

大模型服务平台 TokenHub

模型推理



腾讯云

【 版权声明 】

©2013–2026 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

模型推理

推理模式介绍

TPM 预留

模型单元

模型推理

推理模式介绍

最近更新时间：2026-08-12 17:47:30

大模型服务平台 TokenHub 提供按量付费、TPM 预留、模型单元三种推理模式，覆盖从灵活接入、日常生产到吞吐保障、专属资源的不同诉求。本文介绍三种模式的区别，帮助您根据业务场景选择合适的模式。

推理模式

按量付费

按量付费是平台默认的推理模式。您无需提前预留资源，按照实际消耗的 Token 量后付费，随用随付，适用于业务起步、调用量波动较大或希望灵活控制成本的场景。各模型的按量价格请参见 [模型价格](#)。

TPM 预留

TPM 预留是预付费的吞吐保障模式。您按 kTPM（千 Tokens Per Minute）粒度提前预留吞吐容量，开通后在已购 TPM 额度、适用模型和地域范围内，可提供相应的预留吞吐能力，有助于降低因共享资源高峰导致的限流风险。详细介绍请参见 [TPM 预留](#)。

模型单元

模型单元是资源独占的专属模型服务模式，为您提供专属推理资源及相应的资源隔离能力，并由平台统一托管运维，适用于对数据安全、资源隔离和专属资源有较高要求的业务场景。具体隔离范围、部署形态、网络边界和安全责任以产品文档及订单约定为准。详细介绍请参见 [模型单元](#)。

模式对比

对比项	按量付费	TPM 预留	模型单元
计费模式	后付费，按实际 Token 消耗计费	预付费，按月购买	预付费，按月购买
保障内容	不提供吞吐保障	按 kTPM 粒度预留专属吞吐容量	专属的单位容量推理资源
资源形态	共享资源	共享资源 + 预留吞吐配额	专属推理资源
典型场景	业务起步、调用量波动大	稳定生产吞吐、高峰防限流	资源隔离、数据安全
开通方式	默认开通，无需申请	联系商务经理或 提交工单	联系商务经理或 提交工单

是否独享资源	否	否（共享资源 + 预留吞吐配额）	是
是否承诺固定吞吐	否	在已购额度内承诺	提供单个单元容量预估，按需扩容
适用模型/地域	全部模型、全部地域	GLM 5.2（广州）	通用模型单元 B（广州）
超额或超出保障额度后的处理	不适用	超出预留部分按标准按量付费价格计费	超出单元容量后会出现请求错误
是否支持变配、续费及退费	不适用	支持变配 暂不支持自动续费（后续版本将支持），退费需提交工单，退款按订单剩余价值计算，具体金额以工单审核结果和订单约定为准	支持变配 暂不支持自动续费（后续版本将支持），退费需提交工单，退款按订单剩余价值计算，具体金额以工单审核结果和订单约定为准

如何选择



相关文档

- [TPM 预留](#)

- [模型单元](#)
- [计费方式](#)
- [模型价格](#)

TPM 预留

最近更新时间：2026-08-12 17:47:30

什么是 TPM 预留

TPM 预留是大模型服务平台 TokenHub 提供的预付费吞吐保障模式。TPM（Tokens Per Minute）表示模型每分钟可处理的 Token 量。开通 TPM 预留后，在已购 TPM 或保障包额度、适用模型和地域范围内，可提供相应的预留吞吐能力，有助于降低因共享资源高峰导致的限流风险；超过已购额度或不满足适用条件的请求，按产品规则处理。具体保障范围、限流规则及适用条件以控制台和产品文档为准。

适用场景

- **稳定生产吞吐**：业务已上线生产且调用量相对稳定，需要明确的吞吐保障。
- **高峰防限流**：业务存在明显流量高峰，需要保障高峰时段每分钟可处理的 Token 量。

计费说明

TPM 预留采用预付费模式，按月购买，价格请参见 [TPM 预留价格](#)。超过预留 TPM 额度后，超出部分按标准按量付费价格计费，价格请参见 [模型价格](#)。

如何使用

⚠ 注意：

TPM 预留需联系您的商务经理开通后方可使用。

1. 登录 [TokenHub 控制台](#)，在左侧导航选择**推理服务** > **在线推理**。
2. 单击**创建自定义推理服务**。
3. 在模型列表中选择需要使用的模型。
4. 在**计费方式**中选择 **TPM 预留**（预留固定 TPM 吞吐额度保障稳定调用，超出预留部分按 Token 计费）。
5. 按需配置购买额度（输入与输出 kTPM）与购买时长。TPM 预留暂不支持自动续费，后续版本将支持。
6. 确认费用预估并同意相关服务协议后提交订单，完成创建。

控制台配置示例（GLM 5.2）

TokenHub

快速接入

模型广场

模型体验

语言模型

视觉模型

多模态理解

推理服务

在线推理

批量推理

观测

用量统计

模型监控

平台管理

启用管理

API Key 管理

平台配置

Token Plan

Coding Plan

Token Plan 个人版

Token Plan 企业版

老板控制台

港元大模型 API

DeepSeek API

创建自定义推理服务

选择模型

语言模型

视觉模型

多模态理解

向量模型

语音模型

搜索模型名称

K

与 Kimi K2.7 Code 相同但输出速度约为普通版的 5-6 倍，常规编程场景下（取输入长度中位数）输出速度约 180 Token/s，短上下文场景可达 260 Token/s，支持 256...

Z

GLM-5.2

深度思考

文本生成

GLM-5.2 是智谱最新旗舰模型，支持1M上下文及思考长度控制，在coding及agent场景表现优异

K

Kimi K2.7 Code

文本生成

深度思考

视觉理解

Kimi 迄今最智能的 Coding 模型，在长上下文中更可可靠地遵循指令，能以更高的成功率完成编程任务，同时支持文本、图片与视频输入，思考模式，对话与 Agent 任务。

MiniMax-M3

文本生成

深度思考

视觉理解

MiniMax-M3作为MiniMax最新发布的模型，在编程和智能体等专业任务上达到了前沿的能力，它使用了全新注意力架构 MSA（MiniMax Sparse Attention），最高支持 ...

DeepSeek-V4-Pro

原厂直供

深度思考

文本生成

由 DeepSeek 直接提供的 DeepSeek V4 Pro 模型服务，TokenHub 不对该服务提供 SLA 保障。使用该模型即视为您已知晓并同意遵守 DeepSeek 的服务协议，请您在...

DeepSeek-V4-Pro

文本生成

深度思考

DeepSeek-V4-Pro 1.6T 参数的原生多模态旗舰，通过全新的 CSA+HCA 混合注意力架构，在复杂数学推理、长程代码工程及深度智能体协作领域代表了当前的行业顶...

MiMo-V2.5-Pro

文本生成

深度思考

计费方式

默认开启免费体验额度（剩余额度 0 / 1000000 tokens） 免费额度用完后将按下方选择的计费方式继续计费

收起详情

按 Token 计费

按实际消耗的 Token 数量计费，无最低消费，用多少付多少。

按量计费区

固定收费区

✓ 零门槛，即开即用

✓ 无最低消费，无预付费

✓ 按实际输入/输出 Token 数计费

✓ 适合开发测试和中小流量场景

TPM 预留

预留固定TPM吞吐额度保障稳定调用，超出预留部分按Token计费。

按量计费区

固定收费区

✓ 保障业务稳定调用

✓ 超出部分按量付费

✓ 适用有稳定容量需求的场景

TPM 保障

预留固定TPM吞吐额度保障稳定调用，超出预留部分按Token计费。

按量计费区，性能更低

固定收费区

✓ 更好的性能表现

✓ 保障业务稳定调用

✓ 超出部分按量付费

✓ 适用要求高性能且有稳定容量需求的场景

按模型单元计费

独享 GPU 模型单元，按实例时长计费，完全隔离，延迟最低。

按量计费区

固定收费区

✓ 独享 GPU 模型单元

✓ 完全资源隔离

✓ 延迟最低

✓ 适合大规模生产和高并发场景

计费类型

包月购买 | 预付费

购买额度

输入

10

KTPM

输出

1

KTPM

当前可用额度：100 KTPM

当前可用额度：10 KTPM

购买时长

-

1

+

月

调用额度限制

☐ 启用服务级别限流（开启后可设置服务维度的 TPM / RPM 限制，上限不超过模型值）

标签

标签键

标签值

+

添加标签

🔗

键值粘贴板

🕒

历史记录

模型服务变更

☐ 我已知晓并同意模型服务变更说明

平台上的模型可能因技术迭代、服务优化、合作能力调整、法律法规等原因发生调整、升级、替换、下架或停止提供服务、价格调整的情况；当前可用的模型不代表长期、永久或持续不断地可用、价格不变。请您及时留意官网通知；我们将尽合理努力为您提供优质服务，并将在上述情况发生时，结合具体情况协助您妥善处理已购权益。

协议条款

☐ 我已阅读并同意《大模型服务平台 TokenHub 隐私政策》[《大模型服务平台 TokenHub 服务等级协议》](#)[《自动续费规则》](#)和[《大模型服务平台 TokenHub 服务条款》](#)

注意事项

- 暂不支持变配。
- 暂不支持自助退费。如需申请退费，请 [提交工单](#)，退费规则请参见 [预付费云资源退款规则说明](#)，按订单剩余价值退款，具体退费金额以工单审核结果和订单约定为准。

更多权益边界（是否独享资源、是否承诺固定吞吐、适用模型/地域、超额或超出保障额度后的处理、是否支持变配/续费/退费等）请参见 [推理模式介绍](#)。

相关文档

- [推理模式介绍](#)
- [模型单元](#)
- [模型价格](#)

版权所有：腾讯云计算（北京）有限责任公司

第8 共11页

模型单元

最近更新时间：2026-08-12 17:47:30

什么是模型单元

模型单元是大模型服务平台 TokenHub 提供的专属模型服务模式。模型单元为您提供专属推理资源及相应的资源隔离能力，并由平台统一托管运维，适用于对资源隔离有较高要求的业务场景；具体隔离范围、部署形态、网络边界和安全责任以产品文档及订单约定为准。

适用场景

- **资源隔离**：需要专属推理资源、不与其他用户共享资源的业务。
- **稳定专属资源**：需要长期稳定的专属资源，并希望由平台托管运维的生产业务。

计费说明

模型单元采用预付费模式，按月购买。价格请参见 [模型单元价格](#)。

如何使用

⚠ 注意：

模型单元需联系您的商务经理开通后方可使用。

1. 登录 [TokenHub 控制台](#)，在左侧导航选择**推理服务** > **在线推理**。
2. 单击**创建自定义推理服务**。
3. 在模型列表中选择需要使用的模型。
4. 在**计费方式**中选择**按模型单元计费**。
5. 按需配置资源规格、模型单元数量与购买时长。
6. 确认费用预估并同意相关服务协议后提交订单，完成创建。

TokenHub

快速接入

模型广场

模型体验

语言模型

视觉模型

多模态理解

推理服务

在线推理

批量推理

观测

用量统计

模型监控

平台管理

启用管理

API Key 管理

平台配置

配额管理

网络配置

系统配置

Token Plan

Coding Plan

Token Plan 个人版

Token Plan 企业版

老板控制台

混元大模型 API

DeepSeek API

创建自定义推理服务

按 Token 计费

TPM 预留

TPM 保障

按模型单元计费

Deepseek-v4-flash

DeepSeek-V4-Flash 正式版

Kimi K3

Kimi K2.7 Code HighSpeed

GLM-5.2

Kimi K2.7 Code

MiniMax-M3

计费方式

默认开启免费体验额度（剩余额度 0 / 1000000 tokens） 免费额度用完后将按下方选择的计费方式继续计费

收起详情

计费方式

预付费(包月) 包销

资源规格

PD 分离

模型单元数量

32

必须为 32 的整数倍，当前可购买的最大单元数量为 128

购买时长

1个月 3个月 6个月 12个月 自定义时长

1 月

自动续费

开启自动续费

标签

标签键 标签值

+ 添加标签 | 键值粘贴板 历史记录

模型服务变更

☐ 我已知晓并同意模型服务变更说明

平台上的模型可能因技术迭代、服务优化、合作能力调整、法律法规等原因发生调整、升级、替换、下架或停止提供服务、价格调整的情况；当前可用的模型不代表长期、永久或持续不断地可用、价格不变。请您及时留意官网通知；我们将尽合理努力为您提供优质服务，并将在上述情况发生时，结合具体情况协助您妥善处理已购权益。

注意事项

- 暂不支持变配。
- 暂不支持自助退费。如需申请退费，请 [提交工单](#)，退费规则请参见 [预付费云资源退款规则说明](#)，按订单剩余价值退款，具体退费金额以工单审核结果和订单约定为准。

更多权益边界（是否独享资源、是否承诺固定吞吐、适用模型/地域、超额或超出保障额度后的处理、是否支持变配/续费/退费等）请参见 [推理模式介绍](#)。

相关文档

- [推理模式介绍](#)
- [TPM 预留](#)
- [计费方式](#)