RECIPES FROM THE AI KITCHEN

DELL Technologies

Watch a quick video tutorial ⟶

Run AI models locally with a Dell Precision AI Workstations

As companies embrace AI to make the day-to-day easier, many are drawn to cloud-based solutions due to the ease of setup. But the truth is that no matter how appealing cloud-based solutions sound, they can expose sensitive company data to risk.

At Dell Technologies, we advocate for running AI directly on your device to keep things more secure. AI PCs and AI Workstations not only make you more efficient, but it will also change *how* you get stuff done. The best part? You're not tied to an internet connection or reliant on cloud services. You can get your work done securely, independently, right on your device.

What you need to get started

Dell Precision AI Workstation

With a Precision Workstation with a discrete GPU, your machine can handle AI workloads faster and more efficiently

Shop Precision AI Workstations ⟶

LM Studio (Or GPT4All, Ollama)

Run local Large Language Models (LLMs) through this application on your AI Workstation

Get Started with local AI models

Steps:

1. Make your sure your PC can handle running LLMs locally

If you have 10GB+ of RAM, you should be able to run quantized versions of popular 3B - 7B models comfortably. With an NPU or a high performing discrete GPU your operations will run faster and will be less tasking on your device

2. Download an AI software platform

Directly download the application from lmstudio.ai or O llama or GPT4ALL

3. Load an LLM on LM Studio

Start with a popular open-source model like Llama 3 or Phi-3. You may have to experiment with which sized version of the model works well on your device.

4. Prompt your LLM with the relevant information about your task

Prepare your prompt, whether you are wanting to create formatted tables quickly with raw data or create content in minutes, prompt your LLM as you would with any other chatbot. Here's are a few example for you to easily copy and paste and customize to fit your needs

5. Press Enter and wait for the model to produce a response.

Some rules of thumb: If you have longer prompt it can take longer time to process, if you have a smaller model, it typically will generate a response faster, or if you have larger model, it is usually more capable (but not always true!)

You're now using LLMs locally in LM Studio, with full data privacy and no cost, (and you don't even need to be connected to the internet).

EXAMPLE 1 (Create a formatted table in seconds)
Save time and reduce errors in data entry by automating the conversion of raw and unorganized data into a structured table format.

INPUT EXAMPLE:

```
[Please create a table with the following columns: Name, Date, and Quantity. Populate the table with the data provided.]
```

EXAMPLE 2 (Automate meeting summaries)
Streamline the creation of meeting summaries and follow-up notes with minimal manual effort by inputting your meeting transcript into the LLM, ensuring clarity in your notes.

INPUT EXAMPLE:

```
[Please summarize the meeting notes. Include key points, decisions made, and any assigned responsibilities.]
```

Save the finalized summary. The LLM can then reference this meeting to incorporate additional content from follow-up transcripts

EXAMPLE 3 (Create social campaigns in just minutes)
With on-device LLMs, you can confidently create a comprehensive and secure social media campaign quickly that ensure proprietary and confidential launch information stays protected.

INPUT EXAMPLE:

```
[Create a social campaign to boost engagement for our new summer product line, targeting young adults aged 18-30 who love outdoor activities. The theme is "Summer Fun," with key messages highlighting our products as ideal for summer adventures and outdoor enjoyment
```

What more can you do?

- Turn on GPU acceleration in LM Studio for faster performance
- Experiment with different prompts
- Experiment with different models (e.g. coding models if you're a coder)

Explore use cases on Dell.com/AI