Large Language Models (LLMs) are a class of artificial intelligence systems trained to understand and generate human language. These models are built using deep learning techniques, particularly transformer architectures, and are trained on massive text datasets from sources such as websites, books, articles, and programming repositories.

LLMs can perform various language-related tasks including question answering, summarization, translation, text generation, and code completion. The effectiveness of a model depends primarily on the number of parameters it contains and the diversity and quality of its training data.

Examples of Open-Source LLMs

1. **Mistral 7B**
   Developed by Mistral AI, this model has 7 billion parameters and offers a strong balance between performance and efficiency.
   It uses techniques such as Grouped Query Attention and Sliding Window Attention to process longer texts more effectively.
   Mistral 7B is available under a permissive open-source license and can be deployed fully on-premise.

2. **LLaMA 3 (Meta)**
   Created by Meta AI, LLaMA 3 models are available in various sizes, such as 8B and 70B.
   These models achieve high performance in natural language tasks such as reasoning and dialogue.
   Some versions of LLaMA 3 are available for research and commercial use under specific licensing conditions.

3. **Gemma (Google)**
   Gemma models emphasize safety, efficiency, and transparency.
   They are fine-tuned using methods such as Reinforcement Learning from Human Feedback (RLHF).
   Gemma is typically integrated in cloud environments but can also be used in private contexts.

4. **Phi-3 (Microsoft)**
   Phi-3 is a compact language model developed by Microsoft, optimized for resource-limited environments.
   It performs well on understanding and instruction-following tasks and is suitable for edge or on-premise applications.
   Public access may be limited, depending on version and license.

Why Use LLMs Locally?

Deploying LLMs locally (on-premise) is especially useful in business environments where data privacy and control are essential.
Open-source models like Mistral and LLaMA allow companies to build intelligent assistants without sending sensitive data to external cloud providers.
When combined with a retrieval system (e.g. through Retrieval-Augmented Generation, or RAG), these assistants can answer user queries based on internal company documents, ensuring relevance and accuracy.