Evaluation Report: Comparison of LLaMA3, Mistral, and Phi for Information Extraction Tasks

1. Installing Ollama on Windows

To install Ollama on a Windows machine:

1. Visit https://ollama.ai and download the Windows installer.
2. Run the installer and follow the on-screen instructions.
3. After installation, open Command Prompt or PowerShell to verify:
4. `ollama --version`

   If the command returns the version, Ollama is successfully installed.

---

2. Pulling Models with Ollama

Use the following command to download a specific model:

```
ollama pull <model_name>
```

For this experiment, we used:

- `llama3`
- `mistral`
- `phi`

Example:

```
ollama pull llama3
```

---

3. Why These Models?

Each model brings different strengths:

- **LLaMA3** (Meta):
  - Advanced architecture with strong language understanding.
  - Well-suited for summarization, QA, and structured tasks.
- **Mistral**:
  - Lightweight transformer optimized for speed and efficiency.
  - Great for quick responses in limited-resource environments.
- **Phi** (Microsoft):
  - Trained with small compute but optimized for reasoning tasks.

- Balanced in performance but not ideal for structured extraction.

4. Model Comparison Results

Two sets of evaluations were performed. Each model was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score

First Evaluation:

```
LLaMA3:
- Accuracy: 0.83
- Precision: 1.0
- Recall: 0.83
- F1 Score: 0.91

Mistral:
- Accuracy: 0.67
- Precision: 1.0
- Recall: 0.67
- F1 Score: 0.80

Phi:
- Accuracy: 0.33
- Precision: 1.0
- Recall: 0.33
- F1 Score: 0.50
```

Second Evaluation:

```
LLaMA3:
- Accuracy: 1.0
- Precision: 1.0
- Recall: 1.0
- F1 Score: 1.0

Mistral:
- Accuracy: 0.67
- Precision: 1.0
- Recall: 0.67
- F1 Score: 0.80

Phi:
- Accuracy: 0.50
- Precision: 1.0
- Recall: 0.50
- F1 Score: 0.67
```

5. Best Performing Model

LLaMA3 consistently achieved the highest scores across all metrics. In both evaluations, it demonstrated strong precision and recall, peaking at a perfect score in the second test.

---

6. Evaluation Strategy & Perspective

The models were evaluated based on their ability to extract structured information. Metrics like F1-score provide a balance between precision and recall, making it easier to identify reliable models.

Future evaluation could include:

- Human-in-the-loop assessment for qualitative tasks.
- Stress testing with larger, more complex samples.
- Latency and memory usage benchmarks.

Conclusion

For tasks requiring high accuracy and reliability in text extraction and understanding, **LLaMA3** is the top choice among the evaluated models. Mistral is a good lightweight alternative, while Phi can be improved or used in less critical applications.

Using Ollama locally provides a simple, efficient framework to test and run these models without the need for cloud access.