, most gamed benchmark. Unreliable.
  - Tests on multiple choice questions.
- <https://huggingface.co/spaces/DontPlanToEnd/UGI-Leaderboard>
  - Test for LLM censorship.
  - Most uncensored does not necessarily mean the smartest model.
- <https://chat.lmsys.org/?leaderboard>
  - Leaderboard with votes from users.
  - Mostly tests for SFW one-liners.
  - Can be botted.
- <https://eqbench.com/>
  - Tests on multiple choice questions related to evaluating emotions.
  - Not as gamed as huggingface leaderboard.
- [https://huggingface.co/datasets/ChuckMcSneed/NeoEvalPlusN\\_benchmark](https://huggingface.co/datasets/ChuckMcSneed/NeoEvalPlusN_benchmark)
  - Tests for creative writing and obedience.
  - Low number of questions + strong human factor.
- [https://huggingface.co/datasets/ChuckMcSneed/WolframRavenwolfs\\_benchmark\\_results](https://huggingface.co/datasets/ChuckMcSneed/WolframRavenwolfs_benchmark_results)
  - Tests for ability of the model to answer 18 questions related to German data protection law.
  - Low number of questions + limited field.
- [https://ayumi.m8geil.de/erp4\\_chatlogs/index.html](https://ayumi.m8geil.de/erp4_chatlogs/index.html)
  - Tests for ERP abilities of the models.
  - Questionable quality.
- [https://tatsu-lab.github.io/alpaca\\_eval/](https://tatsu-lab.github.io/alpaca_eval/)
  - LLM evaluation with GPT4

If you don't care and just want me to pick something for you:

Note: you will need to find ggufs yourself

80B+:

- <https://huggingface.co/wolfram/miqu-1-120b>
- <https://huggingface.co/wolfram/miquiliz-120b-v2.0>
- <https://huggingface.co/CohereForAI/c4ai-command-r-plus>
- <https://huggingface.co/alpindale/goliath-120b>
  - <https://huggingface.co/ChuckMcSneed/WinterGoliath-123b>
  - <https://huggingface.co/ChuckMcSneed/Premerge-EX-EX-123B>
  - <https://huggingface.co/ChuckMcSneed/Premerge-XE-XE-123B>

~70B

- <https://huggingface.co/miqudev/miqu-1-70b>
- <https://huggingface.co/Qwen/Qwen1.5-72B-Chat>

30-50B

- <https://huggingface.co/CohereForAI/c4ai-command-r-v01>
- <https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>

<30B

I don't follow the small model meta closely. Bigger quant of

<https://huggingface.co/Sao10K/Fimbulvetr-11B-v2-GGUF/tree/main?>