7 modelos de IA de frontera. Ninguno gana en todo.

Benchmarks reales, precios actualizados, filtraciones de esta semana, y qué significa cada número en tu trabajo del día a día.

Oscar Obando Chaves | Senior AI Counsel | CAIO

oscarobando.ai

Marzo 2026

Siete ecosistemas, un solo trono vacío

Marzo de 2026 fue el mes más intenso en lanzamientos de modelos de IA de la historia. GPT-5.4 llegó el 5 de marzo. Claude Opus 4.6 y Sonnet 4.6 ya estaban desde febrero. Gemini 3.1 Pro apareció el 19 de febrero. Grok 4.20 salió de beta a mediados de marzo. Y del lado open-source, Qwen 3.5 (Alibaba), Mistral Small 4, y GLM-5 (Zhipu AI) completaron una oleada que dejó obsoletos a todos los modelos de hace seis meses.

Los precios de API colapsaron entre 80% y 97% respecto a 2024. Lo que costaba $60 por millón de tokens hace dos años ahora cuesta entre $1 y $2. DeepSeek ofrece capacidades de frontera a $0.28 por millón de tokens de entrada, entre 10 y 60 veces más barato que los modelos premium. La ventana de contexto estándar subió a 1 millón de tokens. Llama 4 Scout de Meta ofrece 10 millones.

Modelo	Empresa	Precio (input/output por 1M tokens)	Contexto	Tipo
GPT-5.4	OpenAI	$2.50 / $15	272K (1M exp.)	Propietario
Claude Opus 4.6	Anthropic	$5 / $25	1M	Propietario
Gemini 3.1 Pro	Google	$2 / $12	1M	Propietario
Grok 4.20	xAI	$2 / $6	128K+	Propietario
DeepSeek V3.2 / R1	DeepSeek	$0.28 / $0.42	128K	Open (MIT)
Llama 4 Scout/Maverick	Meta	~$0.15 / varía	10M / 1M	Open (Llama Lic.)
Qwen 3.5 / Mistral Large 3	Alibaba / Mistral	Gratis-$2 / varía	256K-1M	Open (Apache 2.0)

Fuentes: TechCrunch, Anthropic, Google AI Blog, OpenRouter, Artificial Analysis (mar 2026).

Qué significa cada benchmark en tu día a día

Los benchmarks suenan a laboratorio. SWE-Bench, GPQA Diamond, ARC-AGI. Nadie fuera de la industria sabe qué miden. Aquí está la traducción a situaciones que cualquier profesional reconoce.

Benchmark	Qué mide	Ejemplo real	Humano	Líder IA
SWE-Bench Verified	Resolver bugs reales en repositorios de código	Le das un error del sistema de tu empresa y la IA lo encuentra y corrige sola.	Un ingeniero senior resuelve la mayoría, pero le toma ~1 hora por bug. No hay score formal.	Claude Opus 4.6 (80.8%)
GPQA Diamond	Preguntas de doctorado en física, química, biología	Un profesor le pregunta sobre reacciones químicas avanzadas. Gemini acierta 94 de 100.	PhD en la materia: ~70%. Profesional inteligente sin PhD: 34%. Azar: 25%.	Gemini 3.1 Pro (94.3%)
ARC-AGI-2	Resolver acertijos visuales nunca antes vistos	Le muestras un patrón visual nuevo y debe deducir la regla. Como un acertijo sin instrucciones.	Persona promedio: ~75%. 100% de tareas resueltas por al menos 2 humanos.	Gemini 3.1 Pro (77.1%)
Chatbot Arena (ELO)	6 millones de personas votan cuál respuesta prefieren, a ciegas	Le pides un email o estrategia y eliges cuál te gusta más sin saber qué modelo lo escribió.	Los humanos SON los jueces. No compiten, evalúan.	Claude Opus 4.6 (1503 ELO)
SimpleQA	Preguntas factuales sin alucinar	Le preguntas la capital de un país o quién es el CEO de una empresa. Mide si inventa datos.	Una persona informada acierta casi todo. No hay score formal.	DeepSeek V3.2 (97.1%)
FrontierMath	Matemáticas de nivel investigación	Problemas que solo matemáticos de primer nivel resuelven. Ningún modelo supera el 50%.	Solo matemáticos de investigación activos. No hay baseline público.	GPT-5.4 Pro (47.6%)

Fuentes: Vellum AI Leaderboard (23 mar 2026), LMArena (6M+ votos), Artificial Analysis Index v4.0, Epoch AI, ARC Prize Foundation, arXiv (Rein et al. 2023 para GPQA baselines humanos).

Cuál usar según lo que necesitas hacer

Tu situación	Mejor opción	Por qué
Escribir emails, propuestas, contenido profesional	Claude Opus 4.6	#1 en preferencia humana (6M votos). El tono más natural y matizado. Entiende contexto mejor.
Tu equipo de desarrollo necesita resolver bugs y escribir código	Claude Sonnet 4.6	80% de la calidad de Opus al 60% del precio. Potencia Cursor, Windsurf, GitHub Copilot.
Analizar un documento de 200 páginas o un video de una hora	Gemini 3.1 Pro	Procesa texto, imagen, audio y video como flujo continuo nativo (no frames sueltos como otros modelos). 1M tokens de contexto.
Investigar un tema académico o científico complejo	Gemini 3.1 Pro	94.3% en preguntas de doctorado (GPQA). 77.1% en razonamiento abstracto (ARC-AGI-2). El más preciso en ciencia.
Necesitas procesar miles de documentos al menor costo posible	DeepSeek V3.2	$0.28 por millón de tokens. 50 veces más barato que Claude Opus. Licencia MIT, puedes alojarlo tú mismo.
Quieres un sistema de IA privado, sin enviar datos a terceros	Qwen 3.5 o Llama 4	Gratuitos, código abierto, los instalas en tu servidor. Qwen soporta 201 idiomas bajo Apache 2.0.
Tu empresa quiere implementar IA pero el presupuesto es limitado	Gemini 3.1 Pro	$2/$12 por 1M tokens. Empata en primer lugar del índice compuesto con GPT-5.4. El mejor rendimiento-precio entre propietarios.
Necesitas datos en tiempo real de redes sociales	Grok 4.20	Acceso directo al feed de X (Twitter). Útil para monitoreo de marca y tendencias. $2/$6 por 1M tokens.

El índice compuesto de Artificial Analysis (10 evaluaciones) coloca a Gemini 3.1 Pro y GPT-5.4 empatados en primer lugar con 57 puntos. Claude Opus 4.6 sigue con 53. Pero en preferencia humana real (Chatbot Arena), Claude Opus 4.6 es el #1 indiscutible con 1503 ELO.

Lo que se filtró esta semana y lo que viene después

La última semana de marzo de 2026 trajo dos filtraciones que cambian el panorama. La primera vino de Anthropic. La segunda de OpenAI. Ambas por error.

Claude "Mythos" / Capybara (26 de marzo)

Un error de configuración en el CMS de Anthropic dejó ~3,000 archivos internos accesibles públicamente. Investigadores de seguridad de LayerX y Cambridge los encontraron. Fortune publicó la exclusiva. Los documentos revelaron un modelo llamado "Mythos" bajo un nuevo tier "Capybara", un nivel por encima de Opus. Anthropic lo describió como "un cambio de paso" con "puntuaciones dramáticamente superiores" a Opus 4.6 en codificación, razonamiento y ciberseguridad. La parte alarmante fue la advertencia de que el modelo "plantea riesgos de ciberseguridad sin precedentes" y que está "muy por delante de cualquier otro modelo en capacidades cibernéticas". Las acciones de ciberseguridad cayeron 4-6% al día siguiente. Anthropic confirmó la existencia del modelo pero no hay fecha de lanzamiento público.

OpenAI "Spud" (24-25 de marzo)

The Information reportó que OpenAI completó el pre-entrenamiento de un modelo mayor con nombre clave "Spud". Podría ser GPT-5.5 o GPT-6. Sam Altman dijo internamente que "puede realmente acelerar la economía" y que "un modelo muy poderoso saldrá en unas semanas". Se entrenó en la instalación Stargate en Abilene, Texas, con más de 100,000 GPUs. Coincidiendo, OpenAI cerró Sora (su generador de video que perdía $15 millones por día contra $2.1 millones de ingresos totales de por vida) para redirigir esa capacidad de cómputo a Spud.

Ambas empresas se preparan para OPIs en 2026. La presión por demostrar capacidades de siguiente generación antes de salir a bolsa explica la velocidad. Lo que vemos hoy en los benchmarks podría quedar obsoleto en 60 días.

Fuentes: Fortune exclusive (26 mar 2026), The Information (25 mar 2026), CoinDesk (27 mar 2026), Futurism (27 mar 2026), Tom's Guide (25 mar 2026).

Cuánto cuesta realmente correr un modelo abierto al 100%

Los modelos open-source son gratuitos para descargar. Pero correrlos a capacidad completa necesita hardware serio. Aquí está lo que requiere cada modelo según el número de usuarios, con costos reales de marzo 2026.

Modelo (tamaño real)	1 usuario	1-5 usuarios	5-10 usuarios
Llama 4 Scout 109B total, 17B activos	1x H100 80GB (Q4) ~55-61 GB VRAM ~$1,500-2,200/mes	2x H100 80GB 160 GB VRAM (FP8) ~$3,000-4,400/mes	4x H100 80GB 320 GB VRAM (FP16 completo) ~$6,000-8,800/mes
DeepSeek V3.2 / R1 671B total, 37B activos	8x H100 80GB (AWQ Q4) ~400 GB VRAM ~$12,000-17,000/mes ~33 tok/s	8x H100 (FP8 nativo) ~700 GB VRAM ~$12,000-17,000/mes ~100 tok/s concurrentes	2 nodos de 8x H100 1.28 TB VRAM total ~$24,000-34,000/mes ~620 tok/s concurrentes
Llama 4 Maverick 400B total, 17B activos	3x H100 80GB (Q4) ~200 GB VRAM ~$4,500-6,500/mes	4x H100 80GB 320 GB VRAM ~$6,000-8,800/mes	8x H100 80GB (FP16) 640 GB VRAM ~$12,000-17,000/mes
Qwen 3.5 / Mistral Large 3 397-675B total	Similar a DeepSeek 8x H100 para versión completa ~$12,000-17,000/mes	8x H100 (FP8) ~$12,000-17,000/mes	2 nodos 8x H100 ~$24,000-34,000/mes

Los costos mensuales asumen renta de GPU en proveedores especializados (RunPod, Lambda Labs, Vast.ai) a ~$2-2.50/hora por H100. En AWS o Azure el costo sube 2-3x. La opción más accesible es Llama 4 Scout cuantizado en una sola GPU (~$1,500/mes). Para DeepSeek y los modelos de 671B+ parámetros, el piso realista es $12,000/mes por un clúster de 8 GPUs H100. Comprar el hardware cuesta $200,000-$400,000 por un servidor 8x H100, más electricidad, refrigeración y mantenimiento.

En contraste, una suscripción a Claude Pro, ChatGPT Plus o Gemini Advanced cuesta $20/mes por usuario. Para 10 usuarios son $200/mes. El modelo abierto al 100% empieza a tener sentido económico solo cuando el volumen de procesamiento supera las decenas de miles de consultas diarias, o cuando la privacidad de datos es un requisito no negociable.

Fuentes: GetDeploying.com (41 proveedores H100, mar 2026), GMI Cloud pricing guide, RunPod Llama 4 deployment guide, GitHub dzhsurf/deepseek-v3-r1-deploy-and-benchmarks, Spheron Llama 4 deploy guide (feb 2026), Unsloth quantization benchmarks.

Si solo puedes pagar una suscripción, elige así

Todos los modelos de frontera están disponibles por $20/mes en sus planes básicos. ChatGPT Plus, Claude Pro, Gemini Advanced, SuperGrok. Todos a $20. La diferencia no está en el precio sino en para qué lo vas a usar.

Si tu trabajo es escribir (emails, propuestas, contratos, contenido), Claude Pro es la elección. Es el único modelo que ha ganado la preferencia humana en las tres categorías (texto, código y búsqueda) simultáneamente. La calidad de escritura en español es superior a cualquier competidor según evaluaciones independientes.

Si necesitas investigar, analizar documentos largos, o trabajas con video y audio, Gemini Advanced. Es el modelo con mejor procesamiento nativo de video (entiende el flujo temporal, no solo frames). Tiene el motor de razonamiento científico más preciso del mercado (94.3% GPQA). Y el tier gratuito de Google AI Studio permite hacer pruebas sin pagar nada.

Si trabajas mucho en Excel, hojas de cálculo y análisis de datos, ChatGPT Plus. OpenAI lanzó en marzo "ChatGPT for Excel", un add-in que opera directamente dentro del libro de trabajo. Le dices en lenguaje natural qué necesitas y construye fórmulas, modelos financieros y escenarios. GPT-5.4 tiene el mejor rendimiento en simulaciones de trabajo de oficina real (83% en GDPval, que evalúa 44 ocupaciones). Beta disponible en EE.UU., Canadá y Australia para Plus, Pro, Business y Enterprise.

Si el presupuesto es cero, DeepSeek es gratuito en su web (chat.deepseek.com) y el modelo R1 tiene pensamiento extendido comparable a modelos de $200/mes. Qwen 3.5 también es accesible sin costo. Los dos son modelos chinos, lo que implica consideraciones de privacidad de datos que cada usuario debe evaluar según su contexto.

Lo que no conviene es elegir un solo modelo y asumir que sirve para todo. Los profesionales más productivos en 2026 combinan dos o tres plataformas según la tarea. Claude para escribir y codificar. Gemini para investigar. DeepSeek o ChatGPT para volumen. Cada plataforma hace algo mejor que las demás. Ningún benchmark dice lo contrario.

Fuentes: Chatbot Arena/LMArena (6M votos, mar 2026), Artificial Analysis Index v4.0, LogRocket AI Power Rankings (mar 2026).

OSCAR OBANDO CHAVES

"Vamos a llegar a un punto de commodity. El modelo en sí no va a ser el principal diferenciador."

— Dario Gil, Director de Investigación de IBM (marzo 2026)

oscarobando.ai

linkedin.com/in/oscar-obando-chaves-237495133

Sígueme en LinkedIn para análisis semanal sobre el impacto real de la IA