Fastest and most affordable models

Fastest Models ⓘ

Model	Tokens/second
Llama 4 Scout	~2300
Llama 3.3 70b	~2300
Llama 3.1 70b	~2100
Llama 3.1 8b	~1850
Llama 3.1 405b	~1000

Lowest Latency (TTFT) ⓘ

Model	Seconds to first token
Nova Micro	~0.30
Llama 3.1 8b	~0.32
Llama 4 Scout	~0.33
Gemini 2.0 Flash	~0.33
GPT-4o mini	~0.33

Cheapest Models ⓘ ● Input ● Output

USD per 1M tokens

Model	Input	Output
Nova Micro	~0.05	~0.14
Gemma 9.27b	~0.06	~0.06
Gemini 1.5 Flash	~0.07	~0.30
GPT-4o 20b	~0.08	~0.35