AI Model Benchmark

Guia para Emprendedores Hispanohablantes

Mayo 2026

72 modelos con cobertura ≥50 runs · 9,628+ runs preservados · v2.6.0

72	25	148	9,628+
MODELOS	SUITES	TESTS/MODELO	RUNS TOTALES

Medido desde Santiago, Chile · 8 providers (OpenRouter, OpenAI, Anthropic, Groq, NVIDIA NIM, Xiaomi, Ollama Cloud, MiniMax)

Como se calcula el Score (v2.6.0)

- **50%** Calidad (formato + sustancia + LLM-as-Judge Phi-4 14B local, MIT)
- **20%** Costo (curva log inversa: $0.001/call=8.0, $0.01=5.0, $0.10=2.0)
- **15%** Tool Calling (precision de function calling para agentes)
- **7.5%** Velocidad (tokens por segundo)
- **7.5%** Latencia (tiempo hasta primera respuesta)

📍 **Disclaimer**: este benchmark NO sustituye a HumanEval, MMLU, GSM8K, SWE-bench, NIAH inglés, MT-Bench, LMSYS Arena. Es **complemento** para emprendedores hispanohablantes que deciden producción real (N8N, OpenClaw, Hermes, blogs LATAM, agentes).

Hallazgos clave de Mayo 2026

Insights cuantitativos del benchmark v2.6.0. Cada hallazgo con fecha de descubrimiento o validación.

1. Provider matters cuantificado (28-30 abril 2026)

Mismo modelo en provider directo vs OpenRouter: **+0.16 a +0.25 puntos**. Confirmado en 4 proveedores (Xiaomi direct, Groq direct, NIM gratis, MiniMax direct). **Para producción: provider directo cuando esté disponible**.

2. Thinking forzado EMPEORA agéntica multi-turn (29-30 abril 2026)

8/9 modelos hybrid bajan vs sin thinking: Opus 4.7 -0.67, Sonnet 4.6 -0.50, Hermes 4 70B -0.54. Solo Kimi K2.5 sube (+0.73). **Para N8N/OpenClaw: NO actives thinking default**.

3. Why Opus 4.7 NO está en top 10 (29 abril 2026)

Opus 4.7 quality 8.08 (top 6 entre todos), pero **40-100x más caro y 5-10x más lento**. Score compuesto pondera costo+speed → fuera del top 10. Para producción a volumen LATAM con presupuesto: alternativas.

4. Modelo gigante NO siempre gana (28 abril 2026)

Mistral Large 3 (675B params) saca 6.89. Nemotron Nano 9B v2 saca 6.91. **Modelo 75x más pequeño rinde más**.

5. "1M context declarado" ≠ retrieval efectivo (1 mayo 2026)

Solo **GPT-4.1 procesa 1M tokens efectivamente** via OpenAI/OpenRouter. Llama 4 Scout (10M declarado), DeepSeek V4 Flash (1M), Gemini 3.1 Pro (1M) — todos cap por providers a 128K-256K real.

6. DeepSeek V4 Pro NIM NO funciona en producción (28 abr + 3 mayo 2026)

Cascada 504 reproducible **2 veces**. NIM gateway no maneja modelo gigante con prompts largos. Para V4 Pro: **OpenRouter pagado o Ollama Cloud sub** (97% éxito).

7. MiMo Xiaomi sub family — mejor C/B en español (2-3 mayo 2026)

4 modelos MiMo en top 10 (V2-Omni, V2.5, V2.5-Pro, V2-Flash). Sub **$14/mes** da acceso a 4 modelos competentes con español neutro fuerte. Mejor opción para emprendedores LATAM con presupuesto fijo predecible.

8. Devstral Small lidera NIAH-ES (30 abr + 3 mayo 2026)

Specialist coding model gana retrieval en español: **7.25 vs Opus 4.7 4.98**. **+2.27 puntos a 1/450 del costo**. Tasks de extracción info con contexto <128K: Devstral.

9. Limitación honesta: NO medimos debugging real (30 abril 2026)

Caso reportado: MiniMax M2.7 NO pudo resolver bug en VPS Hetzner. Opus 4.7 sí en minutos. Para debugging agentic real, mira **SWE-bench Verified** (Opus 4.7 = 87.6% top 1 globalmente).

10. GPT-5.5 cobertura completa (3 mayo 2026)

GPT-5.5 single-turn 6.32 (regresión vs GPT-5.4 7.23) + agent_long_horizon + NIAH-ES completos. Confirma patrón: thinking puede empeorar en aplicación real.

Ver hallazgos completos con datos cuantitativos: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/INSIGHTS.md

Top 10 Score Compuesto

#	Modelo	Score	Calidad	Costo$	Tool	Speed	Lat	Provider
1	**Llama 4 Scout 17B (Groq preview)**	**7.69**	7.7	8.32	7.04	7.81	5.87	groq_direct
2	**Llama 3.1 8B Instant (Groq)**	**7.67**	7.33	8.72	7.1	8.02	6.3	groq_direct
3	**Devstral Small**	**7.52**	7.89	7.57	6.81	8.47	5.42	openrouter
4	**Mistral Small 4**	**7.51**	7.88	7.76	7.07	7.3	4.53	openrouter
5	**GPT-OSS 20B (Groq)**	**7.47**	7.1	8.47	6.94	9.26	5.89	groq_direct
6	**MiMo V2-Omni (Xiaomi direct)**	**7.46**	7.27	8.84	6.93	8.93	2.09	xiaomi_direct
7	**MiMo V2.5 (Xiaomi)**	**7.45**	7.63	8.47	7.14	7.8	1.98	xiaomi_direct
8	**Gemini 3.1 Flash Lite**	**7.44**	7.82	7.36	7.08	7.59	4.98	openrouter
9	**Nemotron 3 Nano 30B**	**7.43**	7.79	8.65	6.6	8.21	1.65	openrouter
10	**MiMo V2.5-Pro (Xiaomi)**	**7.42**	7.65	8.83	7.11	6.97	1.44	xiaomi_direct

Top 5 Quality (sin pesar costo)

#	Modelo	Quality	Final compuesto
1	Gemma 4 31B (DGX Spark Q4_K_M)	**8.22**	6.69
2	Gemma 4 31B (local)	**8.22**	6.69
3	Gemma 4 31B	**8.19**	7.30
4	Gemma 4 31B (NIM)	**8.19**	7.30
5	Mistral Large 3 675B (NIM)	**8.18**	6.83

Lectura: el ranking compuesto pondera quality+costo+speed+latencia. Si solo te importa **quality** (e.g. trabajo crítico sin presupuesto), mira la segunda tabla: Opus/Gemini/Hermes 4 suben fuerte. Si el costo importa: top compuesto.

Ranking completo de 72 modelos en https://benchmarks.cristiantala.com/ · datos crudos en https://github.com/ctala/ai-benchmarks-alternativos

Que Modelo Usar — Por Caso de Uso

Recomendaciones basadas en los datos del benchmark + cross-reference con SWE-bench / NIAH inglés. **Stack agente recomendado**: 1 LLM cabecera (orquestador) + N skills especializados.

Agente cabecera (N8N / OpenClaw / Hermes)

Recomendación	Score agéntico	Costo
🥇 GPT-OSS 120B (Ollama Cloud)	**8.15** (#1 ALH)	$30/mes sub
🥈 Llama 3.3 70B Groq	7.60	$0.59/$0.79 per M
🥉 MiMo V2.5 (Xiaomi sub)	7.65	$14/mes sub

Skill: Coding (workflows N8N, plugins, scripts)

Modelo	Notas
Devstral Small	Apache 2.0 · #1 NIAH-ES · Specialist · $0.10/$0.30
Devstral 2 (Dic 2025)	Apache 2.0 · 256K context

Skill: Content (blog, social, newsletter)

Modelo	Notas
MiMo V2.5 (Xiaomi sub)	Sub $14/mes · español neutro fuerte
Gemini 3.1 Flash Lite	$0.25/$1.50 · Google directo

Skill: Customer Support multi-turn

Modelo	Notas
GPT-OSS 120B Cloud	#1 agent_long_horizon (8.15)
Llama 3.3 70B Groq	Latencia ultra baja (270 tok/s)

Skill: Research con tools (Perplexity-style)

Modelo	Notas
DeepSeek V4 Flash (NIM)	GRATIS · 40 RPM · 1M context
Mistral Small 4	$0.15/$0.60 · top quality

Skill: Long-context retrieval (>32K tokens)

Modelo	Notas
GPT-4.1	Único modelo confirmado a 1M efectivo
Gemini 3.1 Pro	Más estable a 256K (5.37 avg)

Para debugging agentic real

> **NO usar nuestro ranking** — usa SWE-bench Verified. Top: Opus 4.7 (87.6%), Sonnet 4.6 (77.2%), GPT-5.x. Caso real reportado: solo Opus resolvió bug Docker complejo.

Recomendaciones detalladas por plataforma + tarea + presupuesto: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/RECOMENDACIONES.md

Rankings por Categoria — Top 5 por área

Razonamiento (single-turn)

#	Modelo	Score
1	Llama 4 Scout 17B (G	**8.26**
2	Gemini 3.1 Flash Lit	**7.89**
3	Mistral Small 4	**7.87**
4	Llama 3.3 70B (Groq)	**7.84**
5	Grok 4.1 Fast	**7.81**

Contenido / Marketing

#	Modelo	Score
1	Llama 3.1 8B Instant	**8.27**
2	Llama 4 Scout 17B (G	**8.26**
3	GPT-OSS 20B (Groq)	**8.25**
4	GPT-OSS 120B (Ollama	**8.07**
5	Llama 3.3 70B (Groq)	**8.06**

Multi-step (agent_long_horizon)

#	Modelo	Score
1	GPT-OSS 120B (Ollama	**8.60**
2	Llama 4 Scout 17B (G	**8.26**
3	Llama 3.1 8B Instant	**8.21**
4	MiMo V2-Omni (Xiaomi	**8.17**
5	Devstral Small	**8.12**

Coding

#	Modelo	Score
1	Llama 4 Scout 17B (G	**8.15**
2	Llama 3.1 8B Instant	**8.09**
3	Devstral Small	**8.09**
4	GPT-OSS 20B (Groq)	**7.92**
5	GPT-4.1	**7.92**

Agentes / Operaciones

#	Modelo	Score
1	Llama 3.1 8B Instant	**8.06**
2	Llama 4 Scout 17B (G	**7.79**
3	Mistral Small 4	**7.70**
4	Llama 3.3 70B (Groq)	**7.60**
5	Grok 4.1 Fast	**7.51**

Aguja-en-Pajar (NIAH-ES)

#	Modelo	Score
1	Devstral Small	**7.24**
2	MiMo V2.5 (Xiaomi)	**7.05**
3	Mistral Small 4	**7.01**
4	Llama 4 Scout 17B (G	**6.91**
5	Gemini 3.1 Flash Lit	**6.78**

Ranking completo + drill-down por subcategorías en https://benchmarks.cristiantala.com/

Precios y Suscripciones

Pay-as-you-go (por millón de tokens)

Modelo	Input $/M	Output $/M
Llama 3.3 70B Groq	$0.59	$0.79
Llama 3.1 8B Groq	$0.05	$0.08
Devstral Small	$0.10	$0.30
Mistral Small 4	$0.15	$0.60
Gemini 3.1 Flash Lite	$0.25	$1.50
Grok 4.1 Fast	$0.20	$0.50
GPT-4.1	$2.50	$10.00
Claude Sonnet 4.6	$3.00	$15.00
Claude Opus 4.7	$15.00	$75.00

NIM Gratis (NVIDIA)

Acceso a **20+ modelos gratis** con rate limit 40 RPM. Suficiente para producción de bajo volumen y benchmarks. Modelos: DeepSeek V4 Flash, Gemma 4 31B, Llama Nemotron family, Qwen 3-Next, Mistral Large 3, etc.

Suscripciones Mensuales Fijas

Plan	Precio	Modelos
Xiaomi MiMo Standard	**$14/ mes**	4 modelos: V2.5, V2.5-Pro, V2-Pro, V2-Omni · 200M credits/mes
MiniMax Agent Pro	$19/ mes	M2.7 Highspeed · generosos límites
Anthropic Pro	$20/ mes	Claude (web only, NO API)
Ollama Cloud Pro	$30/ mes	5 modelos: GPT-OSS 120B, DeepSeek V4 Pro/Flash, Qwen 3.5 397B/default

Estrategia de costo recomendada

Si tienes ~$0/mes y bajo volumen: NIM gratis (40 RPM) + modelos open en Groq direct cheap.
Si tienes ~$14/mes: Sub Xiaomi MiMo (4 modelos en español neutro fuerte).
Si tienes ~$30-50/mes: Sub Ollama Cloud + Sub Xiaomi (cobertura completa).
Si tienes >$100/mes: Pay-as-you-go en OpenRouter con fallback automático.

Detalle de suscripciones + cuándo conviene cada una: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/SUSCRIPCIONES.md

Estrategia Local — segun VRAM/RAM disponible

Modelos open-source para correr en hardware propio. Recomendaciones por capacidad disponible (Mac Apple Silicon usa unified memory; PCs con GPU dedicada usan VRAM; servidores con AI accelerator usan unified RAM como DGX Spark 128GB).

≤8 GB (laptops básicas, RTX 3060)

Modelo	Tamaño Q4
Llama 3.1 8B	~4.5 GB
Phi-4 14B (juez Phi-4)	~8 GB
Mistral 7B / DeepSeek-Coder 6.7B	~4-5 GB

16 GB (M2/M3 base, RTX 4070)

Modelo	Tamaño Q4
Gemma 4 26B MoE	~14 GB
Qwen 3.5 25B	~13 GB
Devstral Small (Mistral 24B)	~13 GB

24-32 GB (M2 Pro/Max, RTX 4090)

Modelo	Tamaño Q4
Gemma 4 31B	~18 GB
Nemotron 3 Nano 30B	~17 GB
Qwen 3.5 32B	~17 GB

48-64 GB (M3 Max, M3 Ultra)

Modelo	Tamaño Q4
Llama 3.3 70B	~40 GB
Qwen 3.5 72B	~40 GB
MiniMax M2.5	~50 GB

128 GB (Mac Studio M3 Ultra, NVIDIA DGX Spark)

Modelo	Tamaño Q4
Nemotron 3 Super 120B	~80 GB ✓
Llama 4 Maverick	~110 GB
DeepSeek V3.2	~120 GB

256 GB+ (servidores dedicados)

Modelos gigantes 600B+: Mistral Large 3 675B, Qwen 3.5 397B (necesitan >256GB Q4). Generalmente no rentables vs API a este tamaño.

Cuándo usar local vs nube

Local SI: privacidad de datos · sin latencia red · costo $0 por call · soberanía.
Nube SI: mejor calidad (modelos > tu hardware) · sin setup ni mantenimiento · escalable.

Modelos validados localmente + scripts setup: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/MODELOS.md

Mapa de Proveedores

8 providers cubiertos en mayo 2026. Ordenados por costo (de más barato a más caro).

Provider	Tipo	Modelos clave	Costo	Rate limit	Notas
NVIDIA NIM	Free tier	20+ modelos: DeepSeek V4 Flash, Gemma 4 31B, Llama-Nemotron, Mistral Large 3, Qwen 3-Next	**$0**	40 RPM	NIM Pro saturado para gigantes (504 cascada)
Local (Ollama)	Self-hosted	Modelos open-source bajo tu HW	**$0 (HW propio)**	Tu HW	Ver "Estrategia Local" según VRAM
Xiaomi MiMo	Sub mensual	4 modelos MiMo (V2.5, V2.5-Pro, V2-Pro, V2-Omni)	**$14/mes**	200M credits/ mes	Mejor C/B en español neutro
MiniMax	Sub + API	M2.7 Highspeed	**$19/mes sub o $0.30/$1.20 per M**	Generosos	Ultra baja latencia (Highspeed)
Ollama Cloud	Sub mensual	5 modelos: GPT-OSS 120B, DeepSeek V4 Pro/Flash, Qwen 3.5 397B	**$30/mes**	Variable	97% éxito en V4 Pro (vs NIM 0%)
Groq direct	Pay-as-you-go	Llama family, GPT-OSS 20B	**$0.05-1.36/M**	30 RPM (alto riesgo)	Ultra rápido (270+ tok/s)
OpenRouter	Aggregator	290+ modelos via 1 API key	Variable (margen +5-15% vs direct)	Bajo	Cap 256K en algunos modelos 1M
OpenAI directo	Pay-as-you-go	GPT-4.1, GPT-5.x family	$2.50-75/M	Bajo	Único confirmado a 1M effective
Anthropic API	Pay-as-you-go	Claude Opus/Sonnet/Haiku 4.x	$0.25-75/M	Bajo	SOTA SWE-bench. Sub Pro $20 NO da API
Google AI Studio	Pay-as-you-go	Gemini 2.5/3.x Flash/Pro	$0.10-12/M	Bajo	Mejor a 256K context (NIAH 5.37)

Estrategia recomendada: 1 sub barata (MiMo Xiaomi $14) para producción base + NIM gratis para modelos especializados + 1 cuenta OpenRouter como fallback con $20-50 de credit para emergencias.

Comparativa detallada de proveedores: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/PROVEEDORES.md

Metodologia — Que Medimos (y que NO)

148 tests por modelo en 25 suites

Tipo	Suites	Tests
Single-turn (4 pilares)	23	91
Multi-step (agent_long_horizon) 🆕	1	12
Aguja en pajar (NIAH-ES) 🆕	2	45 (lite) o 60 (full)
Aguja en pajar 1M (NIAH-ES 1M) 🆕	1	15

4 pilares aplicados (single-turn)

Pilar	Suites
Razonamiento	reasoning, deep_reasoning, hallucination, strategy
Coding	code_generation, structured_output, string_precision, ocr_extraction
Contenido	content_generation, summarization, presentation, startup_content, creativity, news_seo_writing, sales_outreach, translation
Agentes	tool_calling, task_management, customer_support, orchestration, multi_turn, policy_adherence, agent_capabilities

Suites nuevas (mayo 2026)

agent_long_horizon (multi-step): 12 tests con conversaciones de 8+ turnos. Mide context retention, skill orchestration, interruption recovery, goal persistence. Plantilla rígida (script de usuario pre-escrito) para reproducibilidad. Tools simulados via stubs.

NIAH-ES (Needle-in-a-Haystack en español): primer NIAH público en español neutro LATAM. 5 needles × 4 contextos (4K-256K) × 3 posiciones. Mide retrieval de info específica en documentos largos. Datos LATAM realistas (códigos, fechas, identificadores).

Que SI medimos

- ✅ **Calidad** (50%): formato + sustancia + LLM-as-Judge Phi-4 14B local
- ✅ **Costo real** (20%): por provider exacto, curva log inversa
- ✅ **Tool calling** (15%): function calling para agentes
- ✅ **Velocidad** (7.5%): tokens por segundo
- ✅ **Latencia** (7.5%): primer token (medida desde Chile)
- ✅ Multi-turn long-horizon (8+ turnos)
- ✅ Long-context retrieval hasta 256K (1M parcial)

Que NO medimos

- ❌ Debugging agentic real (Docker/K8s) — usa **SWE-bench Verified**
- ❌ Capacidades fundamentales generales — usa **HumanEval/MMLU/GSM8K**
- ❌ Multimodal (imágenes/audio) — limitado
- ❌ Consistencia entre runs (single shot N=1)
- ❌ Razonamiento puro complejo — usa **GPQA Diamond/AIME**

> 📍 **Recordatorio**: este benchmark es **complemento**, no sustituto. Para producción aplicada en español neutro LATAM agrega data que los académicos no cubren. Para investigación académica, prioriza los oficiales.

Metodología completa + reproducibilidad: https://github.com/ctala/ai-benchmarks-alternativos/blob/main/CLAUDE.md

¿Te sirvió este cheatsheet?

Llévatelo a producción, aporta datos o únete a la comunidad

Calculadora interactiva

https://benchmarks.cristiantala.com/

Filtra por presupuesto, calidad, velocidad. Top 72 modelos actualizados.

Comunidad Skool

https://www.skool.com/cagala-aprende-repite

Cágala, Aprende, Repite — workshops, casos reales, stack de emprendedores LATAM.

Cómo aportar al benchmark

1. **Reporta tu caso real** · ¿Modelo X resolvió/falló en tu producción? Compártelo en la comunidad o GitHub Issues.
2. **Sugiere tests nuevos** · El benchmark es open-source MIT. Pull requests bienvenidos.
3. **Replica los benchmarks** · Repo público, datos crudos versionados, scripts reproducibles.
4. **Comparte hallazgos** · Si encontrás un patrón en tu uso, documentalo en INSIGHTS.md.

github.com/ctala/ai-benchmarks-alternativos

cristiantala.com · Mayo 2026 · v2.6.0 · MIT

Este benchmark se mantiene gracias a tiempo personal de Cristian Tala Sánchez (Chile) + suscripciones $45/mes. Si te ayudó, comparte el repo. Si querés contribuir, las contribuciones de comunidad se documentan en CHANGELOG.