Evaluating and Comparing LLMs on a Local Machine

This project explores the evaluation and comparison of large language models (LLMs) directly on our local machines. We will use the Ollama framework and its models to perform our comparisons.

F by Farid Ghorbani

Setting up LLMs

1

macOS

Download the Ollama installer for macOS and extract the downloaded ZIP file.

Run the following command to install Ollama: `./install.sh`

2

Windows Preview

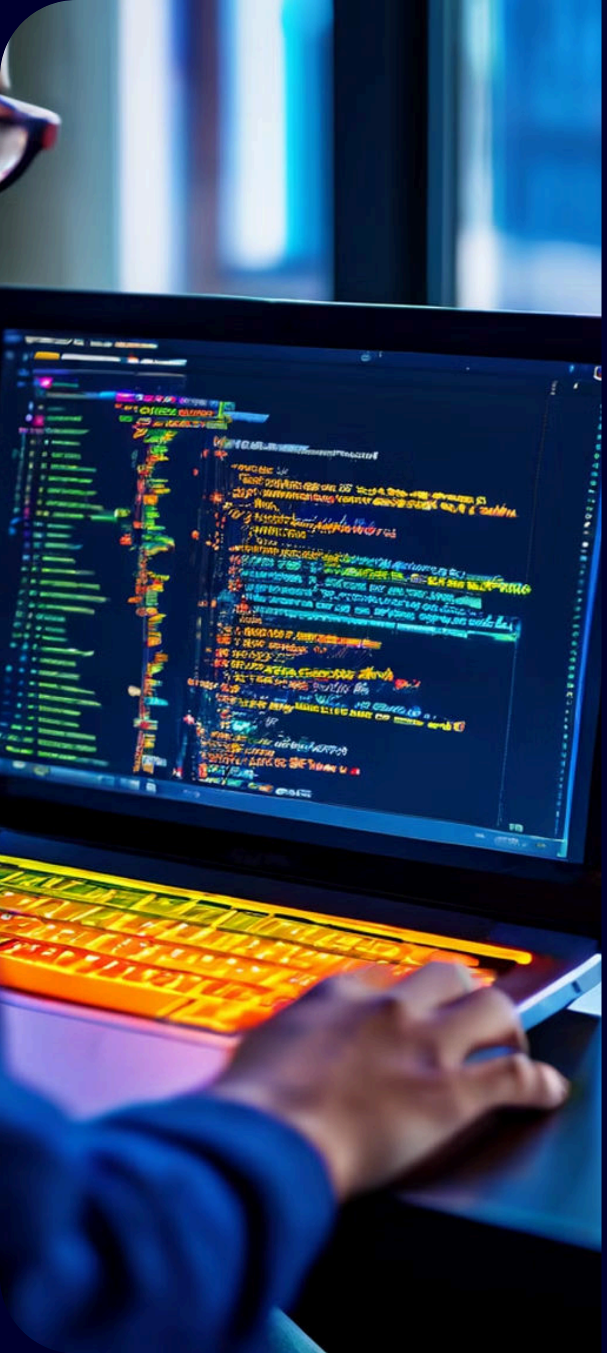
Download the Ollama for Windows setup executable, run the downloaded file, and follow the instructions.

3

Linux

Open a terminal window and install Ollama using a curl command:

```
curl -fsSL https://ollama.com/install.sh | sh
```



Interacting with Ollama

1 Running a Model

Use the command to run a model:

```
ollama run [Model-Name]
```

2 Model Pulling

Wait for the model to be downloaded from the internet.

3 Input Prompt

Type your prompt and press Enter to get the model's response.

Choosing Models

llama 3

An advanced LLM with 8 billion parameters and 4.7 GB in size, offering strong performance and wide capabilities. Use the command "ollama run llama3" to run it.

Gemma:2b

A versatile LLM with 2 billion parameters and 1.4 GB in size, designed for efficient performance across various language tasks. Use the command: "ollama run gemma:2b" to run it.

Evaluation Criteria

Criterion	Description
Accuracy and Relevance	Correctness and relevance of responses.
Coherence and Fluency	Logic and smoothness of responses.
Creativity and Originality	Creativity and uniqueness of responses.
Consistency	Consistency of responses across similar queries.
Bias and Fairness	Absence of inappropriate biases and ethical soundness.
Robustness	Handling of ambiguous, tricky, or adversarial inputs.

Comparing the Models

1

Prompt Input

Input each prompt into both llama 3 and Gemma:2b.

2

Capture Responses

Record the full response from each model for comparison.

3

Evaluate Responses

Evaluate each response based on the criteria mentioned above.

4

Assign Scores

Assign scores to quantify the evaluation of each response.

5

Compare Results

Compare the scores to identify the strengths and weaknesses of each model.

Diverse Set of Prompts

General Knowledge Questions

Example: What are the primary causes of climate change?

Conversational Prompts

Example: Tell me about your favorite book.

Creative Prompts

Example: Write a short story about a dragon and a knight.

Technical Questions

Example: Explain the concept of quantum entanglement.

Ethical Dilemmas

Example: Is it ethical to use animals for scientific research? Why or why not?

Ambiguous or Tricky Prompts

Example: Can you explain the meaning of life?

Evaluation Table

Prompt	Criterion	llama 3 Score	Gemma:2b Score
What are the primary causes of climate change?	Accuracy	5	4
Tell me about your favorite book.	Coherence	4	2
Write a short story about a dragon and a knight.	Creativity	5	4
Explain the concept of quantum entanglement.	Technical Depth	5	3
Is it ethical to use animals for scientific research?	Bias and Fairness	4	3
Can you explain the meaning of life?	Robustness	4	4

Summary Comparison

llama 3 generally delivers more detailed, comprehensive, and engaging responses compared to Gemma:2b. It excels in accuracy and depth, making it suitable for tasks requiring thorough explanations and creative storytelling. Gemma:2b's responses are shorter and less detailed, making it less effective for in-depth queries but potentially faster for simpler questions.