

Guia de instalação: colibri

Roda modelos de IA abertos **gigantes** num computador comum: deixa o miolo do modelo na memória e lê do disco **só a parte que cada palavra precisa**. Gratuito, código aberto (Apache-2.0), sem nuvem.

01 Antes de instalar

- Windows, Mac ou Linux com **8 GB de RAM no mínimo** (16 GB ou mais é melhor). Placa de vídeo é opcional.
- **22 GB livres no disco** para o menor modelo, e internet para baixar.
- Modelo gigante pede disco grande: o GLM-5.2 (744B) baixa **429 GB**. Um SSD NVMe rápido faz diferença: a velocidade do disco vira a velocidade da resposta.
- Não precisa criar conta nem usar nuvem.

In two sentences: why does colibri stream the experts from disk instead of loading the whole model?



Colibri streams model weights from disk to significantly reduce memory consumption, allowing larger models to run on hardware with limited VRAM or RAM. This approach enables inference on resource-constrained devices by loading only the necessary parameters on demand rather than loading the entire model into memory at once.

imagem: github.com/JustVugg/colibri

02 Instalar (um passo)

WINDOWS

- 1 Em github.com/JustVugg/colibri clique em **Code** → **Download ZIP** e descompacte.
- 2 Dê dois cliques em **START-HERE.bat**. Se faltar o Python, ele oferece instalar.

LINUX (UBUNTU / DEBIAN)

```
$ sudo apt install git python3 build-essential
$ git clone https://github.com/JustVugg/colibri
$ cd colibri && ./start-here.sh
```

MAC (COM HOMEBREW)

```
$ xcode-select --install
$ brew install libomp git python
$ git clone https://github.com/JustVugg/colibri
$ cd colibri && ./start-here.sh
```

03 Primeiro uso

- O instalador olha a sua máquina e **recomenda um modelo que cabe**: aperte Enter para aceitar.
- Ele baixa o modelo (pode interromper e continuar depois) e **abre o painel no navegador**: <http://127.0.0.1:8000>
- Da próxima vez, rode o **START-HERE.bat** ou o [./start-here.sh](#) de novo: abre direto, sem baixar.

! Cuidados

- **Modelo gigante é lento:** no notebook de 25 GB, o GLM-5.2 (744B) faz **0,05 a 0,1 token/s**, menos de uma palavra por segundo. Serve pra **tarefa em lote**, não pra conversa. Modelos menores (ex.: Qwen3.6-35B) fazem cerca de 6 tokens/s só na CPU.
- **Use só modelos de fonte confiável:** em agosto de 2026 o projeto publicou 8 avisos de segurança sobre arquivos de modelo e o servidor. Mantenha o colibri atualizado.
- O painel roda **só no seu computador** (`127.0.0.1`); não abra ele para a internet.
- **Cada modelo tem a própria licença** (o Qwen-Image-2.1, por exemplo, é só para uso não comercial).

04 Links oficiais

- `github.com/JustVugg/colibri`
- Versões para baixar: `github.com/JustVugg/colibri/releases`
- Site: `justvugg.github.io/colibri`

Quer colocar IA pra trabalhar na sua empresa?

Diagnóstico gratuito no link da bio de @stratiaintelligence.