本地模型服务说明

接口名称	模型名称	Model ID	上下文长度	特点说明
DeepSeek-V4-Flash	DeepSeek-V4-Flash	deepseek-chat	1048576	DeepSeek 轻量级快速响应模型，2840亿参数，拥有百万字超长上下文，适合高并发对话、实时交互、批量问答等延迟敏感场景，为用户提供毫秒级首 token 响应的流畅体验
DeepSeek-V4-Pro	DeepSeek-V4-Pro	deepseek-pro	1048576	DeepSeek V4 系列高性能版本，1.6万亿参数/激活49B，拥有百万字超长上下文，平衡推理速度与输出质量，适合对响应质量和速度均有要求的生产环境、多轮对话、复杂问答等场景
GLM5.1	GLM-5.1	glm-chat	202752	智谱 GLM 系列旗舰模型，754B MoE 架构/激活40B，专注于 Agent 工程与复杂编码任务，擅长处理长程、复杂的智能体任务，可长时间持续迭代优化，适合文档处理、知识抽取、多轮对话、内容创作
MiniMax2.7	MiniMax M2.7	minimax	196608	MiniMax首个深度参与自身演化的旗舰模型，能够构建复杂的智能体框架，借助智能体团队、复杂技能和动态工具搜索，高效完成软件工程、专业办公等高度精细的生产力任务
DeepSeek-R1 模型接口	DeepSeek-R1-0528	`deepseek-r1:671b`	65,536	DeepSeek 推理旗舰模型，685B 参数 MoE 架构，擅长复杂推理、数学证明、代码生成，支持长链式思维（Chain-of-Thought），适合科研计算、算法设计、逻辑推理任务
DeepSeek-V3.2 模型接口	DeepSeek-V3.2	`deepseek-v3:671b`	131,072	DeepSeek 通用旗舰模型，685B 参数 MoE 架构，平衡速度与质量，擅长中英双语对话、文档理解、知识问答，适合长文档处理、多轮对话、内容创作
Qwen-instruct 模型接口	Qwen3.5-397B-A17B	`qwen-instruct`	262,144	阿里通义千问新一代开源旗舰模型，397B 总参数 / 17B 激活参数，混合注意力 + 稀疏 MoE 架构，为智能体时代设计，支持文本和多模态任务，适合多语言对话、代码生成、智能体应用场景
DeepSeek-Math-V2 模型接口	DeepSeek-Math-V2	`deepseek-math`	163,840	DeepSeek 数学专用模型，专为数学推理优化，擅长公式推导、符号计算、数学证明、科学计算，适合学术研究、工程计算、教育辅导场景
Qwen-code 模型接口	Qwen3-Coder-Next	`qwen-code`	262,144	阿里通义千问编程专用模型，80B 总参数 / 3B 激活参数 MoE 架构，专为编程智能体设计，擅长代码生成、调试优化、多工具编程任务，适合软件开发、代码审查、自动化编程场景
BGE-M3 模型接口	BGE-M3	`bge-m3`	-	智源研究院多语言嵌入模型，支持中文、英文、多语言文本向量化，适合语义检索、文档聚类、相似度计算、RAG 向量数据库构建
bge-reranker-v2-m3 模型接口	bge-reranker-v2-m3	`bge-reranker`	-	智源研究院重排序模型，用于检索结果精排，配合 BGE-M3 使用可提升 RAG 检索精度，适合搜索系统、问答系统、推荐系统

申请说明

本地模型服务依托有限的服务器资源进行交付。为确保服务质量与资源合理配置，申请时请预估对应模型的访问频次（如日均请求次数、峰值频次/分钟），以便我们更好地进行资源规划与服务保障。

最后更新： 2026-04-26