LLM Model Recommendations

# Top 5 Recommended Models to Add (High Impact)

|  |  |  |
| --- | --- | --- |
| Model | Why Add It | Use Case |
| Command R+ | Best open-weight RAG model; optimized for retrieval + generation | Document Q&A, Knowledge Search |
| Mistral 7B / Mixtral MoE | Very fast, open-weight, modular (Mixture of Experts); cost-efficient | Fast inference, multi-user LLM serving |
| Claude 3 Sonnet | Superior long-context reasoning and safety features | Enterprise chat, summarization |
| GPT-4o (API) | Multimodal capability (vision/audio/text), fallback for premium tasks | Unified chat agents, support bots |
| Phi-2 | Lightweight and accurate; ideal for edge/offline prototypes | Edge AI, quick inference tasks |

# Optional/Niche Additions (Low to Medium Priority)

|  |  |  |
| --- | --- | --- |
| Model | Why Consider | Use Case |
| Orca 2 | Academic/research style reasoning, great for COT | Education tools, few-shot inference |
| OpenELM | Tiny, fast Transformer series from Meta | On-device, mobile prototypes |
| Yi / Qwen 1.5 | Newer alternatives with multilingual & code support | Backup models, language diversity |

# Upgrade Suggestions for Your Existing List

1. Qwen 2.5 → Upgrade to Qwen 1.5 or Qwen3: They include stronger multi-turn and coding support.

2. LLaMA 3 8B → Add LLaMA 3 70B: For tasks needing deeper understanding and longer completions.

3. Gemma 4bit → Add full Gemma 7B FP16 version: For higher accuracy when resources permit.

# Strategic Advice

• RAG support: Add Command R+ + embedding models (e.g., BGE, E5) for production search pipelines.

• Multimodal vision/audio: Start experimenting with GPT-4o (API) or Qwen-vl.

• Inference speed at scale: Use Mixtral MoE for balancing performance and GPU usage.

• Compliance-ready models: Use Claude 3 for enterprise use with privacy/regulatory focus.