大模型服务平台 TokenHub

模型推理

腾讯云

【版权声明】

【商标声明】

【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

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

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

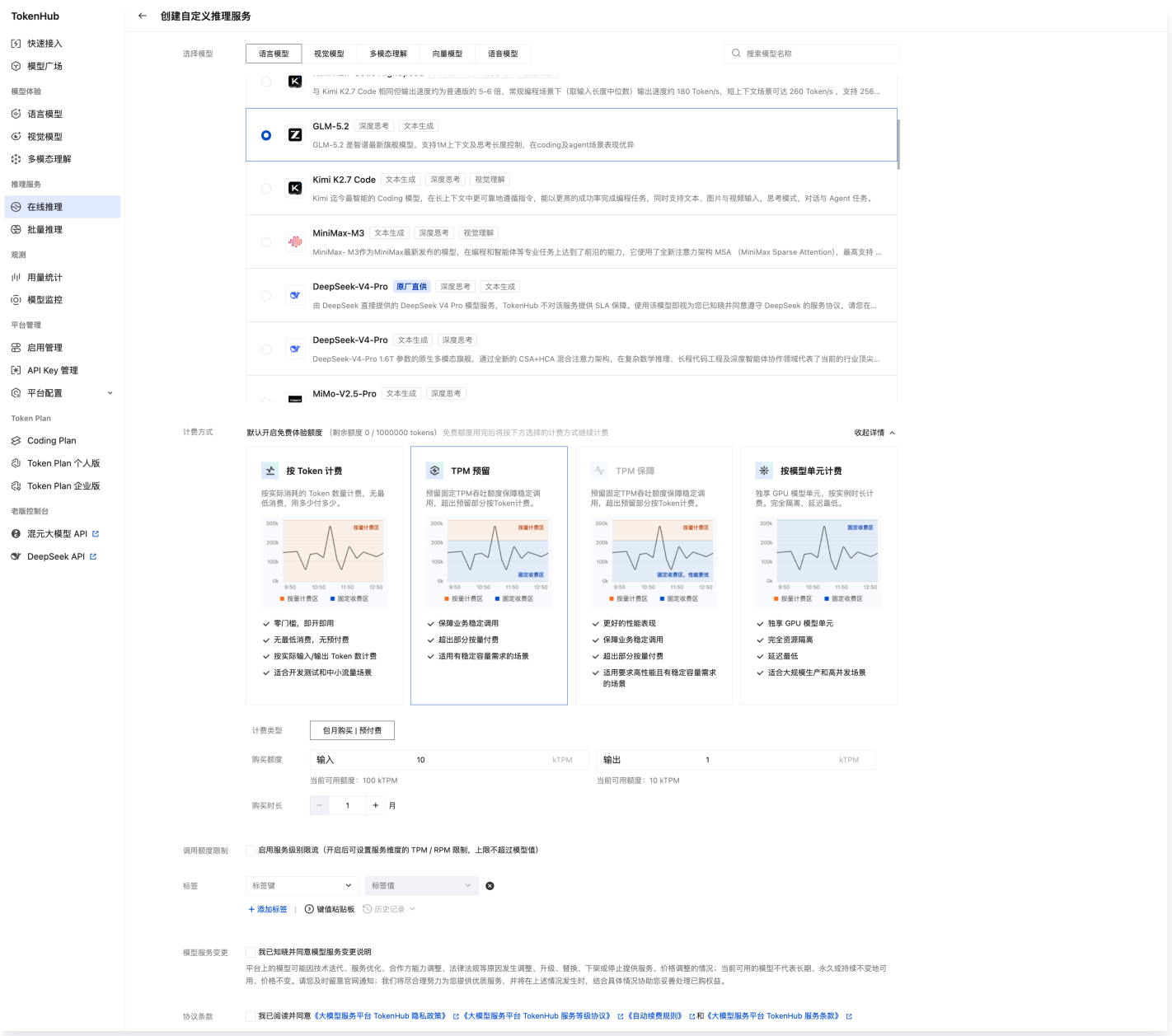
TPM 预留采用预付费模式，按月购买，价格请参见 TPM 预留价格 。超过预留 TPM 额度后，超出部分按标准按量付费价格计费，价格请参见 模型价格 。

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）



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

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。

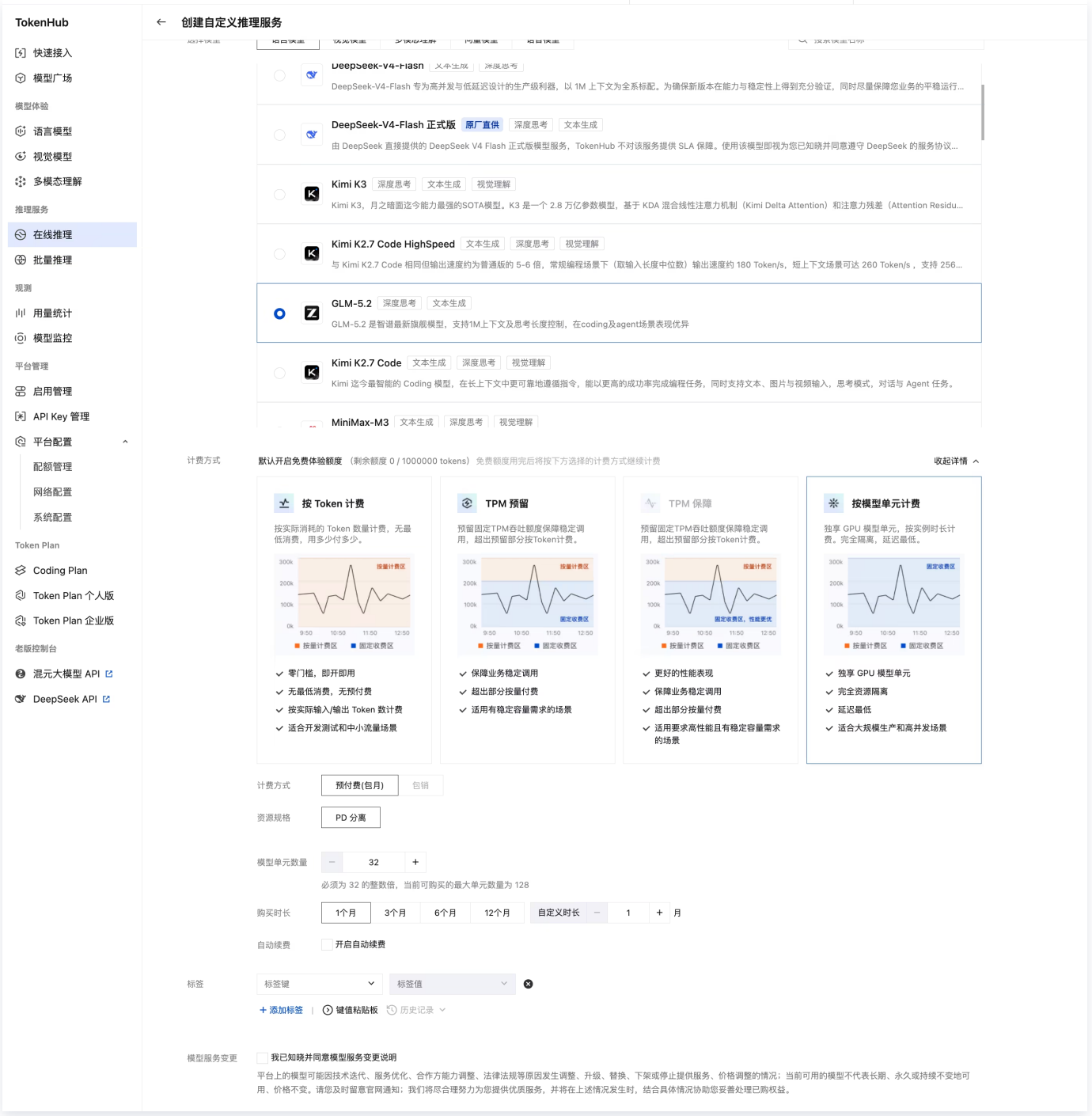
计费说明

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。



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