Estudio Comparativo: Agregadores de Modelos IA

Fecha: 21 junio 2026 | **Investigacion 003**

Resumen Ejecutivo

Comparativa de 8 plataformas agregadoras de modelos de IA analizando modelos gratuitos, precios, limites y peculiaridades.

Ganador Free Tier: Groq

18+ modelos gratuitos, sin tarjeta de credito, hardware LPU propio (560-1000 tokens/s).

1. Comparativa General

Plataforma	Modelos	Free	Peculiaridad
OpenRouter	340+	Muchos :free	Un solo key para Claude, GPT, Gemini, Llama...
ZenMux	136	4 -free	Seguro de calidad: compensacion automatica
Groq	18+	Excelente	Hardware LPU: 560-1000 tok/s. Whisper+TTS gratis
Together AI	26+	Creditos	Batch API 50% descuento
DeepInfra	50+	Ninguno	200 concurrent req. Despliegue desde $0.89/h
Fireworks AI	80+	Pago	FireFunction v2 para tool calling
Replicate	Miles	Trial	Factura por GPU-time. Modelos propios

Novita AI	60+	1 modelo	Alternativa economica. GPU instances

2. Modelos Gratuitos por Plataforma

Groq: 18+ modelos free

Modelo	Contexto	RPM	TPM
llama-3.1-8b-instant	131K	30	6K
llama-3.3-70b-versatile	131K	30	12K
llama-4-scout-17b	131K	30	30K
gpt-oss-120b	131K	30	8K
qwen3-32b	131K	60	6K
whisper-large-v3 (STT)	--	20	--
orpheus-v1 (TTS)	4K	10	1.2K

Limites: 30 RPM / 1K-14.4K RPD / 6K-30K TPM. Sin tarjeta.

OpenRouter: Muchos :free

Destacados: meta-llama/llama-3.3-70b-instruct:free, openrouter/owl-alpha:free, qwen/qwen3-next-80b-a3b-instruct:free, minimax/minimax-m2.5:free

Contexto: 32K-256K. **Rate limits:** ~20 RPM por proveedor.

ZenMux: 4 modelos -free

Modelo	Contexto	Modalidades	Razonamiento
z-ai/glm-5.2-free	1,000,000	Texto	No
stepfun/step-3.7-flash-free	256,000	Texto+Imagen+Video	Si
z-ai/glm-4.7-flash-free	200,000	Texto	Si
z-ai/glm-4.6v-flash-free	200,000	Texto+Imagen+Video	Si

Plus: Seguro de calidad con compensacion automatica.

3. Precios Cruzados: DeepSeek V4 Flash

Proveedor	Input/MTok	Output/MTok	Diferencia
DeepSeek directo	$0.14	$0.28	Referencia
DeepSeek directo (cache hit)	$0.0028	$0.28	50x mas barato input
OpenRouter	$0.09	$0.18	Mas barato (subsidio?)
ZenMux	$0.14	$0.28	= Igual

4. Precios Cruzados: DeepSeek V4 Pro

Proveedor	Input/MTok	Output/MTok	Diferencia
DeepSeek directo	$0.435	$0.87	--
OpenRouter	$0.435	$0.87	= Igual
ZenMux	$0.435	$0.87	= Igual

5. Recomendaciones por Caso de Uso

Caso	Recomendacion	Razon
DeepSeek V4 Flash (tu modelo)	**DeepSeek directo**	Mas barato, cache hit 50x
Maxima velocidad gratuita	**Groq**	LPU: 560-1000 tok/s, 18+ free
Un key para todo	**OpenRouter**	340+ modelos, muchos :free
Produccion con garantia	**ZenMux**	Seguro de calidad incluido
Modelos open-source nicho	**Replicate**	Miles de modelos comunidad
Tool calling intensivo	**Fireworks AI**	FireFunction v2 optimizado

---

Investigacion 003 - Hermes Agent (Nono) - Generado el 2026-06-21

Fuentes: APIs oficiales DeepSeek, OpenRouter, ZenMux + documentacion de cada plataforma