

大模型服务 版本说明

产品版本: v7.0.2

发布日期: 2026-10-08

目录

1 版本说明	1
1.1 版本说明书	1

1 版本说明

1.1 版本说明书

版本信息

产品名称	产品版本	发布日期
Beacon	V7.0.2	2026-09-30

更新说明

Beacon V7.0.2 是 Beacon 大模型服务的首个正式发布版本。

Beacon 用于统一管理模型资产、部署模型推理服务，并向业务应用提供统一的模型访问能力。本版本提供模型广场、模型仓库、推理引擎、模型部署、AI 服务提供者、模型网关、模型体验、基准测试、套餐管理、API Key 和 Token 用量等功能，支持模型的统一接入、部署、调用和管理。

新增功能

模型广场

支持创建和管理模型，并按照模型提供者、模型类型和任务类型检索模型。创建模型时，可以绑定模型仓库中的已有模型文件，也可以导入新的模型文件。

模型仓库

支持通过 ModelScope、Hugging Face 地址导入模型文件或从本地上传模型文件，并通过公共仓库和私有仓库统一存储及管理模型文件。支持设置仓库容量、查看任务进度和管理 LoRA 微调模型文件。

推理引擎

支持配置模型部署所需的推理环境，包括 CPU、内存、AI 加速卡型号与数量、系统架构、推理镜像、启动命令、启动参数和环境变量。

支持 NVIDIA GPU、海光 HCU 和昇腾 NPU 等 AI 加速卡，并提供基于 vLLM 和 SGLang 的推理引擎镜像。

模型部署

支持选择模型文件和推理引擎，将模型文件部署为在线推理服务。支持配置部署副本数及模型名称，并对模型部署执行启动、停止、重命名和删除操作。

支持查看模型部署的 Pod 状态、运行监控、容器日志和终端。

AI 服务提供者

支持创建内部和外部 AI 服务提供者。内部 AI 服务提供者可以关联平台中的模型部署；外部 AI 服务提供者可以配置 OpenAI 兼容的访问地址和访问密钥。

模型网关

支持通过模型网关为模型服务创建统一访问入口，并配置路由域名、对外模型及 AI 服务提供者后端。

同一条路由可以包含多个模型，同一模型可以关联多个后端，并按照优先级和权重进行服务选择。路由还可以关联 IP 黑白名单进行访问控制。

模型体验

支持通过 API Key、模型网关路由和模型直接验证模型服务能力，体验对话、图片生成、视觉或 OCR、语音等模型能力。

基准测试

支持选择已就绪的模型部署创建基准测试，并查看输入吞吐、输出吞吐、首 Token 延迟、单 Token 生成耗时、请求延迟、错误率和测试耗时等结果。

套餐管理

支持通过套餐管理模型服务的访问范围和调用限额。用户可以配置请求限速、Token 限速及套餐允许访问的模型网关路由，并将套餐共享给其他项目。

API Key

支持创建 API Key 并自动签发密钥。API Key 可以绑定套餐，并通过签发的密钥访问套餐授权的模型网关路由。

同一 API Key 可以持有多把密钥，并支持对密钥单独进行禁用和删除，以便进行密钥轮换。

Token 用量

支持按照部门、项目、用户、API Key、入口域名和模型查看模型服务的调用汇总、最近请求及 Token 消耗情况，并支持导出报表和查看模型用量趋势。

依赖说明

- 平台版本至少为 V7.0.2。
- 安装本产品前需要安装安全容器服务 V7.0.2、容器镜像服务 V7.0.2。

咨询热线：400-100-3070

北京易捷思达科技发展有限公司：

北京市海淀区西北旺东路10号院东区23号楼华胜天成科研大楼一层东侧120-123

南京分公司：

江苏省南京市雨花台区软件大道168号润和创智中心B栋一楼西101

上海office：

上海黄浦区西藏中路336号华旭大厦22楼2204

成都分公司：

成都市高新区天府五街168号德必天府五街WE602

邮箱：

contact@easystack.cn (业务咨询)

partners@easystack.cn(合作伙伴咨询)

marketing@easystack.cn (市场合作)

training@easystack.cn (培训咨询)

hr@easystack.cn (招聘咨询)