La nueva generación de modelos locales con Ollama: más allá de Llama 3

Durante mucho tiempo, hablar de modelos de lenguaje en local era hablar casi exclusivamente de Llama 3, Mistral y poco más. Sin embargo, en los últimos meses el ecosistema ha evolucionado de forma notable: han aparecido modelos más eficientes, más rápidos y, en algunos casos, incluso mejores para tareas específicas como programación o razonamiento.

Si estás utilizando Ollama, esto te interesa.

Qué está cambiando realmente

El cambio clave no es solo la aparición de nuevos modelos, sino el enfoque. Ya no se trata de encontrar “el mejor modelo”, sino de combinar varios según el uso que necesites: velocidad, calidad, consumo de recursos o especialización.

El estándar actual: Llama 3 (8B)

Sigue siendo la referencia para uso general.

Por qué destaca:

- Muy buena calidad de respuesta
- Buen rendimiento en programación
- Equilibrio entre recursos y capacidad
- Funciona correctamente en equipos domésticos

 Ideal si quieres un único modelo para todo.

Instalación:

```
ollama run llama3
```

⚡ La gran sorpresa: Phi-3 (mini)

Uno de los modelos más interesantes actualmente.

Ventajas:

- Extremadamente rápido
- Muy bajo consumo de RAM
- Buen rendimiento en tareas de código

 Perfecto para portátiles, automatización y entornos con pocos recursos.

Instalación:

```
ollama run phi3
```

El equilibrado: Mistral (7B)

Un clásico que sigue funcionando muy bien.

Dónde destaca:

- Explicaciones claras
- Buen razonamiento
- Estabilidad en respuestas largas

 Ideal para documentación, formación y redacción técnica.

Instalación:

ollama run mistral

La tendencia: modelos pequeños que sí funcionan

Modelos como Gemma han demostrado que no siempre necesitas grandes recursos para obtener resultados útiles.

Claves:

- Versiones pequeñas → rápidas pero limitadas
- Versiones grandes → mucho más competentes

 Útiles para pruebas, entornos ligeros o despliegues rápidos.

El nuevo enfoque: combinar modelos

Aquí está el verdadero cambio.

En lugar de depender de uno solo, cada vez más usuarios utilizan varios modelos según la tarea:

Modelo	Uso recomendado
Llama 3	Asistente general
Phi-3	Velocidad y automatización
Mistral	Explicaciones y documentación

Gemma	Pruebas ligeras

Recomendación práctica

Un setup básico y muy funcional sería:

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

Con esto cubres:

- Uso general
- Velocidad
- Explicación y documentación

Lo que realmente marca la diferencia

Más allá del modelo, el valor está en cómo lo integras:

- Automatización con scripts (Python, Bash)
- Asistentes privados sin depender de APIs externas
- Procesamiento de datos local
- Integración con tus propios proyectos

Aquí es donde Ollama deja de ser una herramienta experimental y se convierte en una solución práctica.

Conclusión

Los modelos locales han dejado de ser una simple curiosidad técnica. Hoy son una alternativa real para trabajar con inteligencia artificial de forma privada, eficiente y flexible.

Si todavía no estás aprovechando este enfoque, probablemente estás dejando pasar una de las evoluciones más interesantes del ecosistema actual.