Mejores modelos para usar Ollama (en local)

🥇 1. Mejor equilibrio general

Llama 3 – llama3:8b

Por qué es el más recomendado:

- Muy buena calidad de respuestas
- Funciona en ordenadores normales
- Mucho conocimiento general
- Excelente para programar

Muchos rankings lo colocan como el mejor modelo local equilibrado entre calidad y recursos, especialmente la versión **8B** para uso doméstico.

📥 Instalar:

```
ollama run llama3
```

⚡ 2. Muy rápido y ligero

Phi-3 – phi3:mini

Ventajas:

- Muy rápido
- Perfecto para Macs con poca RAM
- Buen rendimiento para código

En pruebas con equipos de 8 GB RAM suele ser el modelo más rápido manteniendo calidad razonable.

 Instalar:

```
ollama run phi3
```

---

3. Muy bueno para texto y chat

Mistral – mistral:7b

Ventajas:

- Muy equilibrado
- Buen razonamiento
- Respuestas claras

 Instalar:

```
ollama run mistral
```

---

4. Gemma

Gemma

El modelo **gemma3:1b** es:

- muy pequeño
- rápido
- pero bastante limitado

Los modelos Gemma más grandes (9B o 27B) funcionan mucho mejor.

Mi recomendación

Instala estos **3 modelos**:

```
ollama pull llama3
ollama pull mistral
ollama pull phi3
```

Y tendrás:

modelo	uso
llama3	asistente general
mistral	escribir / explicar
phi3	rápido y ligero