

# LLM Verfügbarkeitsübersicht

aihub.biteno.com | Stand: 2026-06-07 | 18 lokal + 45 remote Modelle

## Lokal gehostete Modelle (18)

| Modell                   | Input / 1M Token | Output / 1M Token | Kontext |
|--------------------------|------------------|-------------------|---------|
| deepseek-r1:14b          | \$0.5000         | \$0.9000          | —       |
| deepseek-r1:32b          | \$0.5000         | \$0.9000          | —       |
| gemma3:27b               | \$0.5000         | \$0.9000          | —       |
| gemma4:31b               | \$0.3000         | \$0.9000          | —       |
| glm-4.7-flash:latest     | \$0.2500         | \$0.5000          | —       |
| gpt-oss:20b              | \$0.2500         | \$0.5000          | —       |
| lfm2.5-thinking:1.2b     | \$0.5000         | \$0.9000          | —       |
| lfm2:24b                 | \$0.5000         | \$0.9000          | —       |
| llama3.1-8b              | \$0.2500         | \$0.4000          | —       |
| minimax-m2.7             | \$0.3500         | \$1.2500          | —       |
| mistral-small-24k:24b    | \$0.5000         | \$0.9000          | —       |
| mistral-small3.2-32k:24b | \$0.2500         | \$0.5000          | —       |
| nemotron-3-nano:4b       | \$0.2500         | \$0.4000          | —       |
| phi4:latest              | \$0.2500         | \$0.5000          | —       |
| qwen3-vl-8B-instruct     | \$0.2500         | \$0.4000          | —       |
| qwen3.6:35b-a3b          | \$0.2500         | \$0.5000          | —       |
| qwen3:14b                | \$0.2500         | \$0.5000          | —       |
| qwen3:32b                | \$0.5000         | \$0.9000          | —       |

## Remote / API Modelle (45)

| Modell                   | Input / 1M Token | Output / 1M Token | Kontext |
|--------------------------|------------------|-------------------|---------|
| remote-claude-3-5-haiku  | —                | —                 | —       |
| remote-claude-3-5-sonnet | —                | —                 | —       |

| Modell                           | Input / 1M Token | Output / 1M Token | Kontext |
|----------------------------------|------------------|-------------------|---------|
| remote-claude-3-7-sonnet-latest  | —                | —                 | —       |
| remote-claude-4-sonnet-20250514  | \$3.0000         | \$15.0000         | 64k     |
| remote-claude-haiku-4-5          | \$1.0000         | \$5.0000          | 64k     |
| remote-claude-opus-4-5           | \$5.0000         | \$25.0000         | 64k     |
| remote-claude-opus-4-6           | \$5.0000         | \$25.0000         | 128k    |
| remote-claude-opus-4-7           | \$5.0000         | \$25.0000         | 128k    |
| remote-claude-opus-4-8           | \$5.0000         | \$25.0000         | 128k    |
| remote-claude-sonnet-4-5         | \$3.0000         | \$15.0000         | 64k     |
| remote-claude-sonnet-4-6         | \$3.0000         | \$15.0000         | 64k     |
| remote-deepseek-chat             | \$0.2800         | \$0.4200          | 8k      |
| remote-deepseek-coder            | \$0.1400         | \$0.2800          | 4k      |
| remote-deepseek-reasoner         | \$0.2800         | \$0.4200          | 65k     |
| remote-gemini-2.0-flash-exp      | —                | —                 | —       |
| remote-gemini-2.0-flash-thinking | —                | —                 | —       |
| remote-gemini-2.5-flash          | \$3.0000         | \$15.0000         | 65k     |
| remote-gemini-2.5-flash-lite     | \$3.0000         | \$15.0000         | 65k     |
| remote-gemini-3-pro-config       | \$3.0000         | \$15.0000         | 65k     |
| remote-gemini-3.1-pro-preview    | \$3.0000         | \$15.0000         | 65k     |
| remote-gemini-3.5-flash          | \$1.5000         | \$9.0000          | 65k     |
| remote-gpt-4                     | \$30.0000        | \$60.0000         | 4k      |
| remote-gpt-4-turbo               | \$10.0000        | \$30.0000         | 4k      |
| remote-gpt-4.1-mini              | \$0.4000         | \$1.6000          | 32k     |
| remote-gpt-4.5-preview           | —                | —                 | —       |
| remote-gpt-4o                    | \$2.5000         | \$10.0000         | 16k     |
| remote-gpt-4o-mini               | \$0.1500         | \$0.6000          | 16k     |
| remote-gpt-5                     | \$1.2500         | \$10.0000         | 128k    |
| remote-gpt-5-mini                | \$0.2500         | \$2.0000          | 128k    |
| remote-gpt-5.2-codex             | \$1.7500         | \$14.0000         | 128k    |

| Modell                                | Input / 1M Token | Output / 1M Token | Kontext |
|---------------------------------------|------------------|-------------------|---------|
| remote-gpt-5.2-pro                    | \$21.0000        | \$168.0000        | 128k    |
| remote-grok2-latest                   | \$2.0000         | \$10.0000         | 131k    |
| remote-kimi-k2.5                      | \$0.6000         | \$3.0000          | 262k    |
| remote-kimi-k2.6                      | \$0.9500         | \$4.0000          | 262k    |
| remote-minimax-m2.7                   | \$0.3000         | \$1.2000          | —       |
| remote-o1                             | \$15.0000        | \$60.0000         | 100k    |
| remote-openrouter-deepseek-v3         | \$0.3000         | \$1.5000          | 8k      |
| remote-openrouter-gemini-2.0-flash    | \$0.3000         | \$1.5000          | 8k      |
| remote-openrouter-gemini-2.5-pro      | \$0.3000         | \$1.5000          | —       |
| remote-openrouter-gemini-3-pro        | \$0.3000         | \$1.5000          | 65k     |
| remote-perplexity-sonar-deep-research | —                | —                 | —       |
| remote-perplexity-sonar-reasoning-pro | —                | —                 | —       |
| remote-xai-grok-3                     | \$3.0000         | \$15.0000         | 131k    |
| remote-xai-grok-3-mini                | \$0.3000         | \$0.5000          | 131k    |
| remote-xai-grok-4                     | \$3.0000         | \$15.0000         | 256k    |