Local Model Services Description

API Name	Model Name	Model ID	Context Length	Description
DeepSeek-V4-Flash	DeepSeek-V4-Flash	deepseek-chat	1,048,576	DeepSeek lightweight fast-response model, 284B parameters with million-token context, optimized for high-concurrency dialogue, real-time interaction, and batch Q&A scenarios where latency is critical, delivering millisecond-level first-token response
DeepSeek-V4-Pro	DeepSeek-V4-Pro	deepseek-pro	1,048,576	DeepSeek V4 series high-performance model, 1.6T total / 49B active parameters with million-token context, balancing inference speed and output quality, suitable for production environments, multi-turn dialogue, and complex Q&A where both quality and speed are required
GLM5.1	GLM-5.1	glm-chat	202,752	Zhipu GLM series flagship model, 754B MoE / 40B active, focused on Agent engineering and complex coding tasks, excelling at long-range and complex agent tasks with sustained iterative optimization, ideal for document processing, knowledge extraction, multi-turn dialogue, and content creation
MiniMax2.7	MiniMax M2.7	minimax	196,608	MiniMax's first flagship model with deep self-evolution capability, capable of building complex agent frameworks, efficiently handling highly nuanced production tasks in software engineering and professional office scenarios through agent teams, complex skills, and dynamic tool search
DeepSeek-R1 API	DeepSeek-R1-0528	`deepseek-r1:671b`	65,536	DeepSeek reasoning flagship model, 685B MoE architecture, excelling at complex reasoning, mathematical proofs, and code generation, supporting Chain-of-Thought reasoning, suitable for scientific computing, algorithm design, and logical reasoning tasks
DeepSeek-V3.2 API	DeepSeek-V3.2	`deepseek-v3:671b`	131,072	DeepSeek general flagship model, 685B MoE architecture, balancing speed and quality, excelling at bilingual (Chinese/English) dialogue, document understanding, and knowledge Q&A, suitable for long document processing, multi-turn dialogue, and content creation
Qwen-instruct API	Qwen3.5-397B-A17B	`qwen-instruct`	262,144	Alibaba Qwen next-generation open-source flagship model, 397B total / 17B active parameters with hybrid attention + sparse MoE architecture, designed for the agent era, supporting text and multimodal tasks, suitable for multilingual dialogue, code generation, and agent application scenarios
DeepSeek-Math-V2 API	DeepSeek-Math-V2	`deepseek-math`	163,840	DeepSeek mathematics-specialized model, optimized for mathematical reasoning, excelling at formula derivation, symbolic computation, mathematical proofs, and scientific computing, suitable for academic research, engineering computation, and educational tutoring
Qwen-code API	Qwen3-Coder-Next	`qwen-code`	262,144	Alibaba Qwen coding-specialized model, 80B total / 3B active parameters MoE architecture, designed for coding agents, excelling at code generation, debugging and optimization, multi-tool programming tasks, suitable for software development, code review, and automated programming scenarios

API Name	Model Name	Model ID	Context Length	Description
BGE-M3 API	BGE-M3	`bge-m3`	-	BAAI multilingual embedding model, supporting Chinese, English, and multilingual text vectorization, suitable for semantic search, document clustering, similarity computation, and RAG vector database construction
bge-reranker-v2-m3 API	bge-reranker-v2-m3	`bge-reranker`	-	BAAI reranking model for precision ranking of search results, used in conjunction with BGE-M3 to significantly improve RAG retrieval accuracy, suitable for search systems, Q&A systems, and recommendation systems

Application Notes

Local model services are delivered relying on limited server resources. To ensure service quality and reasonable resource allocation, please estimate the access frequency (e.g., daily requests, peak requests per minute) when applying, so we can better plan resources and ensure service quality.

Last Updated: 2026-04-26