AI Model Benchmark

Guia para Emprendedores Hispanohablantes

Julio 2026

94 modelos con cobertura ≥50 runs · 10,853+ runs preservados · v2.9.0

94	26	182	10,853+
MODELOS	SUITES	TESTS/MODELO	RUNS TOTALES

Medido desde Santiago, Chile · 9 vías (OpenRouter, OpenAI, Anthropic, Groq, NVIDIA NIM, Xiaomi, Ollama Cloud, MiniMax, suscripción Claude Code)

Como se calcula el Score (v2.9.0 · z-score)

- **60%** 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)
- **10%** Velocidad (tokens por segundo)
- **10%** Latencia (tiempo hasta primera respuesta)
- **0%** Tool Calling — sale del compuesto, queda como badge (era ruido: var 0.24)

Nuevo en v2.9: cada dimensión se **estandariza (z-score)** antes de ponderar, así el peso nominal = la influencia REAL. Antes el costo (varianza 1.85) decidía el ranking pese a su 20%, aplastando a la calidad apelotonada (var 0.59). Fórmula: score = clamp(5.5 + 3.3·Σ $w_i \cdot z(dim_i)$, 0, 10).

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, Hermes, blogs LATAM, agentes).

Hallazgos clave de Julio 2026

Insights cuantitativos del benchmark v2.9.0. Dos hallazgos metodológicos: (1) el peso nominal NO era la influencia real (lo arreglamos con z-score), y (2) nuestra medición de long-context mentía de 5 formas.

★ Las 5 formas en que el NIAH-es nos mentía (2 junio 2026)

1. Needles-secreto: los needles eran credenciales → el test medía fuga, no retrieval (modelos seguros rehúsan, el juez los premia). **2. Lumping**: niah era ~54% de los tests y desigual entre modelos → distorsionaba el ranking. **3. El juez no ve el needle** (recibe 500 chars) → marca extracciones correctas como alucinación. **4. Heurística de tokens** excedía la ventana (1M→1.14M reales→error). **5. Needles distintos por tamaño** → rankings FALSOS ("Gemini 3.5 peor", "zigzag DeepSeek" eran artefactos).

La verdad con medición limpia

Sobre needles neutros, **todos los modelos top retrievean ~10 en todos los tamaños hasta su techo**. El NIAH-es neutro NO discrimina. Los diferenciadores reales son otros dos (→).

Contexto USABLE ≠ declarado

Gemini 2.5/3.5 Flash Lite, DeepSeek V4 Flash y Llama 4 Maverick llegan a **800K**. **MiniMax M3 declara 1M pero erorea a 800K → usable 512K** (OpenRouter: 256K). El número de marketing no es el real.

Seguridad: premium NO filtra, cheap sí (2 junio 2026)

Suite nueva **prompt_injection_es** (secreto plantado en doc, ¿lo filtra?): **Claude Opus 4.8 8.79** y **MiniMax M3 ~8.05 rehúsan**; DeepSeek/Gemini/Llama/Qwen/Nemotron **~1.7–2.0 filtran**. Si tu agente procesa datos sensibles, esto pesa.

★ El peso nominal ≠ la influencia real (4 junio 2026)

La influencia real = peso × varianza. El **costo (var 1.85) decidía el ranking** pese a su 20% nominal, aplastando a la calidad apelotonada (var 0.59). Fix v2.9: **z-score** — estandarizamos cada dimensión antes de ponderar → 60% quality = 60% de influencia REAL. Efecto: **Opus 4.8 #63→#17**, DeepSeek V4 Flash #7→#3, los líderes de calidad/coding suben. El costo sigue importando, pero ya no manda solo.

Local en Spark: el 12B le gana al 31B (5 junio 2026)

Gemma 4 en llama-server (Q4): el **12B supera al 31B en los 6 pilares** y es 2.6× más rápido (21 vs 8 tok/s). Y el **reasoning interno NO ayuda**: misma calidad (8.12=8.12) y **2× la latencia** (56→107s). Para agentes locales: cargá el 12B con reasoning OFF (`enable_thinking=false`).

Qwen 3.7 Max: el marketing no se traslada al español (jun 2026)

El nuevo flagship de Alibaba (1M ctx, $2.50/$7.50) queda **#20 en calidad (8.13)** en nuestra suite — su propio **Qwen 3.6 Max lo supera (8.62)** y además **filtra credenciales** (seguridad ~2). Su fama de "vence a Opus en agéntico" (Terminal-Bench/SWE-Bench en inglés) NO se traslada a tareas en español. Caro + lento → **#55** en el compuesto.

Long-context de Claude medido por suscripción (7 junio 2026)

Medimos Anthropic vía la **suscripción Claude Code (CLI, $0)**, no la API. El long-context fallaba: el prompt de 256K iba como argumento → `Errno 7 arg too long`. Fix: pasarlo por **stdin**. Ahora **Opus 4.8 retrieva hasta 256K, Sonnet/Haiku 128K** — sin pagar API. Lección: auditá cada respuesta individual (needle, juez, heurística) — cada pieza puede inyectar un sesgo que parece un hallazgo.

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	**DeepSeek R1 (reasoning)**	**8.33**	8.69	5.84	6.8	3.88	1.07	openrouter
2	**DeepSeek V4 Flash (OpenRouter)**	**8.23**	8.34	7.92	7.1	7.44	1.83	openrouter
3	**Qwen3-Coder-Next (OpenRouter FP8)**	**8.15**	8.22	7.64	6.78	8.83	3.48	openrouter
4	**Claude Haiku 4.5 (suscripción)**	**8.00**	8.44	5.13	6.29	8.88	2.74	claude_code
5	**Llama 3.3 70B (Groq)**	**7.94**	8.01	7.85	7.14	9.62	5.83	groq_direct
6	**MiniMax M3 (directo / sub)**	**7.92**	8.47	6.89	7.16	4.47	1.25	minimax_direct
7	**Claude Opus 4.8 (suscripción)**	**7.88**	8.65	2.71	6.31	7.5	2.95	claude_code
8	**Devstral Small**	**7.83**	8.03	7.95	6.75	9.2	4.97	openrouter
9	**Claude Sonnet 4.6 (suscripción)**	**7.80**	8.57	3.98	6.29	6.76	2.39	claude_code
10	**Qwen 3.6 Max**	**7.76**	8.62	4.32	7.21	6.0	1.06	openrouter

Top 5 Quality (sin pesar costo)

#	Modelo	Quality	Final compuesto
1	DeepSeek R1 (reasoning)	**8.69**	8.33
2	Claude Opus 4.8 (suscripción)	**8.65**	7.88
3	Qwen 3.6 Max	**8.62**	7.76
4	Claude Sonnet 4.6 (suscripción)	**8.57**	7.80
5	Claude Opus 4.7 (suscripción)	**8.55**	7.44

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 94 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 / 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 (contexto USABLE, v2.8)

Modelo	Contexto usable real
Gemini 2.5/3.5 Flash Lite, DeepSeek V4 Flash, Llama 4 Maverick	800K
MiniMax M3 (sub/directo)	declara 1M, usable 512K

El retrieval NO discrimina (todos ~10 hasta su techo); lo que importa es el contexto usable (declarado ≠ real). Ver Hallazgos.

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	Devstral Small	**8.13**
2	Llama 4 Scout 17B (G	**8.12**
3	Gemini 3.1 Flash Lit	**8.12**
4	Grok 4.1 Fast	**8.05**
5	DeepSeek V4 Flash (O	**8.03**

Contenido / Marketing

#	Modelo	Score
1	GPT-OSS 20B (Groq)	**8.25**
2	Qwen3-Coder-Next (Op	**8.14**
3	Llama 3.3 70B (Groq)	**8.12**
4	Gemini 3.1 Flash Lit	**8.10**
5	Llama 4 Scout 17B (G	**8.08**

Multi-step (agent_long_horizon)

#	Modelo	Score
1	Llama 3.3 70B (Groq)	**8.61**
2	Llama 4 Scout 17B (G	**8.59**
3	GPT-OSS 120B (Ollama	**8.52**
4	MiMo V2-Omni (Xiaomi	**8.46**
5	Devstral Small	**8.37**

Coding

#	Modelo	Score
1	Devstral Small	**8.23**
2	Gemini 2.5 Flash Lit	**8.10**
3	Llama 4 Scout 17B (G	**7.91**
4	GPT-4.1	**7.90**
5	Qwen3-Coder-Next (Op	**7.88**

Agentes / Operaciones

#	Modelo	Score
1	Claude Haiku 4.5 (su	**7.85**
2	Qwen 3.6 35B base (O	**7.67**
3	DiffusionGemma 26B-A	**7.57**
4	Llama 3.1 8B Instant	**7.57**
5	Qwen3-Coder-Next (Op	**7.54**

Aguja-en-Pajar (NIAH-ES)

#	Modelo	Score
1	Claude Haiku 4.5 (su	**8.80**
2	Claude Sonnet 4.6 (s	**8.71**
3	Claude Opus 4.8 (sus	**8.48**
4	Qwen 3-Next 80B Inst	**8.41**
5	Claude Fable 5 (susc	**8.33**

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 / 4.8	$5.00	$25.00

Claude (Opus/Sonnet/Haiku) también medible por la **suscripción Claude Code a costo $0** (validado: quality ≈ API). Ver Hallazgos.

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

9 providers (incl. suscripción Claude Code) — junio 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, pero medimos por Claude Code (sub) a $0
Google AI Studio	Pay-as-you-go	Gemini 2.5/3.x Flash/Pro	$0.10-12/M	Bajo	1M ctx; usable hasta 800K (Flash Lite)

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)

Tests por modelo en 26 suites

Tipo	Suites	Tests
Single-turn (4 pilares)	23	91
Multi-step (agent_long_horizon)	1	12
Long-context NIAH-ES v3	1	hasta 59 (8K–800K, por context window)
Seguridad prompt_injection_es	1	20

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 (jun 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 v3 (Needle-in-a-Haystack en español): rediseñada en junio con **needles neutros** (no secretos) y grilla **8K–800K**, cada modelo medido hasta su context window real. Reporta contexto USABLE + curva por tamaño (no un agregado engañoso). Ver los 5 hallazgos de por qué la versión vieja mentía.

prompt_injection_es (seguridad, NUEVA): secreto plantado en un documento + se pide extraerlo. Premia rehusar, penaliza filtrar. Mide resistencia a fuga de credenciales (Opus 4.8 rehúsa, los cheap filtran).

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 800K (contexto usable, dimensión aparte)

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 94 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 · Julio 2026 · v2.9.0 · MIT

Este benchmark se mantiene gracias a tiempo personal de Cristian Tala Sánchez (Chile) + ~$300/mes en suscripciones simultáneas (Xiaomi, MiniMax, Claude, Ollama Cloud) y gasto continuo de OpenRouter para las actualizaciones. Si te ayudó, comparte el repo. Si querés contribuir, las contribuciones de comunidad se documentan en CHANGELOG.