expedient AI CTRL Rate Sheet

Generating PDF...

★ STANDARD MODEL

Universal Credits powered by GPT-5 mini

100 CR Credits = **117.65 M** Tokens @

$250 Cost

Based on standard usage patterns with a 9:1 input-to-output ratio. Credits work across all models—token allocations vary by model capability.

● OpenAI ● Anthropic ● Google ● Perplexity ▲ More tokens than GPT-5 mini

OpenAI

GPT-5, GPT-4.1, GPT-4o, and o-series reasoning models

MODEL	INPUT TOKENS / 100 CR	OUTPUT TOKENS / 100 CR
GPT-5 mini ★ STANDARD Balanced model for general AI tasks	200 M	25 M

GPT-4.1 Flagship coding & instruction model	25 M	6.25 M
GPT-4.1 mini Fast, cost-effective	125 M	31.25 M
GPT-4.1 nano Fastest & cheapest	500 M	125 M
GPT-4o Multimodal flagship	20 M	5 M
GPT-4o mini Lightweight multimodal	333.33 M	83.33 M
GPT-5 nano Fastest for summarization & classification	1 B	125 M
GPT-5 pro Smartest and most precise model	3.33 M	416.67 K
GPT-5.1 Best for coding and agentic tasks	40 M	5 M
o3 Advanced reasoning	25 M	6.25 M
o4-mini Efficient reasoning	45.45 M	11.36 M

Anthropic

Claude Opus, Sonnet, and Haiku model families

MODEL	INPUT TOKENS / 100 CR	OUTPUT TOKENS / 100 CR
Claude Haiku 3 Fastest & cheapest	200 M	40 M
Claude Haiku 3.5 Efficient	62.5 M	12.5 M
Claude Haiku 4.5 Near-frontier performance	50 M	10 M
Claude Opus 4.5 Maximum intelligence	10 M	2 M
Claude Sonnet 4 Balanced capability	16.67 M	3.33 M
Claude Sonnet 4.5 Best for coding	16.67 M	3.33 M

Google Gemini

Gemini Pro, Flash, and Flash-Lite models

MODEL	INPUT TOKENS / 100 CR	OUTPUT TOKENS / 100 CR
Gemini 2.0 Flash Balanced multimodal	500 M	125 M
Gemini 2.0 Flash-Lite Ultra-low cost	666.67 M	166.67 M
Gemini 2.5 Flash Hybrid reasoning, 1M context	166.67 M	20 M

Gemini 2.5 Flash-Lite Most cost-effective	500 M	125 M
Gemini 2.5 Pro Coding & complex reasoning	40 M	5 M
Gemini 3 Pro Most powerful agentic model	25 M	4.17 M

Perplexity

Sonar search-grounded models

MODEL	INPUT TOKENS / 100 CR	OUTPUT TOKENS / 100 CR
Sonar Lightweight search	50 M	50 M
Sonar Deep Research Exhaustive research reports	25 M	6.25 M
Sonar Pro Deep content understanding	16.67 M	3.33 M
Sonar Reasoning Step-by-step logic	50 M	10 M
Sonar Reasoning Pro Enhanced multi-step reasoning	25 M	6.25 M

Understanding This Rate Card

- Higher-capability models consume more credits per token
- Efficient models stretch your credits further
- All values based on 100 Credits
- GPT-5 mini is the baseline standard model
- Green values indicate more tokens than GPT-5 mini
- Rates are subject to change by providers
- Actual usage varies based on conversation patterns
