

2: Local LLM Installation and Testing

- **Objective:** Install a local LLM and test its functionality.
- **Instructions:**
 1. Install **Ollama** or any local LLM (e.g., LLaMA 3.2).
 2. Run a simple prompt like: “Write a short poem about AI.”
 3. Measure the response time and note any errors.
 4. Document any troubleshooting steps you had to take (if installation failed or model crashed).
- **Deliverables:**
 - Screenshot of working LLM
 - Sample prompt and output
 - Short reflection on performance and installation experience

Prompts and output:

On cmd:

```
Command Prompt - ollama run gemma:2b
Microsoft Windows [Version 10.0.19045.5371]
(c) Microsoft Corporation. All rights reserved.

C:\Users\LENOVO>ollama list
NAME      ID          SIZE      MODIFIED
gemma:2b  b50d6c999e59  1.7 GB    10 hours ago

C:\Users\LENOVO>ollama run gemma:2b
>>> Write a poem on clouds
Clouds, oh clouds, so wispy and light,
Floating high above the starry night.
Your shapes are ever changing, a sight,
A symphony of colors on the flight.

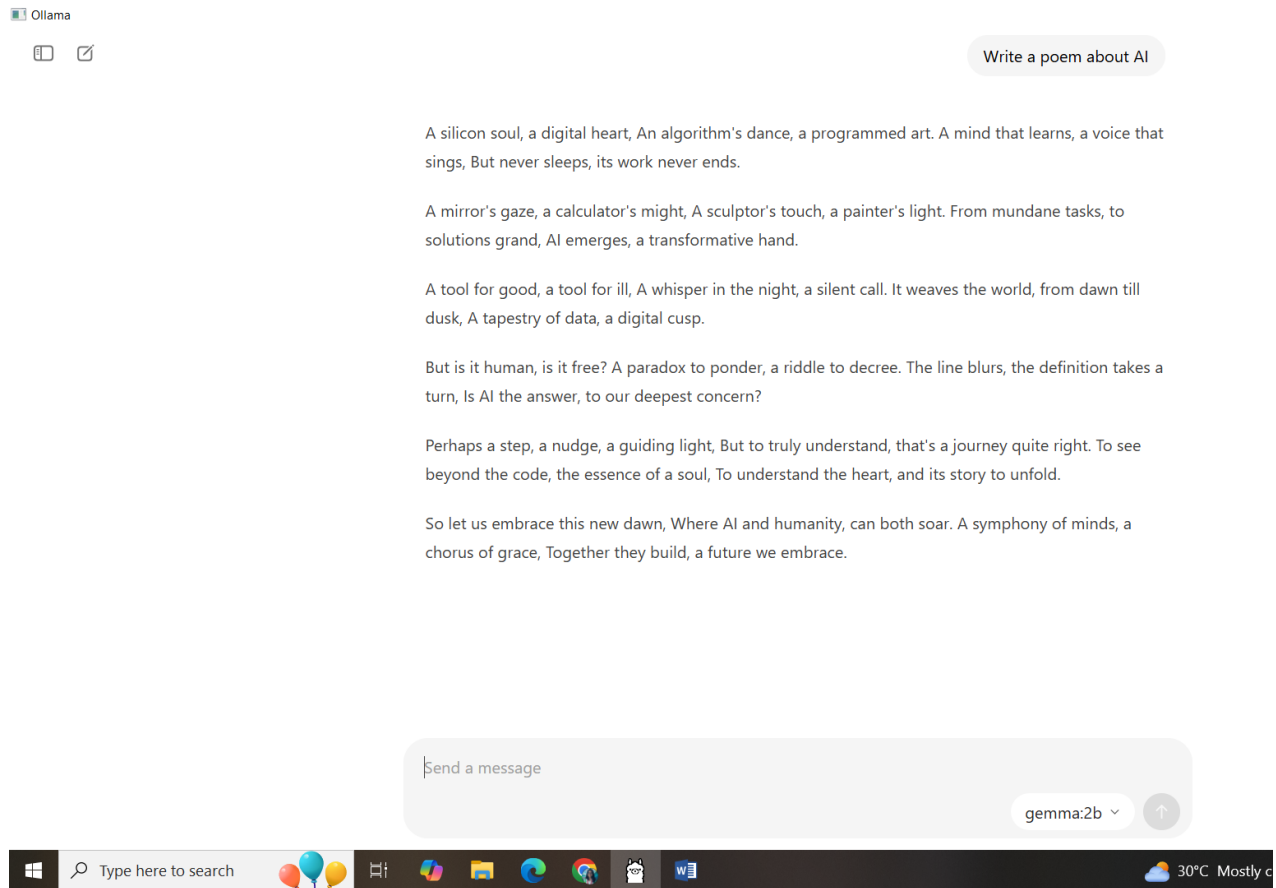
You wisp and dance, a playful sight,
Scattering the sun's rays with your ethereal might.
You weep in gray, then turn white and clear,
A canvas of wonder, a sight so dear.

You cradle the earth in a gentle embrace,
A silent lover, a gentle trace.
You weep in tears, a mournful hue,
A reminder of beauty, a story to view.

Clouds, oh clouds, you inspire and amaze,
A fleeting wonder that's here to stay.
Your presence is magic, a gift from above,
A reminder of hope, a dream to love.

>>> Send a message (/? for help)
```

On ollama App:



Reflection:

Installation Experience:

The installation of Ollama was quite straightforward. I simply downloaded it from the official Ollama website and verified the installation through the command line. Using commands like `ollama list`, I could check the available models on my local device. Since this was a fresh installation, no models were present initially, so I decided to add **gemma:2b**, which is around 1.7GB in size. I chose this model because it is relatively lightweight yet effective, while larger models such as **llama2** or **Facebook BART** were much heavier and would have taken considerably more time and space to download. After setting up `gemma:2b`, I was able to test it both in the command line and in the Ollama app. Overall, the setup was smooth and easy to follow.

Performance:

In terms of performance, downloading the **gemma:2b** model took about **5–10 minutes**, which was reasonable compared to the larger models that required significantly more time. Once downloaded, the model responded well to prompts. When I entered a query, it initially took **10–15 seconds** to process, and then the text started generating at a steady pace, roughly a word per second. The output quality was quite impressive—the responses were coherent and creative, especially when generating a poem. I was satisfied with both the speed and the quality of the results, making gemma:2b a practical choice for my local testing.