■ Internship Application Manager

Ollama AI Setup Guide (Windows)

This guide will walk you through installing and setting up Ollama so you can use the AI Assistant in the Internship Application Manager. Follow each step carefully and you’ll be ready to go in minutes.

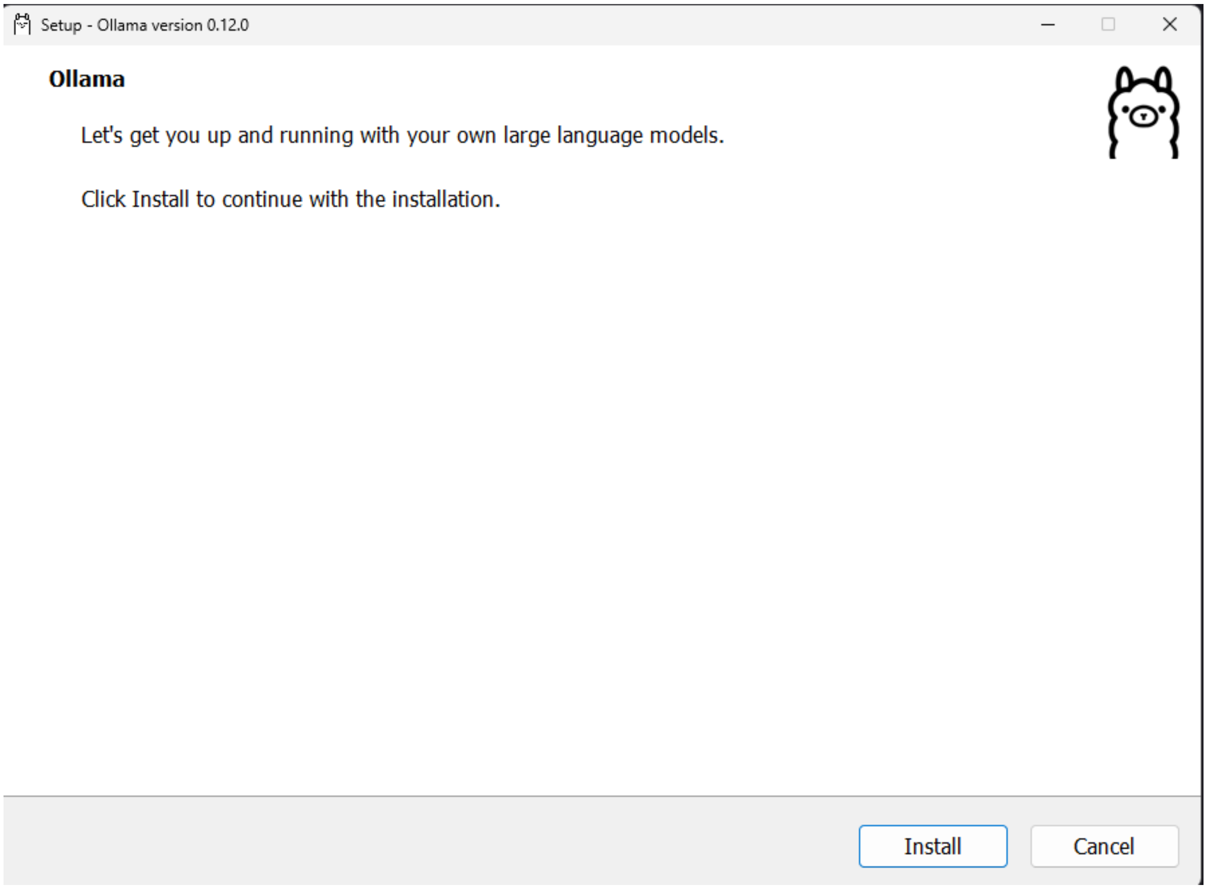
Prepared for: Project Team

Version 1.0

Step 1: Install Ollama

1. Download the Ollama installer for Windows from https://ollama.com/download

2. Run the installer and click **Install**:



Setup - Ollama version 0.12.0

Ollama

Let's get you up and running with your own large language models.

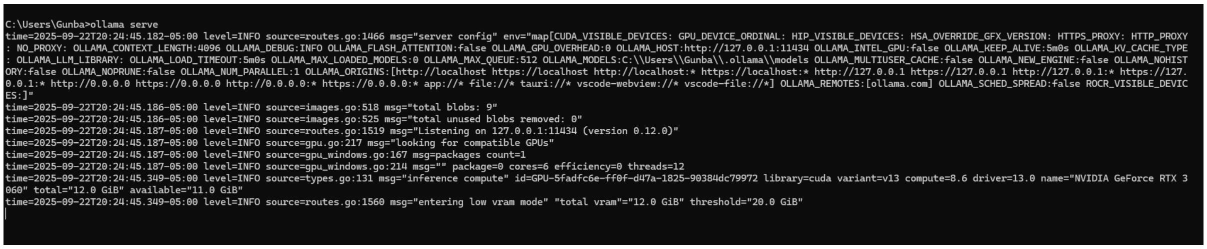
Click Install to continue with the installation.

Step 2: Start Ollama

You can start Ollama in two ways:

- **Option 1:** Open the Ollama app from the Start Menu.
- **Option 2:** Use Command Prompt and run:

```
ollama serve
```

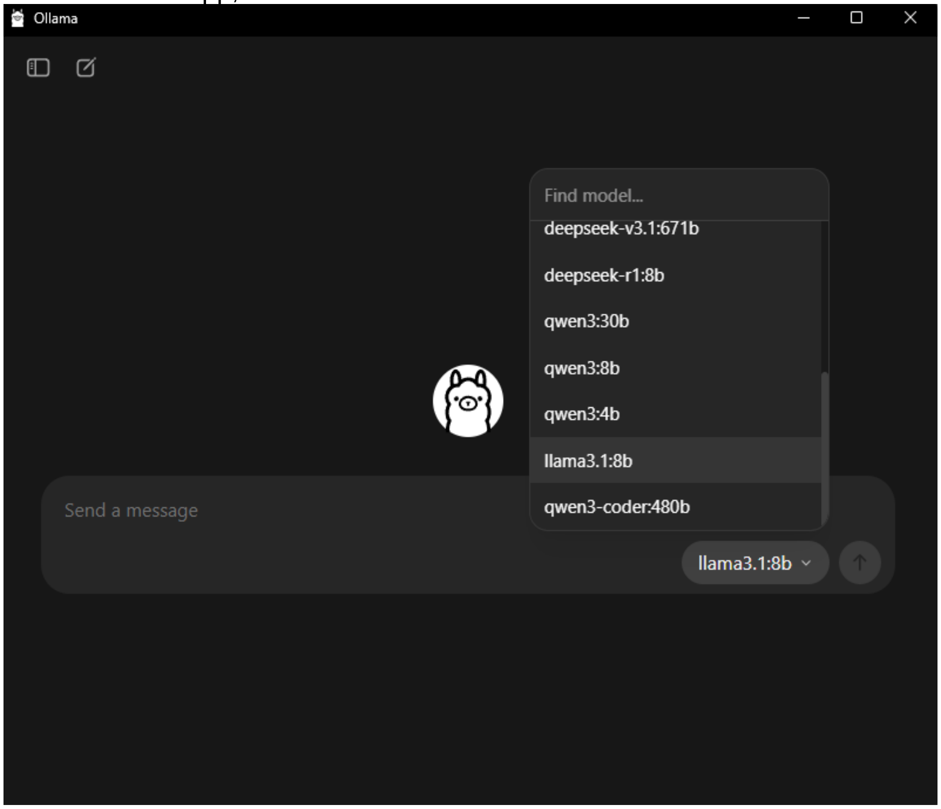


```
C:\Users\Gunba>ollama serve
time=2025-09-22T20:24:45.182-05:00 level=INFO source=routes.go:1466 msg="server config" env="map[CUDA_VISIBLE_DEVICES: GPU_DEVICE_ORDINAL: HIP_VISIBLE_DEVICES: HSA_OVERRIDE_GFX_VERSION: HTTPS_PROXY: HTTP_PROXY
: NO_PROXY: OLLAMA_CONTEXT_LENGTH:4096 OLLAMA_DEBUG:INFO OLLAMA_FLASH_ATTENTION:false OLLAMA_GPU_OVERHEAD:0 OLLAMA_HOST:http://127.0.0.1:11434 OLLAMA_INTEL_GPU:false OLLAMA_KEEP_ALIVE:5m0s OLLAMA_KV_CACHE_TYPE
: OLLAMA_LLM_LIBRARY: OLLAMA_LOAD_TIMEOUT:5m0s OLLAMA_MAX_LOADED_MODELS:0 OLLAMA_MAX_QUEUE:512 OLLAMA_MODELS:C:\\Users\\Gunba\\.ollama\\models OLLAMA_MULTIUSER_CACHE:false OLLAMA_NEW_ENGINE:false OLLAMA_NOHIST
ORY:false OLLAMA_NOPRUNE:false OLLAMA_NUM_PARALLEL:1 OLLAMA_ORIGINS:[http://localhost https://localhost http://localhost:* https://localhost:* http://127.0.0.1 https://127.0.0.1 http://127.0.0.1:* https://127.
0.0.1:* http://0.0.0.0 https://0.0.0.0 http://0.0.0.0:* https://0.0.0.0:* app://* file://* tauri://* vscode-webview://* vscode-file://*] OLLAMA_REMOTES:[ollama.com] OLLAMA_SCHED_SPREAD:false ROCR_VISIBLE_DEVIC
ES:]"
time=2025-09-22T20:24:45.186-05:00 level=INFO source=images.go:518 msg="total blobs: 9"
time=2025-09-22T20:24:45.186-05:00 level=INFO source=images.go:525 msg="total unused blobs removed: 0"
time=2025-09-22T20:24:45.187-05:00 level=INFO source=routes.go:1519 msg="Listening on 127.0.0.1:11434 (version 0.12.0)"
time=2025-09-22T20:24:45.187-05:00 level=INFO source=gpu.go:217 msg="looking for compatible GPUs"
time=2025-09-22T20:24:45.187-05:00 level=INFO source=gpu_windows.go:167 msg=packages count=1
time=2025-09-22T20:24:45.187-05:00 level=INFO source=gpu_windows.go:214 msg="" package=0 cores=6 efficiency=0 threads=12
time=2025-09-22T20:24:45.349-05:00 level=INFO source=types.go:131 msg="inference compute" id=GPU-5fadfc6e-ff0f-d47a-1825-90384dc79972 library=cuda variant=v13 compute=8.6 driver=13.0 name="NVIDIA GeForce RTX 3
060" total="12.0 GiB" available="11.0 GiB"
time=2025-09-22T20:24:45.349-05:00 level=INFO source=routes.go:1560 msg="entering low vram mode" "total vram"="12.0 GiB" threshold="20.0 GiB"
```

Step 3: Download the Required Model

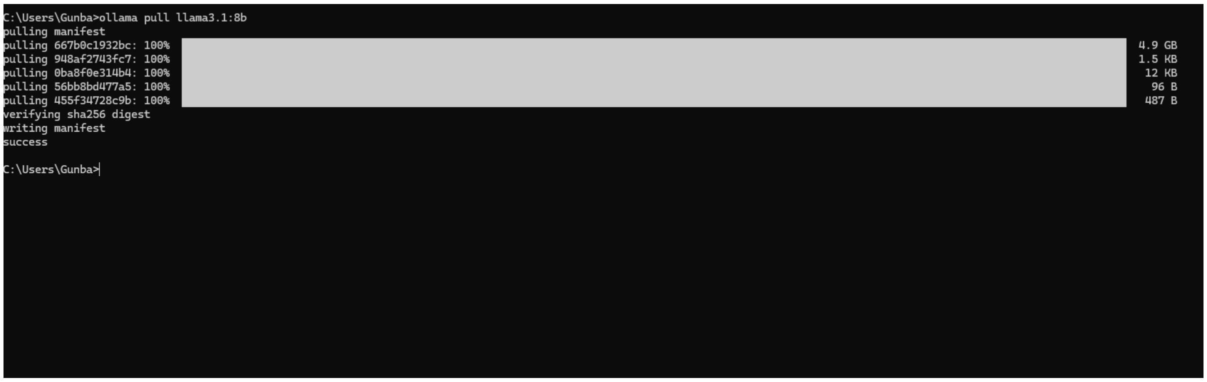
The Internship Application Manager uses the **llama3.1:8b** model.

Option 1: From the Ollama App, search for llama3.1:8b:



Option 2: From Command Prompt, run:

```
ollama pull llama3.1:8b
```



```
C:\Users\Gunba>ollama pull llama3.1:8b
pulling manifest
pulling 667b0c1932bc: 100%                                   4.9 GB
pulling 948af2743fc7: 100%                                   1.5 KB
pulling 0ba8f0e314b4: 100%                                    12 KB
pulling 56bb8bd477a5: 100%                                    96 B
pulling 455f34728c9b: 100%                                   487 B
verifying sha256 digest
writing manifest
success

C:\Users\Gunba>
```

Running the Internship Application Manager

■■ Before running the app, make sure Ollama is running — either the app is open or you see it in Task Manager (Windows).

From the repository root, run:

```
dotnet run
```

This will start the web app locally. Open http://localhost:5000/ in your browser.

Troubleshooting

- If you see 'server not running' or 'connection refused', make sure Ollama is running and the model is downloaded.
- Restart both Ollama and your backend if needed.
- Check firewall/antivirus settings.
- Run 'ollama serve' manually to view error messages.

■ Quick Checklist

✔	Ollama is installed
✔	Ollama is running (app or ollama serve)
✔	Model llama3.1:8b is downloaded
✔	Backend is running (dotnet run)
✔	Access http://localhost:5000/ in your browser

■ All Done!

If you followed all the steps, hooray! Your Internship Application Manager with Ollama AI is ready to go. Enjoy tailoring resumes and exploring internship postings!