大模型知识引擎

购买指南

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2024 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】

腾讯云

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

版权所有：腾讯云计算（北京）有限责任公司

购买指南
计费概述

最近更新时间：2024-05-27 18:28:51

大模型知识引擎提供大模型应用构建平台，根据用户所使用的 token 数、知识库容量、并发数等资源进行计费，目前为公测阶段，注册开通体验即可获得一定量的免费额度。如需购买或扩容，请联系架构师或 官网客服 。

开通方式

大模型知识引擎的开通使用需要先通过 腾讯云企业实名认证 或者 腾讯云个人实名认证 。通过实名认证后，首次在 大模型知识引擎产品页 单击**产品体验**时，即可开通大模型知识引擎使用权限，有效期2个月。

免费额度

通过实名认证的主账号首次登录大模型知识引擎产品时，获得一定量的免费体验额度，详情如下图所示。

资源类别	免费额度
精调知识大模型标准版	开通大模型知识引擎服务即获赠累计50万 token 的免费调用额度，有效期2个月；以资源包的形式发放到您的腾讯云账号中，优先扣除。
精调知识大模型高级版	
混元大模型标准版	
混元大模型高级版	
医学大模型标准版	
金融大模型标准版	
知识库容量	开通大模型知识引擎服务即获赠累计300万字符数的免费知识库容量，主账户无 token 消耗记录半年后回收。
原子能力-多轮改写	开通大模型知识引擎服务即获赠累计50万 token 多轮改写免费额度，用完即止。
原子能力-Embedding	开通大模型知识引擎服务即获赠累计50万 token Embedding 免费额度，用完即止。

说明：
在以下场景/功能中发生交互时，会对 tokens 产生消耗：

- 应用配置管理：
  - 应用配置-角色设定中，一键优化功能

- 知识库管理-问答-导入问答-文档生成问答对
- 知识库管理-任务流程-插入节点-自动生成询问语
- 知识库管理-任务流程-插入节点-自动生成答案的预览

- 问答过程中，在应用配置中测试、应用发布后调用应用 API /使用体验链接、应用评测、应用体验：
  计算消耗内容：包括用户的输入+系统 prompt +召回信息（含文档、问答、开启搜索引擎后搜索召回的内容）+输出

备注：

1. token 换算方式：1 token ≈ 1~1.5个汉字。
2. 基于已配置的任务流程会调用任务型专属模型，消耗对应的 token。
3. 应用对话交互中触发的应用配置的欢迎语、输出配置-回复设置中自定义的保守回复、敏感词拦截后的回复不计入 token 消耗。
4. 如不使用任务流程和搜索引擎，建议关闭"应用配置-知识来源"中的相关开关，以减少额外的 token 消耗。
5. 应用配置-知识来源中，文档和问答的召回数量会影响召回 token ，设定的数量越高，拼接输入到大模型的召回片段越多，消耗 token 数量相应增加。

⚠ 注意：

- 在账单结算时，系统将按照免费资源包 > 预付费资源包的顺序进行结算，即免费资源包是优先扣除的。
- 若您欠费或因违禁原因停服后，将不能继续享受免费额度，只有服务重新开启后才可继续享受免费额度。
- 精调知识大模型标准版、精调知识大模型高级版、混元大模型标准版、混元大模型高级版、金融行业大模型标准版、医学行业大模型标准版共用50万 token 免费额度。

产品价格

预付费 tokens 资源包

定义：一次性购买一定数量的 tokens 资源包，有效期内调用模型服务时优先抵扣资源包当中的 tokens 余量，tokens 资源包如果到期未用完，会当做过期作废处理。
资源包有效期：1年，1年后未使用的资源包清零。
付费方式：预付费，购买 tokens 资源包越大，单价越低。
适用范围：稳定调用，具有一定规模的业务体量。

token资源包	精调知识大模型标准版	精调知识大模型高级版	混元大模型标准版	混元大模型高级版	医学大模型标准版	金融大模型标准版
1000万 token	￥120	￥1,200	￥120	￥1,200	￥120	￥180

5000万 token	￥600	￥6,000	￥600	￥6,000	￥600	￥900
1亿 token	￥1,180	￥11,800	￥1,180	￥11,800	￥1,180	￥1,770
5亿 token	￥5,900	￥59,000	￥5,900	￥59,000	￥5,900	￥8,850
10亿 token	￥11,700	￥117,000	￥11,700	￥117,000	￥11,700	￥17,550

注意：

- 购买的预付费 tokens 资源包有效期为1年，1年后未使用的资源包清零。
- 账户基础购买 tokens 预付费配额后，服务并发数保持不变，如果无法满足实际的业务需求，可按需增购并发。

知识库容量

定义：知识库容量计算用户所有应用上传的文档和问答的总字符数，删除的文档和应用不占用知识库容量。
资源包有效期：1年，1年后如需继续使用，请续费。已扩容的知识库不支持除删除之外的操作。
付费方式：包年。

知识库扩容包	价格
1000万字符	￥1,800
1亿字符	￥16,000

注意：

- 字符数不计算文档中的空格，文档中存在图片的将转存为图片链接，每张图大约换算为150个字符。
- 一篇3万字的硕士论文，大约相当于4万字符，50页文档。1000万字符约为250篇硕士论文，1亿字符约为2500篇硕士论文。

搜索服务包

知识引擎提供搜索服务，如您在知识来源中勾选了搜索引擎，则每次调用搜索引擎将会扣减搜索服务次数。搜索服务资源包根据业务量级划分为不同规格，可一次性付费购买，自购买日起一年内有效，一年内若资源包未使用完，则过期作废。

资源包规格	价格
10万次	￥2,200

50万次	￥11,000
100万次	￥20,000
500万次	￥90,000
1000万次	￥170,000

并发资源

定义：同时进行的会话数量，从请求发起到流式返回全部结果的整个过程都算在占用并发，一个对话占用并发时间约3-15秒不等。
模型：仅支持精调知识大模型标准版。
付费方式：包月包年。

并发类型	并发数量	包月	包年	备注
共享并发	1并发增购	￥800	￥9,200	账号默认5并发，可在此基础上增购
	2并发增购	￥1,250	￥15,000	
	5并发增购	￥3,000	￥35,000	
专属并发	1并发	17,000	￥200,000	5并发起购（增购可按照1、2、5并发增购）
	2并发	24,000	￥280,000	
	5并发	34,000	￥400,000	

原子能力

知识引擎支持以 API 形式提供原子能力接口，支持具有开发能力的用户自行搭建大模型应用，拓展大模型能力边界。
原子能力资源包根据业务量级划分为不同规格，可一次性付费购买，自购买日起一年内有效，一年内若资源包未使用完，则过期作废。

> ⚠ **注意：**
> 调用量的扣费顺序为：免费额度->资源包，即先消耗账号中的免费额度，免费额度耗尽后再消耗资源包。

版权所有：腾讯云计算（北京）有限责任公司

Embedding

Embedding 原子能力可通过 API 进行调用，点击 Embedding 查看API文档。资源包自购买日起一年内有效，一年内若资源包未使用完，则过期作废。

资源包大小	价格
5000万 token	￥100
1亿 token	￥200
5亿 token	￥900
10亿 token	￥1,800

多轮改写

多轮改写原子能力可通过 API 进行调用，点击多轮改写查看API文档。资源包自购买日起一年内有效，一年内若资源包未使用完，则过期作废。

资源包大小	价格
1000万 token	￥60
5000万 token	￥300
1亿 token	￥600
5亿 token	￥2,800
10亿 token	￥5,600

版权所有：腾讯云计算（北京）有限责任公司

购买方式

最近更新时间：2024-04-09 16:21:21

1. 在购买大模型知识引擎-知识库问答产品之前，您需要先 注册腾讯云账号 并通过实名认证，详情请参见 注册腾讯云 。
2. 通过 售前咨询 联系产品售前架构师，购买大模型知识引擎_知识库问答产品服务。
3. 选择并确认版本、购买时长、套数后，线下签订合同并下单，完成支付后即可开通账号进行体验。

版权所有：腾讯云计算（北京）有限责任公司

续费说明

最近更新时间：2024-04-01 11:32:11

若您购买的产品资源耗尽或到达产品有效期，请通过 售前咨询 联系产品售前架构师，进行续费。选择并确认版本、购买时长、套数后，线下签订合同并下单，完成支付后续费成功。

退费说明

最近更新时间：2024-04-01 11:32:11

暂不支持退款，请认真阅读合同条款后判断购买。