1NCE AI **Pricing**

Regular Models - Germany Hosted (eu-central-1)

Model ID	AWS model	Input (USD / 1M)	Output (USD / 1M)
anthropic.claude-opus-4-8	Claude Opus 4.8	$5.50	$27.50
anthropic.claude-opus-4-7	Claude Opus 4.7	$5.50	$27.50
anthropic.claude-opus-4-6-v1	Claude Opus 4.6	$5.50	$27.50
anthropic.claude-sonnet-4-6	Claude Sonnet 4.6	$3.30	$16.50
anthropic.claude-haiku-4-5-20251001-v1:0	Claude Haiku 4.5	$1.10	$5.50
amazon.nova-2-lite-v1:0	Nova 2.0 Lite	$0.429	$3.597
amazon.nova-pro-v1:0	Nova Pro	$1.05	$4.20
mistral.devstral-2-123b	Devstral	$0.48	$2.40
mistral.pixtral-large-2502-v1:0	Pixtral Large 25.02	$2.00	$6.00
openai.gpt-oss-120b-1:0	gpt-oss-120b	$0.20	$0.79
openai.gpt-oss-20b-1:0	gpt-oss-20b	$0.09	$0.40
qwen.qwen3-235b-a22b-2507-v1:0	Qwen3 235B A22B 2507	$0.29	$1.16
qwen.qwen3-coder-30b-a3b-v1:0	Qwen3 Coder 30B A3B	$0.20	$0.79

All prices in USD per 1 million tokens

Regular Models - Germany Hosted (eu-central-1)

Model ID	AWS model	Input (USD / 1M)	Output (USD / 1M)
minimax.minimax-m2.5	MiniMax M2.5	$0.36	$1.44
minimax.minimax-m2.1	Minimax M2.1	$0.36	$1.44
zai.glm-4.7-flash	GLM 4.7 Flash	$0.08	$0.48
nvidia.nemotron-super-3-120b	NVIDIA Nemotron 3 Super 120B A12B	$0.18	$0.78

All prices in USD per 1 million tokens

Embedding Models - Germany Hosted (eu-central-1)

Embeddings are input-only (no output token charge).

Model ID	AWS model	Input (USD / 1M)	Notes
cohere.embed-multilingual-v3	Cohere Embed 3 Multilingual	$0.10	+ $0.0001 per image
cohere.embed-v4:0	Cohere Embed 4	$0.12	text input tokens
amazon.titan-embed-image-v1	Titan Embeddings G1 Image	$1.00	+ $0.0001 per image
amazon.titan-embed-text-v1	Titan Embeddings G1 Text	$0.20	text input tokens

All prices in USD per 1 million tokens

Optional - OpenAI Frontier Models (US Hosted via AWS Bedrock - US East: N. Virginia us-east-1 / Ohio us-east-2)

Model	Input (USD / 1M)	Output (USD / 1M)	Cached input (USD / 1M)
OpenAI — GPT-5.5	$5.50	$33.00	$0.55
OpenAI — GPT-5.4	$2.75	$16.50	$0.275

All prices in USD per 1 million tokens

Notes

- "Standard" = on-demand, synchronous (excludes batch, prompt-cache, flex/priority, and provisioned throughput). The OpenAI frontier table also lists the cached-input rate as provided.
- Claude rates are the Geo / In-region (EU) profile. The Global cross-region profile is ~10% cheaper (e.g. Opus 4.x $5.00/$25.00, Sonnet 4.6 $3.00/$15.00, Haiku 4.5 $1.00/$5.00) but can route outside the EU.
- API-sourced (eu-central-1, on-demand, verified): Nova, Devstral, Pixtral, gpt-oss, Qwen3, MiniMax, GLM, Nemotron, Titan embeddings. Website-sourced: Claude + Cohere (Frankfurt), OpenAI GPT-5.x (US East).