

Model	Pricing model	Free Tier	License	Self-host
Mistral Large 3	Per-token API (\$0.5/M input, \$1.5/M output)	No verified API-specific free tier found on Mistral's site	Open-weight	Yes — models can run on customer infrastructure via vLLM
AI21 Jamba (Jamba2 / Jamba Large 1.7)	Per-token API, tiered by model (\$0.2/\$0.4 per M for Mini; \$2/\$8 per M for Large)	\$10 in credit for three months on new accounts	Apache 2.0 (Jamba2); AI21's own open licensing framework (Jamba Large 1.7)	Yes — self-hosting highlighted for regulated data
ZeroEntropy zerank-2	Per-token API (\$0.025/ M tokens)	No free tier found on ZeroEntropy's site	Non-commercial license; commercial use requires a separate agreement with ZeroEntropy	Yes — on-premise/VPC deployment for regulated industries
BGE Reranker (v2-m3)	No licensing cost (self-hosted)	Free, no extra cost	Open-source	Yes — model runs locally with no API key or billing step required
Jina Reranker (m0)	Usage-based billing beyond free allowance	10 million free tokens per new API key	CC-BY-NC 4.0 for self-hosted open weights; no restrictions via Jina's official API or cloud marketplace listings	Yes — open-weight version can be self-hosted