JULIO 2026

Modelos de IA Locales

Guia comparativa · GGUF & Ollama

Mas de 50 modelos analizados · 5 categorias · Benchmarks reales del Open LLM Leaderboard

64	5	30	60
MODELOS	CATEGORIAS	CON BENCHMARKS	EN OLLAMA

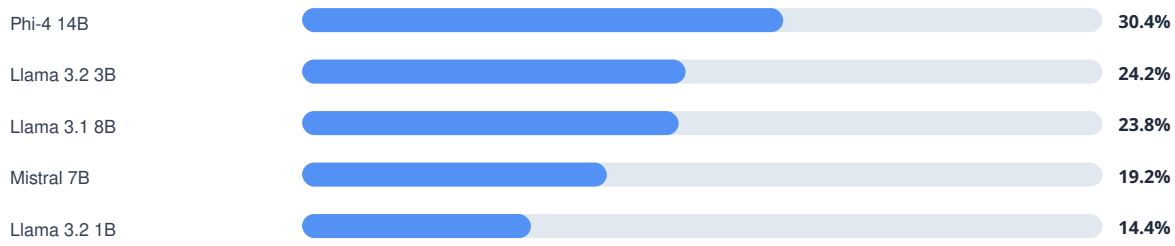
Indice

Propósito general

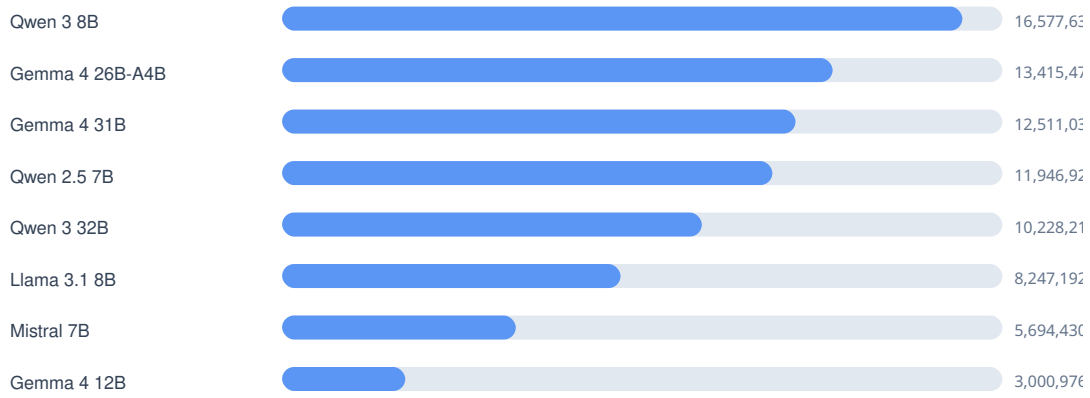
Modelos versátiles para chat, escritura, análisis y tareas generales. Cubren desde 1B hasta 235B parámetros. 24 modelos analizados

MODELO	PARAMS	CONTEXTO	VRAM ESTIMADA	GGUF	OLLAMA	LEADERBOARD	DESCARGAS
Llama 3.1 8B	8B	131K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	**23.8%**	8,247,192
Llama 3.1 70B	70B	131K	~45 GB (Q4) / 148 GB (FP16)	✓	✓	**43.4%**	43,570
Llama 3.2 1B	1B	131K	~0 GB (Q4) / 2 GB (FP16)	✓	✓	**14.4%**	1,832,186
Llama 3.2 3B	3B	131K	~2 GB (Q4) / 6 GB (FP16)	✓	✓	**24.2%**	534,135
Llama 3.2 11B	11B	131K	~7 GB (Q4) / 23 GB (FP16)	✓	✓	—	112,148
Llama 3.3 70B	70B	131K	~45 GB (Q4) / 148 GB (FP16)	✓	✓	**44.9%**	—
Qwen 2.5 7B	7B	131K	~4 GB (Q4) / 16 GB (FP16)	✓	✓	**35.2%**	11,946,929
Qwen 2.5 32B	32B	131K	~21 GB (Q4) / 68 GB (FP16)	✓	✓	**46.6%**	1,796,895
Qwen 2.5 72B	72B	131K	~47 GB (Q4) / 152 GB (FP16)	✓	✓	**48.0%**	526,870
Qwen 3 8B	8B	40K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	—	16,577,631
Qwen 3 32B	32B	40K	~21 GB (Q4) / 68 GB (FP16)	✓	✓	—	10,228,216
Qwen 3 235B-A22B	235B-A22B (MoE)	40K	~152 GB (Q4) / 493 GB (FP16)	✓	—	—	875,892
Gemma 3 2B	2B	131K	~1 GB (Q4) / 5 GB (FP16)	✓	✓	—	—
Gemma 3 12B	12B	131K	~7 GB (Q4) / 25 GB (FP16)	✓	✓	—	1,205,062
Gemma 3 27B	27B	131K	~17 GB (Q4) / 57 GB (FP16)	✓	✓	—	853,942
Gemma 4 12B	12B	131K	~7 GB (Q4) / 25 GB (FP16)	✓	✓	—	3,000,976
Gemma 4 26B-A4B	26B-A4B (MoE)	131K	~16 GB (Q4) / 54 GB (FP16)	✓	✓	—	13,415,471
Gemma 4 31B	31B	131K	~20 GB (Q4) / 65 GB (FP16)	✓	✓	—	12,511,030
Mistral 7B	7B	32K	~4 GB (Q4) / 15 GB (FP16)	✓	✓	**19.2%**	5,694,430
Mixtral 8x22B	8x22B (MoE)	65K	~91 GB (Q4) / 295 GB (FP16)	✓	✓	**33.9%**	78,333
Mistral Small 3.2 24B	24B	131K	~15 GB (Q4) / 50 GB (FP16)	✓	✓	—	280,696
Phi-4 14B	14B	4K	~9 GB (Q4) / 30 GB (FP16)	✓	✓	**30.4%**	722,515
Dolphin 3 8B	8B	131K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	—	72,436
Olmo 2 13B	13B	4K	~8 GB (Q4) / 27 GB (FP16)	✓	✓	—	14,117

Puntuacion Open LLM Leaderboard v2



Popularidad en Hugging Face (descargas)

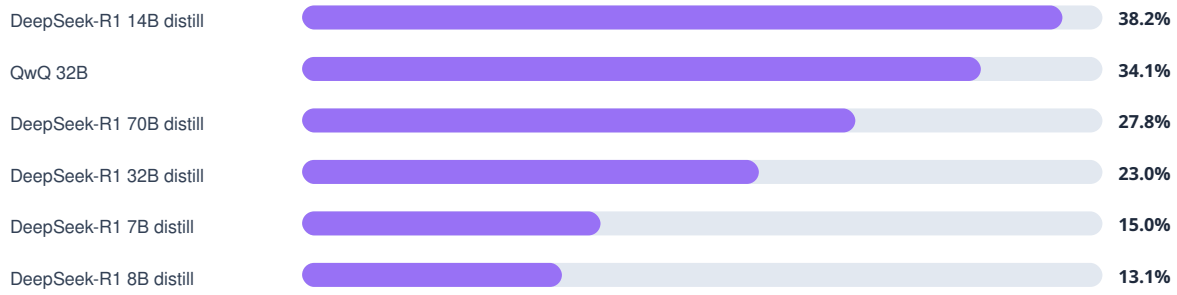


Razonamiento

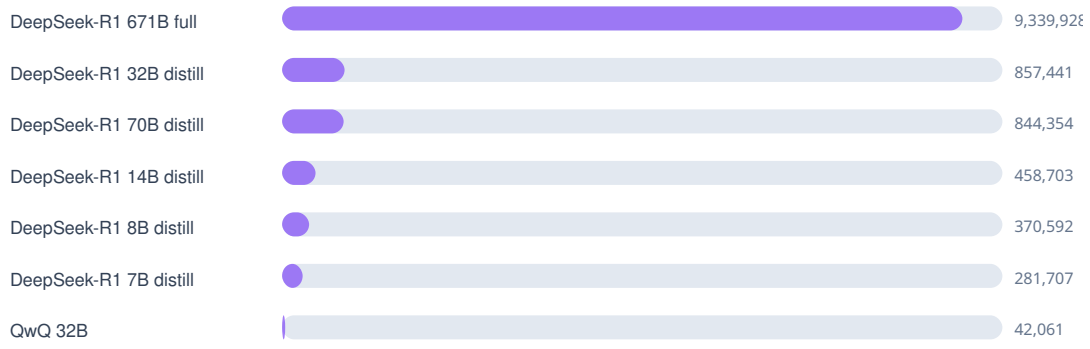
Modelos especializados en pensamiento lógico, matemáticas y problemas complejos con cadenas de razonamiento (CoT). 7 modelos analizados

MODELO	PARAMS	CONTEXTO	VRAM ESTIMADA	GGUF	OLLAMA	LEADERBOARD	DESCARGAS
DeepSeek-R1 7B distill	7B	131K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	**15.0%**	281,707
DeepSeek-R1 8B distill	8B	131K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	**13.1%**	370,592
DeepSeek-R1 14B distill	14B	131K	~9 GB (Q4) / 31 GB (FP16)	✓	✓	**38.2%**	458,703
DeepSeek-R1 32B distill	32B	131K	~21 GB (Q4) / 68 GB (FP16)	✓	✓	**23.0%**	857,441
DeepSeek-R1 70B distill	70B	131K	~45 GB (Q4) / 148 GB (FP16)	✓	✓	**27.8%**	844,354
DeepSeek-R1 671B full	671B (MoE)	131K	~436 GB (Q4) / 1409 GB (FP16)	—	✓	—	9,339,928
QwQ 32B	32B	32K	~21 GB (Q4) / 68 GB (FP16)	✓	✓	**34.1%**	42,061

Puntuacion Open LLM Leaderboard v2



Popularidad en Hugging Face (descargas)

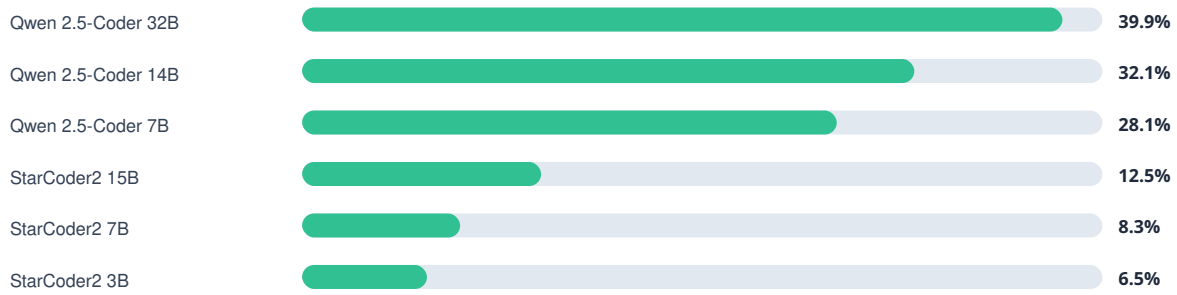


Código

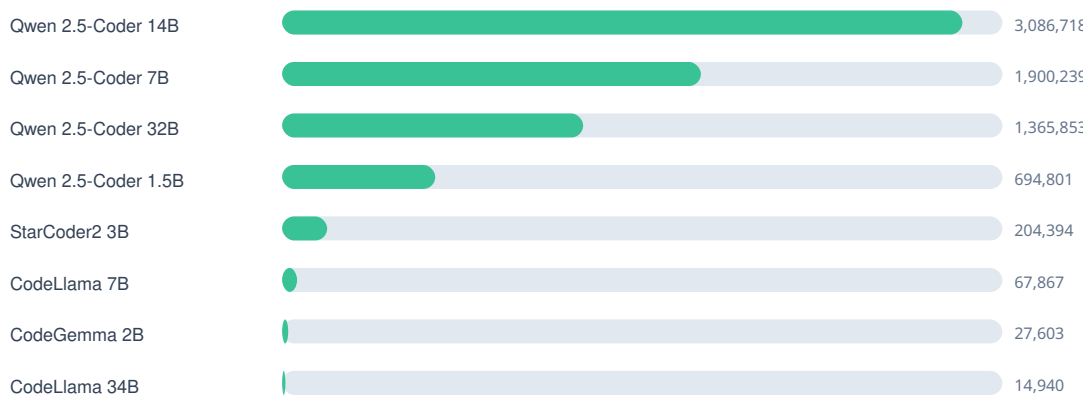
Optimizados para generación, revisión y depuración de código fuente en múltiples lenguajes de programación. 13 modelos analizados

MODELO	PARAMS	CONTEXTO	VRAM ESTIMADA	GGUF	OLLAMA	LEADERBOARD	DESCARGAS
Qwen 2.5-Coder 1.5B	1.5B	32K	~1 GB (Q4) / 3 GB (FP16)	✓	✓	—	694,801
Qwen 2.5-Coder 7B	7B	131K	~4 GB (Q4) / 15 GB (FP16)	✓	✓	**28.1%**	1,900,239
Qwen 2.5-Coder 14B	14B	131K	~9 GB (Q4) / 31 GB (FP16)	✓	✓	**32.1%**	3,086,718
Qwen 2.5-Coder 32B	32B	131K	~21 GB (Q4) / 68 GB (FP16)	✓	✓	**39.9%**	1,365,853
DeepSeek-Coder V2 16B	16B	131K	~10 GB (Q4) / 33 GB (FP16)	✓	✓	—	5,472
CodeGemma 2B	2B	8K	~1 GB (Q4) / 4 GB (FP16)	✓	✓	—	27,603
CodeGemma 7B	7B	8K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	—	1,717
CodeLlama 7B	7B	16K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	—	67,867
CodeLlama 13B	13B	16K	~8 GB (Q4) / 27 GB (FP16)	✓	✓	—	7,260
CodeLlama 34B	34B	16K	~22 GB (Q4) / 71 GB (FP16)	✓	✓	—	14,940
StarCoder2 3B	3B	16K	~1 GB (Q4) / 6 GB (FP16)	✓	✓	**6.5%**	204,394
StarCoder2 7B	7B	16K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	**8.3%**	8,616
StarCoder2 15B	15B	16K	~10 GB (Q4) / 33 GB (FP16)	✓	✓	**12.5%**	8,400

Puntuacion Open LLM Leaderboard v2



Popularidad en Hugging Face (descargas)

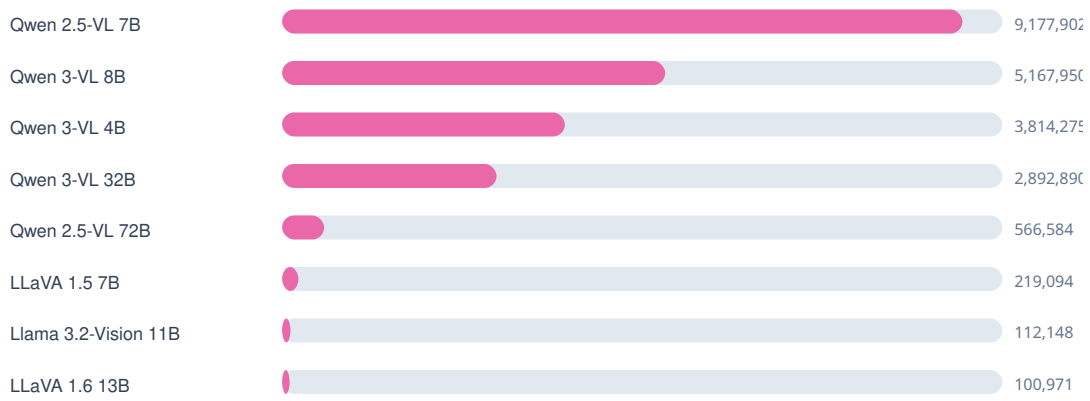


Visión/Multimodal

Capaces de procesar y entender imágenes, documentos escaneados, gráficos y video junto con texto. 12 modelos analizados

MODELO	PARAMS	CONTEXTO	VRAM ESTIMADA	GGUF	OLLAMA	LEADERBOARD	DESCARGAS
Llama 3.2-Vision 11B	11B	131K	~7 GB (Q4) / 23 GB (FP16)	✓	✓	—	112,148
Llama 3.2-Vision 90B	90B	131K	~58 GB (Q4) / 189 GB (FP16)	✓	—	—	49,274
Qwen 2.5-VL 7B	7B	131K	~4 GB (Q4) / 15 GB (FP16)	✓	✓	—	9,177,902
Qwen 2.5-VL 72B	72B	131K	~47 GB (Q4) / 152 GB (FP16)	✓	✓	—	566,584
Qwen 3-VL 4B	4B	32K	~2 GB (Q4) / 8 GB (FP16)	✓	✓	—	3,814,275
Qwen 3-VL 8B	8B	131K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	—	5,167,950
Qwen 3-VL 32B	32B	131K	~21 GB (Q4) / 68 GB (FP16)	✓	—	—	2,892,890
MiniCPM-V 2.4B	2.4B	4K	~1 GB (Q4) / 5 GB (FP16)	✓	—	—	72,908
MiniCPM-V 8B	8B	4K	~5 GB (Q4) / 16 GB (FP16)	✓	✓	—	30,521
LLaVA 1.5 7B	7B	4K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	—	219,094
LLaVA 1.6 7B	7B	4K	~4 GB (Q4) / 14 GB (FP16)	✓	✓	—	8,793
LLaVA 1.6 13B	13B	4K	~8 GB (Q4) / 27 GB (FP16)	✓	✓	—	100,971

Popularidad en Hugging Face (descargas)

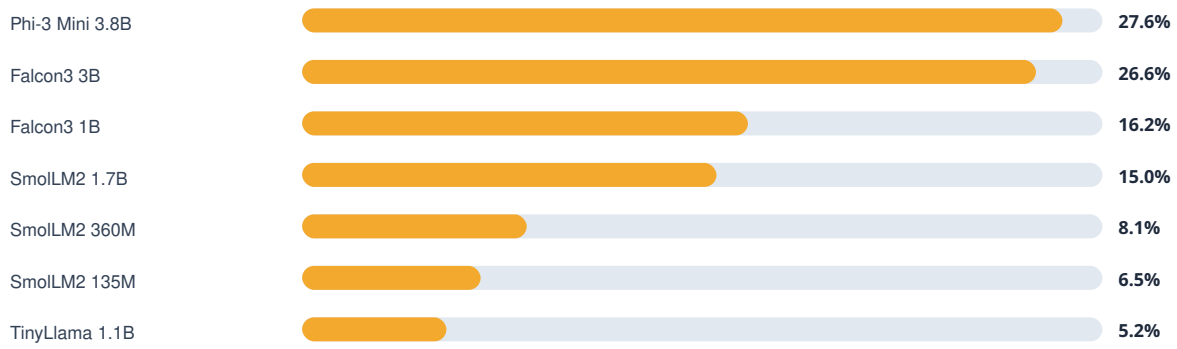


Miniaturas (< 3B)

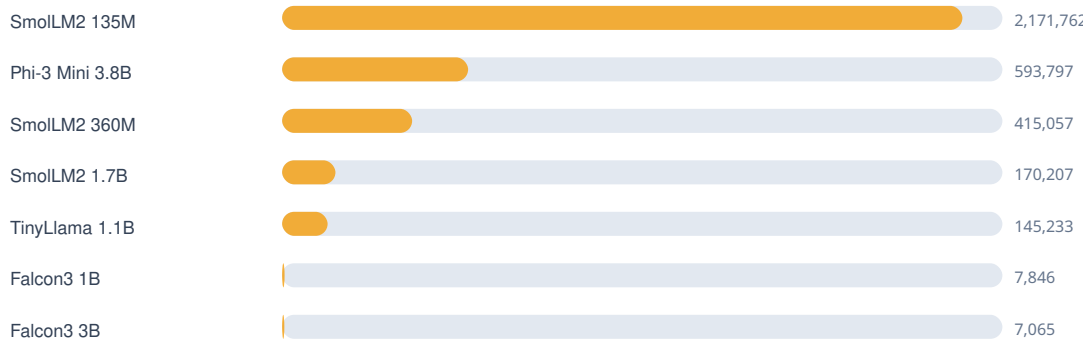
Modelos ultraligeros para CPU, dispositivos móviles y edge computing sin sacrificar demasiada calidad. 8 modelos analizados

MODELO	PARAMS	CONTEXTO	VRAM ESTIMADA	GGUF	OLLAMA	LEADERBOARD	DESCARGAS
TinyLlama 1.1B	1.1B	2K	~0 GB (Q4) / 2 GB (FP16)	✓	✓	**5.2%**	145,233
SmolLM2 135M	135M	8K	~0 GB (Q4) / 0 GB (FP16)	✓	✓	**6.5%**	2,171,762
SmolLM2 360M	360M	8K	~0 GB (Q4) / 0 GB (FP16)	✓	✓	**8.1%**	415,057
SmolLM2 1.7B	1.7B	8K	~1 GB (Q4) / 3 GB (FP16)	✓	✓	**15.0%**	170,207
Falcon3 1B	1B	8K	~0 GB (Q4) / 2 GB (FP16)	✓	✓	**16.2%**	7,846
Falcon3 3B	3B	8K	~1 GB (Q4) / 6 GB (FP16)	✓	✓	**26.6%**	7,065
Phi-3 Mini 3.8B	3.8B	4K	~2 GB (Q4) / 8 GB (FP16)	✓	✓	**27.6%**	593,797
Orca-Mini 3B	3B	4K	~1 GB (Q4) / 6 GB (FP16)	✓	✓	—	—

Puntuacion Open LLM Leaderboard v2



Popularidad en Hugging Face (descargas)



Ranking Global: Top 15 Modelos Locales

Segun puntuacion media del Open LLM Leaderboard v2

#	MODELO	PARAMS	CATEGORIA	SCORE
1	**Qwen 2.5 72B**	72B	Propósito general	**48.0%**
2	**Qwen 2.5 32B**	32B	Propósito general	**46.6%**
3	**Llama 3.3 70B**	70B	Propósito general	**44.9%**
4	**Llama 3.1 70B**	70B	Propósito general	**43.4%**
5	**Qwen 2.5-Coder 32B**	32B	Código	**39.9%**
6	**DeepSeek-R1 14B distill**	14B	Razonamiento	**38.2%**
7	**Qwen 2.5 7B**	7B	Propósito general	**35.2%**
8	**QwQ 32B**	32B	Razonamiento	**34.1%**
9	**Mixtral 8x22B**	8x22B (MoE)	Propósito general	**33.9%**
10	**Qwen 2.5-Coder 14B**	14B	Código	**32.1%**
11	**Phi-4 14B**	14B	Propósito general	**30.4%**
12	**Qwen 2.5-Coder 7B**	7B	Código	**28.1%**
13	**DeepSeek-R1 70B distill**	70B	Razonamiento	**27.8%**
14	**Phi-3 Mini 3.8B**	3.8B	Miniaturas (< 3B)	**27.6%**
15	**Falcon3 3B**	3B	Miniaturas (< 3B)	**26.6%**

Comparativa visual del ranking global

Recomendaciones por Caso de Uso

Seleccion practica segun tu hardware y necesidades

CASO DE USO	VRAM <= 6GB	VRAM <= 12GB	VRAM <= 24GB	VRAM >= 48GB
Chat / Asistente	Llama 3.2 3B - Gemma 3 2B - Phi-4 14B (Q4)	Qwen 3 8B - Llama 3.1 8B - Gemma 4 12B	Qwen 2.5 32B - Gemma 4 31B - Mistral Small 3.2	Qwen 3 235B MoE - Llama 3.3 70B - Qwen 2.5 72B
Razonamiento / Matematicas	DeepSeek-R1 7B - Qwen 3-VL 4B	DeepSeek-R1 14B - QwQ 32B (Q4)	DeepSeek-R1 32B - QwQ 32B	DeepSeek-R1 70B - DeepSeek-R1 671B
Programacion	Qwen 2.5-Coder 1.5B - CodeGemma 2B	Qwen 2.5-Coder 7B - Qwen 3-Coder	Qwen 2.5-Coder 14B - DeepSeek-Coder V2 16B	Qwen 2.5-Coder 32B - DeepSeek-Coder V2 236B
Vision / Imagenes	MiniCPM-V 2.4B - Qwen 3-VL 4B	Qwen 2.5-VL 7B - Llama 3.2-Vis 11B	Qwen 2.5-VL 32B - Qwen 3-VL 32B	Qwen 2.5-VL 72B - Llama 3.2-Vis 90B
Edge / Movil	SmolLM2 360M - TinyLlama - Falcon3 1B	SmolLM2 1.7B - Phi-3 Mini - Gemma 3 2B	—	—

Nota: Las estimaciones de VRAM asumen cuantizacion Q4_K_M (4-bit) y batch size 1. Para FP16, multiplica los parametros x 2. Para Q8, x 1.25. Los modelos MoE (Mezcla de Expertos) activan solo una fraccion de sus parametros por token.

Recursos y Referencias

Donde seguir explorando y comparando modelos

RECURSO	URL	PARA QUE SIRVE
Open LLM Leaderboard v2	hf.co/spaces/open-llm-leaderboard	Benchmarks objetivos y reproducibles
Chatbot Arena (LMSYS)	lmarena.ai	Ranking por preferencia humana (ELO)
Artificial Analysis	artificialanalysis.ai	Comparativa calidad/precio/velocidad
Ollama Library	ollama.com/library	Descarga y ejecucion local directa
r/LocalLLaMA	reddit.com/r/LocalLLaMA	Comunidad, experiencias reales
Hugging Face GGUF	hf.co/models?library=gguf	Buscar cuantizaciones GGUF

Datos recopilados en Julio 2026 - Fuentes: Open LLM Leaderboard, Hugging Face API, Ollama