

Choosing AI Software

I did some research (<https://duck.ai>, Claude, CoPilot). Following is the result after the AI and I beat my head against the wall until I figured some stuff out. Also, thanks to Randell for giving me assurance I was not on a dead end path here.

The Front End

For an Artificial Intelligence server, you need one piece of software installed that knows how to manage whatever model(s) you use. [Ollama](#) is a local runtime and API for running LLMs on your own machine. It gives you a simple way to download, run, and serve models locally (direct copy from their web site). It basically provides the front end to your model.

The Model

The front end (Ollama) is just that; it interfaces between the outside world and the engine that learns and replies. This is called a model.

llama3.2:3b is a small model aimed at local use, with a 3B parameter size and about a 2.0 GB download in the Ollama library listing. From Meta, it is a text-only 3-billion-parameter version. Better still, it is packaged in Ollama's library so easy to install.

This should be a model that will not strain my little no-GPU server.

What does the '3b' mean

This is a direct copy/paste from an AI when I asked that question. Somewhat edited by me where necessary.

3B vs 8B vs 70B means the model has about 3 billion, 8 billion, or 70 billion parameters. More parameters usually means better capability, but also more RAM/VRAM, slower inference, and more cost.

In plain English:

- 3B: small, fast, cheap to run locally; good for simple chat, summarizing, extraction
- 8B: a strong middle ground; noticeably smarter than 3B for many tasks
- 70B: much more capable, especially on harder reasoning and writing, but needs far more hardware

For Llama 3.2 specifically, the 3B model is a text-only multilingual model meant for efficient local and on-device use. Meta describes the 1B and 3B models as lightweight and capable for multilingual text generation and tool use.

The tradeoff is basically:

- bigger = smarter, slower, heavier
- smaller = faster, lighter, less capable

===== Different Models

And note, *llama* is not the only one available. *qwen* (from Alibaba Cloud) is supposed to be even better than llama for what I want to do, but llama is simpler to use, from what I read.

I can always add a new model, be it llama 8B, or qwen something.

Note: determine your task first. My task is to analyze and summarize daily e-mail reports from servers. If I wanted assistance coding, or needed to analyze images from cameras, I'd choose a different model.

llama3.2 and qwen3 are supposed to be good at text analysis, so they were my only choice. Again, llama is reputed to be the easiest, with qwen a little harder. That, plus the origins (Meta vs Alibaba) made my initial choice fairly easy. And, the ability to simply add a new engine did not hurt.

From:

<https://kb.unixservertech.com/> - **Unix Server Tech Knowledge Base**

Permanent link:

<https://kb.unixservertech.com/software/ai/choosing>

Last update: **2026/10/27 03:01**

