Open-Source Local Running LLM Models: Evaluation Report

1. Introduction

This report evaluates various open-source Large Language Models (LLMs) that can be deployed locally for our codebase-related tasks. The assessment is based on key factors such as latency, model size, computational requirements, and suitability for different applications.

2. Evaluated Models

The following models have been considered:

- **GPT-Neo (EleutherAI)**
- **GPT-J (EleutherAI)**
- **LLaMA (Meta AI)**
- **BLOOM (BigScience)**
- **OPT (Meta AI)**
- **Alpaca (Stanford, fine-tuned LLaMA)**

3. Evaluation Criteria

3.1 Latency

- **Low Latency**: Suitable for real-time applications.
- **Moderate Latency**: Acceptable for most interactive tasks.
- **High Latency**: Requires powerful hardware to maintain responsiveness.

3.2 Model Size & Hardware Requirements

- **Small Models (<2B parameters)**: Can run on consumer GPUs or high-end CPUs.
- **Medium Models (2B - 13B parameters)**: Require mid-range GPUs with sufficient VRAM (e.g., RTX 3090, A100).
- **Large Models (>13B parameters)**: Need enterprise-level hardware.

3.3 Use Cases

- **Code Understanding & Generation**
- **Summarization & Documentation**
- **Instruction Following & Fine-Tuning Potential**

4. Model Comparisons

Model	Latency	Model Size	Hardware Requirement	Use Cases
GPT-Neo (1.3B, 2.7B, 6B)	Low (1.3B), Moderate (6B)	1.3B - 6B	Consumer GPUs	General NLP, Code Completion
GPT-J (6B)	Moderate	6B	High-end consumer GPUs	Code Generation, Summarization

LLaMA (7B, 13B, 30B, 65B)	Low (7B), Moderate (13B)	7B - 65B	High-end GPUs for larger models	Code Completion, Advanced NLP
BLOOM (Various, up to 176B)	High	1B - 176B	Enterprise GPUs (for 176B)	Multilingual NLP, Large-scale applications
OPT (125M - 175B)	Low (125M - 6B), High (175B)	125M - 175B	Scales from CPU to enterprise GPUs	General NLP, Summarization
Alpaca (7B, fine-tuned LLaMA)	Low	7B	Consumer GPUs	Instruction Following, Task-Specific Fine-Tuning

5. Recommendations

5.1 For Low Latency & Small Size

- **GPT-Neo (1.3B)** or **Alpaca (7B)** are best suited for lightweight deployments.

5.2 For Balanced Performance

- **GPT-J (6B)** or **LLaMA (7B)** provide an optimal balance between efficiency and accuracy.

5.3 For High-Performance Needs

- **LLaMA (13B)** or **OPT (175B)** can be used if extensive computational resources are available.

6. Conclusion

After evaluating various LLM models, I have decided to go with **GPT-Neo**, as it provides a good balance between low latency, reasonable computational requirements, and strong performance for codebase-related tasks. Its smaller models (1.3B and 2.7B) can run efficiently on consumer hardware, making it an ideal choice for local deployment.