LLM Models: Current Usage vs Future Considerations

# Overview

This page documents the Large Language Models (LLMs) currently in use and those under consideration for future adoption. It also categorizes models based on their types, each optimized for specific workloads or deployment goals.

# LLM Types Explained

|  |  |
| --- | --- |
| Type | Description |
| Instruct | Models fine-tuned to follow natural language instructions (e.g., chat, summarization). |
| Base | Pre-trained models without instruction tuning. Ideal for custom fine-tuning. |
| Coden / Code | Optimized for programming-related tasks such as code generation, completion, debugging. |
| Distilled | Smaller, compressed versions of large models offering faster inference. |
| Q4 / 4-bit / Quantized | Compressed model formats that reduce memory usage, enabling use on edge devices or GPUs with limited VRAM. |
| Multimodal | Can handle text, images, audio, and video (e.g., GPT-4o). |
| Mixture of Experts (MoE) | Uses subsets of parameters per inference for efficiency and scalability. |
| RAG (Retrieval-Augmented Generation) | Combines search-based context retrieval with generative responses. |

# Section 1: Current LLM Models in Use

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Size | Type | Primary Use Case |
| Mistral | 7x7B | Instruct | General instruction tuning |
| Qwen | 32B | Base | Base model for further fine-tuning |
| LLaMA 3.3 | 70B | Instruct | Complex instruction and QA tasks |
| Qwen 2.5 | 32B | Coden, Instruct | Code generation + multi-turn instructions |
| Gemma | 2–22B | Q4 (Quantized) | Lightweight inference |
| LLaMA 4 Scout | 17B–16E | Instruct | In-progress evaluation |
| Pythia | 12B | Base | Research, experimentation |
| Qwen v1 | 7B | Instruct | Legacy instruct model |
| DeepSeek | — | Distilled LLaMA 8B | Efficient inference |
| LLaMA 3 | 8B | Base | Lightweight general model |
| Gemma3 | 4bit | Quantized | Ultra-low resource inference |

# Section 2: Recommended/Future LLM Models to Consider

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Size | Type | Potential Use Case | Why Consider |
| Claude 3 Opus | — | Instruct, Reasoning | Long context chat, RAG | High accuracy and safety |
| GPT-4o | — | Multimodal | Vision + Text + Audio tasks | Unified multimodal interaction |
| Mistral (MoE) | 12.9B act. | Sparse Mixture of Experts | Low latency, high throughput | Scalable and efficient |
| Command R+ | 35B | RAG | Retrieval-based document search | Great for knowledge-heavy tasks |
| Phi-2 | 2.7B | Instruct (Tiny) | On-device, edge inference | Highly compact |
| Orca 2 | 13B | Reasoning | Chain-of-thought, few-shot tasks | Teacher-style training |
| OpenELM | 1–3B | Tiny Transformers | Lightweight mobile/IoT scenarios | Meta's efficient small models |
| Cohere Command R | — | RAG, Instruct | Enterprise-grade RAG systems | Fast structured output generation |

# Next Steps

• Identify top 2–3 models for experimentation based on business fit.

• Establish benchmarks on latency, throughput, and memory usage.

• Plan for infrastructure alignment with future model requirements.