腾讯云数据库 AI 服务

ContextTurbo 应用指南



腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2026 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】

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

ContextTurbo 应用指南
- ContextTurbo 介绍
- 计费说明
- 快速入门
- 创建接入点
- 管理接入点
- 通过 API 或模型网关接入
- 通过 AI 工具接入
- 查看用量与节省效果
- 常见问题

版权所有：腾讯云计算（北京）有限责任公司

ContextTurbo 应用指南
ContextTurbo 介绍

最近更新时间：2026-10-28 17:46:00

产品简介

ContextTurbo 是腾讯云面向 AI 应用提供的上下文增强服务，帮助 AI 应用更高效地组织和使用上下文。
在 AI Coding、工具调用和多轮任务场景中，代码片段、工具日志、检索文档和对话历史持续累积，冗余内容不断增长。ContextTurbo 通过上下文压缩能力，在模型调用前由低成本模型先整理本次请求，精简冗余内容、保留关键语义，再由原有大模型继续推理与回答，帮助用户减少冗余输入、降低调用成本。

> **说明：**
> 上下文压缩当前限时免费，仅收取模型调用费用。适用范围及后续调整以正式活动和计费公告为准。

核心能力

上下文压缩

上下文压缩面向当前任务的模型输入。ContextTurbo 在请求发送给模型前，通过压缩模型精简上下文，减少模型需要读取的输入 Token。压缩以保留关键指令和任务信息为目标，根据内容特点精简工具日志、代码片段、检索文档和对话历史中的冗余部分。
用户可将 ContextTurbo 接入点配置到现有模型网关，也可由业务应用或 AI 工具直接调用。上下文压缩默认开启，可按接入点关闭或恢复，并通过用量统计查看 Token 与费用变化。

使用场景

AI Coding

开发者使用 Claude Code、Codex、OpenCode 等 AI 编程工具时，工具日志、代码搜索结果、测试输出和多轮对话历史持续累积。ContextTurbo 精简重复的构建日志、无关代码片段和冗长的工具输出，保留当前编码任务所需的错误信息、关键代码和任务指令，降低每轮调用的输入 Token 消耗。

多轮工作助手

在客服对话、文档问答、数据分析等多轮交互场景中，对话历史和检索结果逐轮叠加。ContextTurbo 精简早期对话中的重复背景和低相关段落，保留用户意图、关键约束和最新上下文，减少模型需要处理的输入量。

Agent 与工具调用

Agent 在执行复杂任务时频繁调用外部工具，每次工具返回的日志、查询结果和中间输出都会写入上下文。ContextTurbo 对重复的工具输出和过程信息进行精简，保留当前决策所需的关键结果和错误信息，帮助 Agent 在长链路任务中控制上下文规模和调用成本。

版权所有：腾讯云计算（北京）有限责任公司

RAG 检索增强生成

RAG 应用将检索到的多段文档拼入 Prompt。当检索结果冗长或包含重复段落时，ContextTurbo 精简低相关内容，保留回答所需的核心依据，减少输入 Token 同时维持回答质量。

核心优势

优势	说明
语义级压缩	基于低成本模型理解上下文语义，按内容特点选择保留、提炼或裁剪，而非简单截断。
无侵入接入	以 Proxy 形式部署在调用链路中，调用方只需替换请求地址，无需修改业务代码和 Prompt 结构。
灵活的模型连接	支持平台统一模型、腾讯云 TokenHub 和自定义模型服务，兼容 OpenAI 协议，一个接入点可调用多个模型。
按接入点管理	每个接入点独立配置压缩开关、API 来源和访问凭证，支持按业务线或环境隔离管理。
实时效果可观测	提供成本监控、Token 监控和使用明细，可按接入点和模型维度查看压缩前后的用量与费用变化。
压缩限时免费	上下文压缩当前不收取额外费用，用户仅需承担模型调用本身的费用。

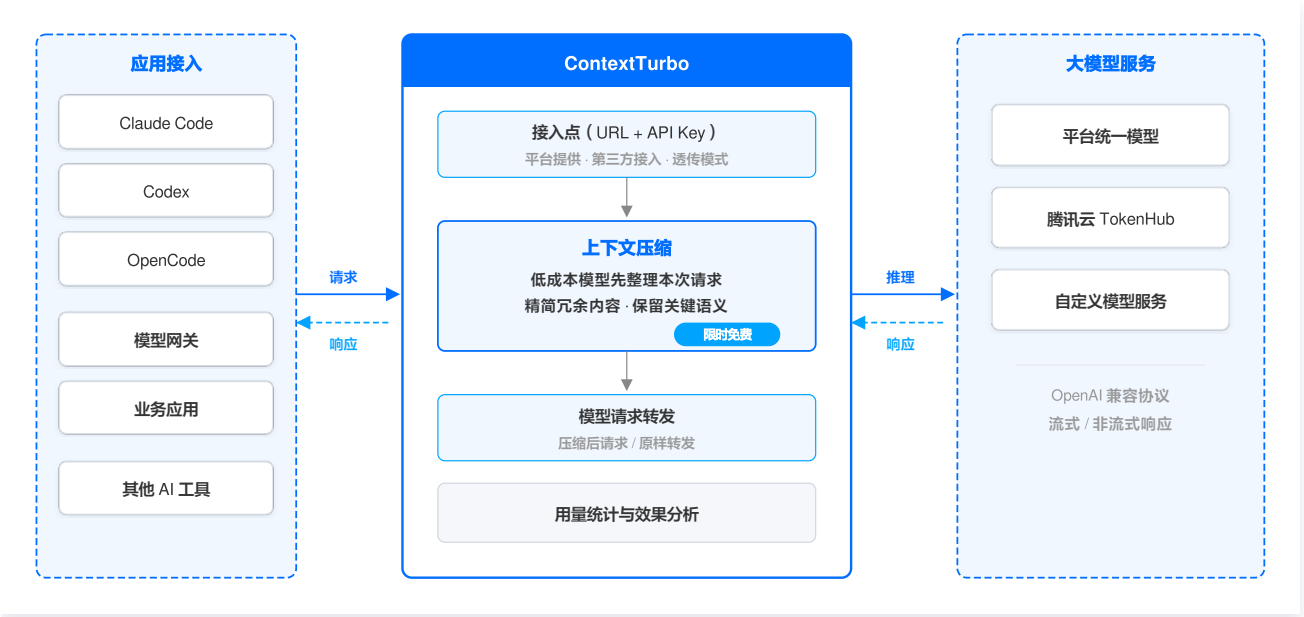
使用价值

价值维度	说明
降低调用成本	减少每次请求的输入 Token，直接降低大模型调用费用；多轮任务中历史上下文的累积精简可带来持续收益。
聚焦关键信息	精简重复日志和低相关片段，让模型输入更紧凑，有助于模型聚焦当前任务的核心内容。
支撑更长任务	控制上下文增长速度，使 AI 工具和 Agent 在长链路任务中更从容地使用上下文窗口。
连接现有工具链	兼容 Claude Code、Codex、OpenCode 等主流 AI 工具及模型网关，可直接融入已有工作流。
即开即用	创建接入点后默认开启压缩，无需逐项配置处理规则，无需调整业务代码逻辑。
效果可量化	通过用量统计对比压缩前后的 Token 用量、Credit 消耗和节省比例，支持数据驱动的成本决策。

产品架构

ContextTurbo 以上下文管理 Proxy 为核心，在应用与大模型服务之间提供上下文增强能力。架构由应用接入、上下文增强内核和大模型服务三部分组成，兼容现有 Agent 工作流与模型服务。

版权所有：腾讯云计算（北京）有限责任公司



架构组成说明：

- **应用接入**：支持 Claude Code、Codex、OpenCode 等 AI Coding 工具，以及模型网关和业务应用。调用方使用接入点 URL 和 API Key 发送请求。
- **ContextTurbo 核心**：接收请求后，根据压缩状态决定是否精简上下文，然后转发至模型服务。同时记录用量数据，支持用户查看压缩效果。
- **大模型服务**：支持平台统一模型、腾讯云 TokenHub 和自定义模型服务，兼容 OpenAI 兼容协议，支持流式和非流式响应。

上下文压缩工作原理

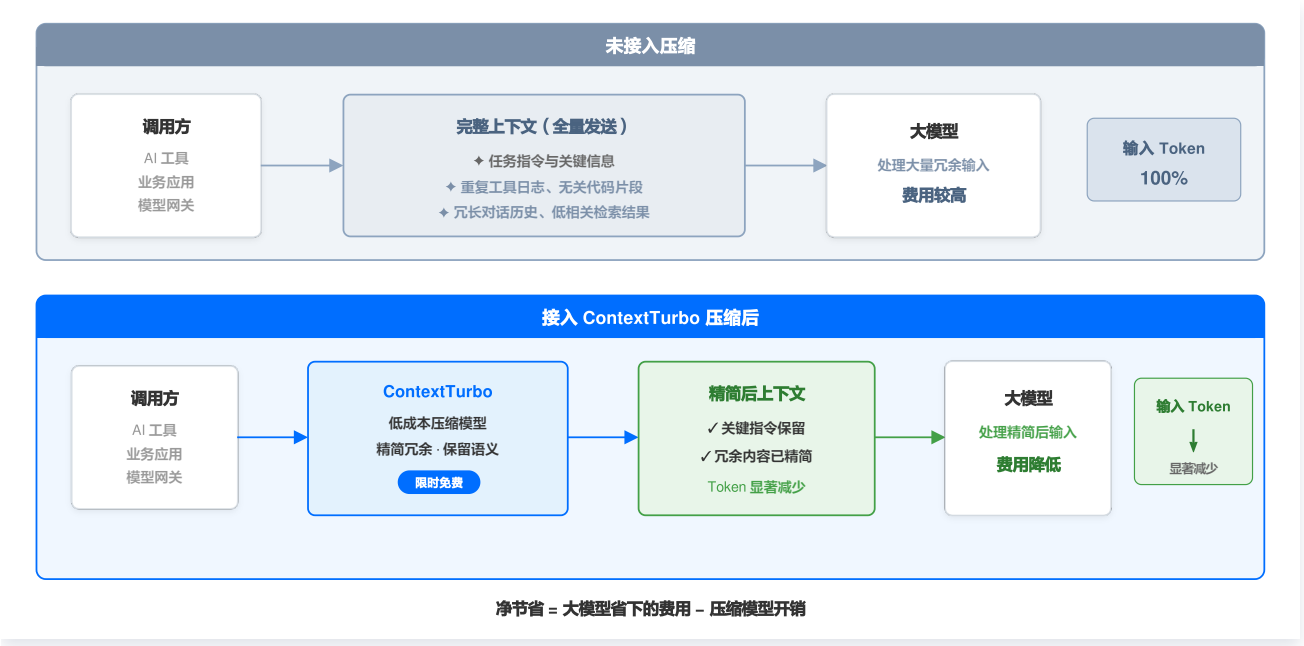
一次模型请求的处理流程如下：



1. 调用方将请求发送至 ContextTurbo 接入点。
2. 接入点验证接入点 API Key，根据 API 来源确定模型服务。
3. 开启压缩时，ContextTurbo 处理请求上下文，将精简后的请求发送给模型服务；关闭压缩时，按原请求转发。
4. 模型生成响应，ContextTurbo 将响应返回调用方。
5. 用户在用量统计中查看输入 Token 及费用变化。

上下文压缩降本机制

上下文压缩通过低成本压缩模型先整理请求，减少大模型需要读取的输入 Token，从而降低模型调用费用。



> **说明：**
> 上图为原理示意，不代表固定压缩比例。净收益以整个任务的实际费用为准。

功能管理

上下文压缩按接入点管理，创建接入点时默认开启。

压缩状态	请求处理方式
已开启	对请求上下文进行压缩处理后转发给模型
已关闭	按原请求转发给模型

关闭压缩后，调用方仍使用该接入点访问模型服务。若需停止接入点接受调用，应执行停用接入点操作。
具体操作请参见 管理接入点 。

使用限制

限制项	说明
支持地域	广州、北京、上海
支持协议	OpenAI 兼容协议（Chat Completions）
API 来源	平台提供、第三方接入（腾讯云 TokenHub / 自定义）、透传模式
网络地址	公网 URL 和内网 URL
压缩费用	当前限时免费

相关文档

- 快速入门
- 创建接入点
- 查看用量与节省效果
- 常见问题

版权所有：腾讯云计算（北京）有限责任公司

计费说明

最近更新时间：2026-10-28 17:46:00

费用组成

使用 ContextTurbo 的上下文压缩能力时，需要区分模型调用费用与上下文压缩费用。模型负责执行推理，上下文压缩负责精简模型输入，两者对应不同的服务消耗。

费用类别	说明
模型调用费用	按所用模型服务的计费规则计算，可能包含输入、缓存输入和输出等计费项。
上下文压缩费用	当前 ContextTurbo 不收费。后续如有调整，将以正式公告为准。

⚠ **注意：**
当前 ContextTurbo 不收费，不代表模型调用免费。使用第三方接入或透传模式时，仍需关注模型账户的额度与计费规则。

版权所有：腾讯云计算（北京）有限责任公司

快速入门

最近更新时间：2026-10-28 17:46:00

操作场景

本文介绍如何创建 ContextTurbo 接入点，将其配置到现有模型调用链路，并查看上下文压缩效果。

前提条件

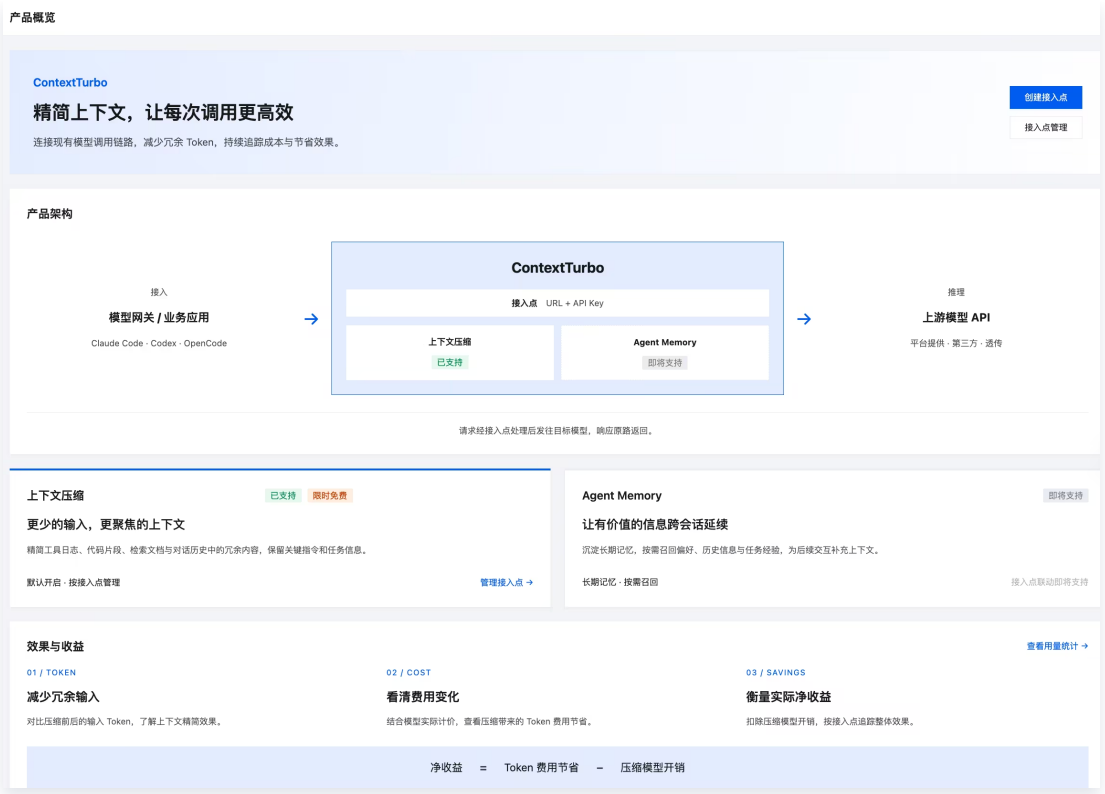
- 已注册腾讯云账号并完成实名认证。
- 已具备 ContextTurbo 控制台 访问权限。
- 已确定使用平台提供、第三方接入或透传模式。
- 使用第三方接入或透传模式时，已获取有效的模型 Base URL、模型 API Key 和 Model ID。
- 调用环境能够访问接入点；使用内网地址时，已具备相应网络连通条件。

步骤一：创建接入点

1. 登录 ContextTurbo 控制台 。
2. 在左侧导航栏选择**接入点管理**，切换至所需地域。
3. 单击**创建接入点**。
4. 填写接入点名称，并选择 API 来源。
5. 根据所选来源完成配置后，单击**创建接入点**。

API 来源	配置内容
平台提供	无需填写模型 Base URL 和模型 API Key。
第三方接入	选择接入平台，填写模型 Base URL 和模型 API Key，可执行连通性验证。
透传模式	创建时无需配置模型 Base URL 和模型 API Key，调用时通过 Header 传入。

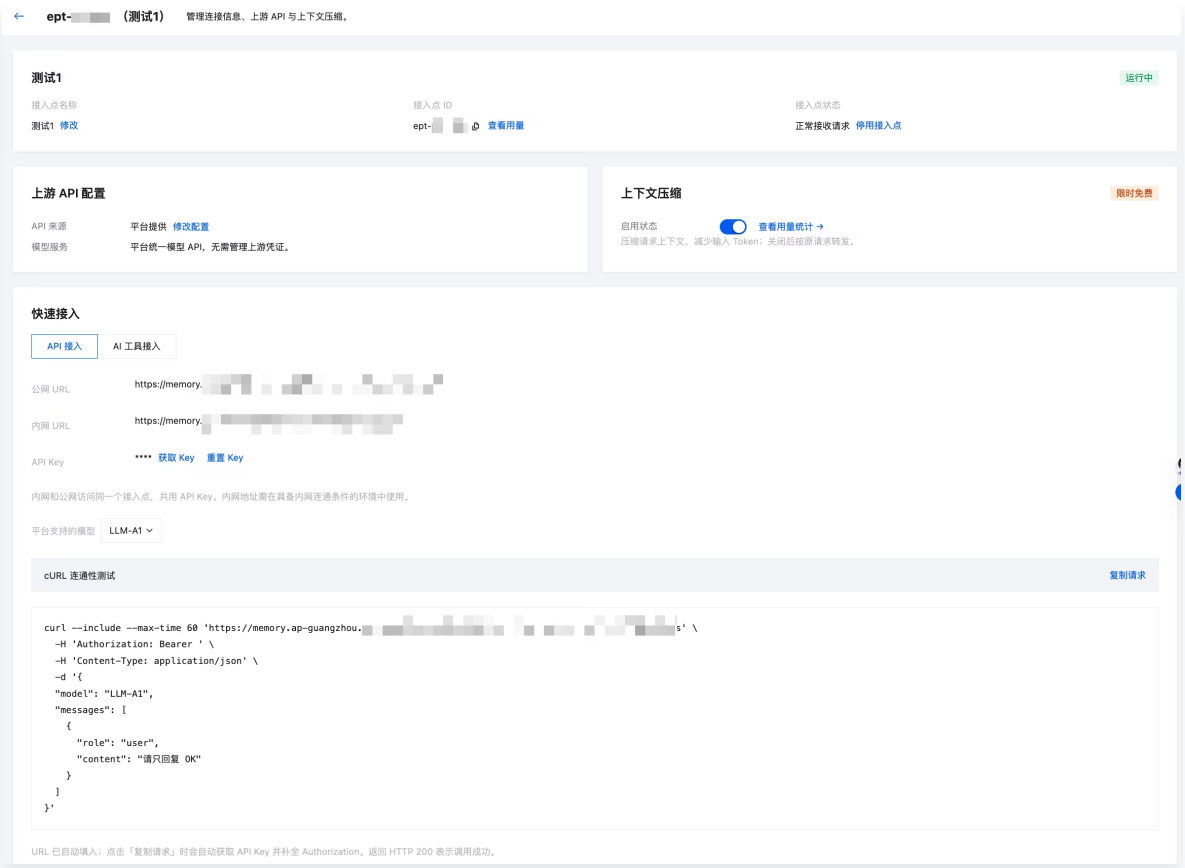
创建后，上下文压缩默认开启。详细操作请参见 创建接入点 。
下图为控制台产品概览页面，展示了接入点管理入口和创建接入点按钮。



步骤二：获取连接信息

1. 在接入点列表中找到目标接入点，单击**查看详情**。
2. 获取该接入点的访问 URL 和 API Key。
3. 根据调用环境选择公网或内网地址。
4. 确认模型服务支持的接口和 Model ID。平台提供模式下，可在快速接入中查看支持的模型。

   下图为接入点详情页面，展示了公网 URL、内网 URL、API Key 及快速接入指引。



步骤三：配置调用方

根据实际使用方式完成配置。

调用方式	配置方法
模型网关	将接入点作为网关中的模型服务连接，填写接入点 URL 和 API Key，并配置目标模型。
业务应用	按接入点快速接入指引，配置模型请求地址、接入点鉴权与 `model` 参数。
AI 工具	按控制台中对应工具的接入指引，配置服务地址、接入点凭证和模型。

透传模式还需由调用方在每次请求中添加携带模型 Base URL 和模型 API Key 的 Header，字段以接入点快速接入指引为准。

> **说明：**
> 接入点 API Key 用于访问 ContextTurbo，模型 API Key 用于访问模型服务，两者不能混用。

步骤四：验证模型调用

1. 使用接入点快速接入中的请求示例，发送一条简单测试请求。
2. 确认请求可以返回所选模型的响应。
3. 再发起一段具有代表性的多轮任务，覆盖实际使用的工具调用和上下文内容。

   以下为平台提供模式下的 cURL 请求示例：

```
curl --include --max-time 60 'https://memory.ap-
guangzhou.tencentdai.com/contextturbo/{接入点ID}/chat/completions' \
  -H 'Authorization: Bearer {接入点API Key}' \
  -H 'Content-Type: application/json' \
  -d '{
  "model": "{Model ID}",
  "messages": [
    {
      "role": "user",
      "content": "请只回复 OK"
    }
  ]
}'
```

> **说明：**
> 将 `{接入点 ID}`、`{接入点 API Key}`、`{Model ID}` 替换为实际值。URL 已自动填入，单击控制台快速接入中的**复制请求**时会自动获取 API Key 并补全 Authorization。返回 HTTP 200 表示调用成功。

简单测试用于确认调用链路；压缩收益应使用实际业务任务验证。

步骤五：查看效果

1. 在接入点列表中，单击目标接入点的**查看用量**；也可从左侧导航选择**上下文压缩 > 用量统计**。
2. 选择对应接入点和统计时间范围。
3. 查看输入 Token、费用及净收益等数据，结合任务完成情况评估效果。

   详细说明请参见 查看用量与节省效果 。

相关文档

- ContextTurbo 介绍
- 创建接入点
- 通过 API 或模型网关接入
- 通过 AI 工具接入

创建接入点

最近更新时间：2026-10-28 17:46:00

操作场景

本文介绍如何在 ContextTurbo 控制台创建接入点，配置模型 API 来源并开启上下文压缩。接入点是调用方与 ContextTurbo 的连接入口，每个接入点拥有独立的访问地址和 API Key。

前提条件

- 已具备 ContextTurbo 控制台 访问权限。
- 使用第三方接入或透传模式时，已准备好模型 Base URL 和模型 API Key。

三种 API 来源对比

对比项	平台提供	第三方接入	透传模式
模型 Base URL	平台预设	创建时填写	每次请求通过 Header 传入
模型 API Key	无需管理	创建时填写	每次请求通过 Header 传入
模型选择	平台支持的 Model ID	已配置模型服务支持的 Model ID	请求指定模型服务支持的 Model ID
连通性验证	不需要	支持	不适用
适用场景	快速体验、无需管理模型凭证	使用已有模型服务、固定模型连接	模型网关按请求管理模型服务

操作步骤

步骤一：进入创建页面

1. 登录 ContextTurbo 控制台 。
2. 在左侧导航中选择**接入点管理**。
3. 选择目标地域。
4. 单击**创建接入点**。



步骤二：配置 API 来源

填写便于识别业务用途的接入点名称，例如 `coding-production`，然后选择 API 来源。

平台提供



使用平台统一模型 API，无需选择供应商或填写模型 API Key。创建后，在接入点快速接入中查看平台支持的模型与 Model ID，调用时通过 `model` 参数指定。

第三方接入



用于接入已开通的模型服务，需填写以下信息。

配置项	说明
接入平台	按实际使用的服务选择。当前提供腾讯云 TokenHub 与自定义接入选项。
模型 Base URL	已开通的模型服务 API 基础地址，可按实际地域或套餐调整。
模型 API Key	访问模型服务的有效凭证。
连通性验证	单击**验证连通性**，检查模型 API 是否可访问及凭证是否有效；验证不会修改当前配置。

> **说明：**
> 模型 Base URL 应按对应模型服务的接口规范填写。不要将完整业务请求路径、控制台页面地址或其他接入点地址误填为模型服务基础地址。

透传模式



无需预先配置模型 Base URL 或模型 API Key。调用方在每次请求中通过 Header 传入模型 Base URL 和模型 API Key，适用于由模型网关按请求管理模型服务的场景。
创建后，请在接入点快速接入中查看透传 Header 配置方式。

步骤三：确认压缩状态

上下文压缩默认开启，创建后可在接入点详情中关闭或恢复。

步骤四：完成创建

1. 核对接入点名称和模型连接配置。
2. 单击**创建接入点**。
3. 创建成功后，在接入点列表中查看资源，进入详情获取访问 URL 和 API Key。

后续操作

- 通过 API 或模型网关接入
- 通过 AI 工具接入
- 管理接入点

管理接入点

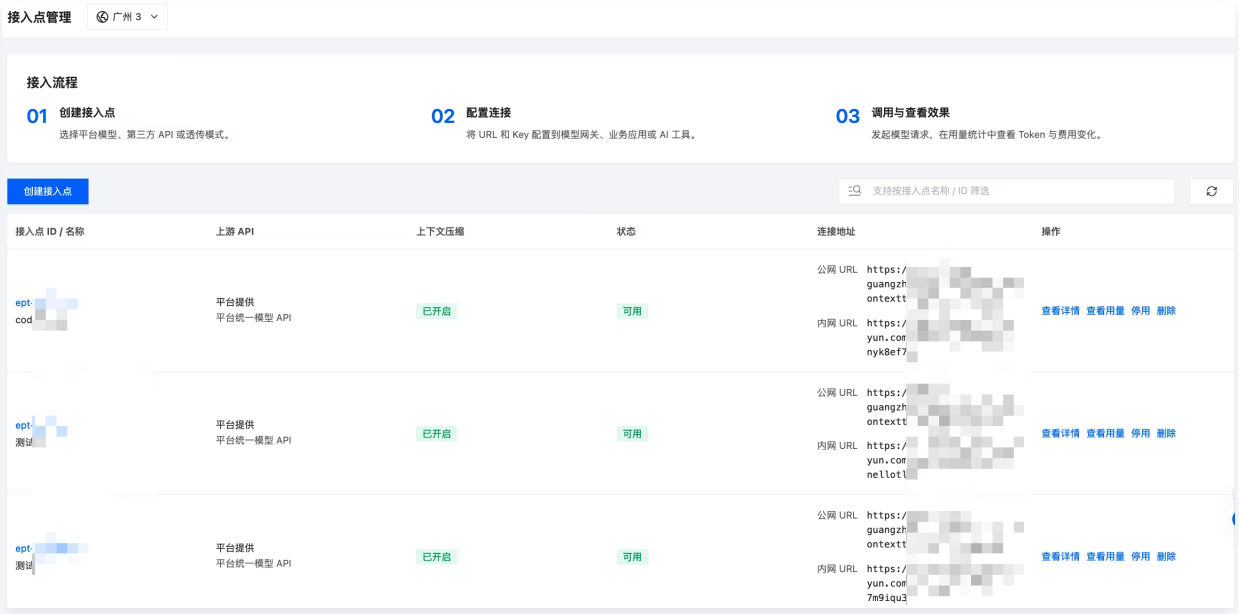
最近更新时间：2026-10-28 17:46:00

操作场景

本文介绍如何查看接入点信息、管理压缩状态，以及停用或删除接入点。

查看接入点

1. 登录 ContextTurbo 控制台 。
2. 在左侧导航栏中选择**接入点管理**。
3. 选择接入点所在地域。
4. 按接入点名称或 ID 筛选目标资源。



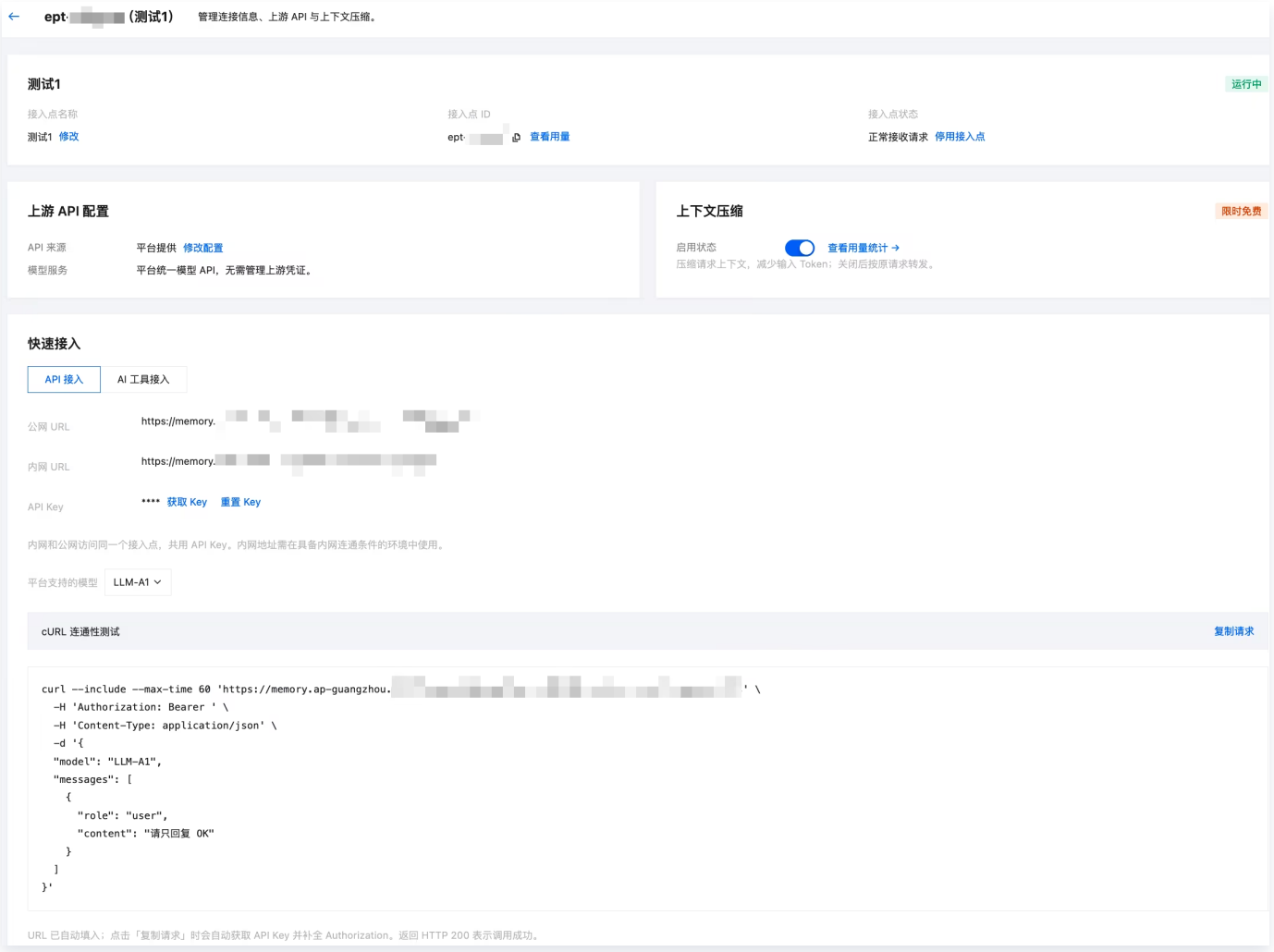
列表展示以下信息。

字段	说明
接入点 ID / 名称	用于识别接入点，资源 ID 以 `ept-` 开头。
上游 API	当前接入点的 API 来源（平台提供 / 第三方接入 / 透传模式）。
上下文压缩	当前是否开启压缩。
状态	接入点当前是否可用。
连接地址	公网 URL 和内网 URL。

版权所有：腾讯云计算（北京）有限责任公司

操作	查看详情、查看用量、停用或删除等操作。

5. 在**操作**列单击**查看详情**可进入接入点详情，查看连接信息与接入指引。



获取连接地址

根据调用环境选择公网 URL 或内网 URL。

地址类型	使用说明
公网 URL	在能够访问公网服务的调用环境中使用。
内网 URL	在具备对应内网连通条件的环境中使用；接入前需验证域名解析与网络可达性。

请直接使用当前接入点展示的地址。接口路径和鉴权方式应与快速接入中的示例保持一致。

关闭或恢复上下文压缩

1. 进入目标接入点详情。
2. 在上下文压缩区域调整启用状态，按页面提示完成操作。
3. 确认新的压缩状态，并使用业务请求验证。

   关闭压缩后，接入点按原请求转发至模型服务。恢复压缩后，后续请求按压缩流程处理。



> **说明：**
> 关闭上下文压缩与停用接入点是两种操作。关闭压缩后仍可通过接入点调用模型；停用会影响通过该接入点发起的后续调用。

修改接入点配置

在接入点详情页面，单击接入点名称旁的**修改**可更改接入点名称。单击上游 API 配置区域的**修改配置**可调整模型连接信息。



停用接入点

不再需要某个接入点接受调用时，可在接入点列表中单击**停用**，按页面提示确认操作。停用前应检查业务调用方，并完成流量切换，避免影响正在使用该接入点的业务。



删除接入点

1. 在接入点列表中找到目标资源。
2. 确认业务应用、模型网关和 AI 工具已不再使用该接入点。
3. 单击**删除**，阅读页面提示后完成操作。

> **警告：**
> 删除后，调用方不能继续依赖原接入点配置访问服务。删除操作不可恢复，请在操作前确认所有调用方已完成迁移。

查看接入点用量

在接入点列表中单击**查看用量**，进入对应的用量统计页面。具体指标请参见 查看用量与节省效果 。

相关文档

- 创建接入点
- 查看用量与节省效果
- 常见问题

通过 API 或模型网关接入

最近更新时间：2026-10-28 17:46:00

操作场景

ContextTurbo 可接入业务应用，也可作为现有模型网关中的一个模型服务连接。本文介绍连接信息、鉴权凭证和三种 API 来源的配置方法。

前提条件

- 已创建可用接入点。
- 已获取接入点访问地址和 API Key。
- 已确认目标模型的 Model ID，以及调用方和模型服务共同支持的接口协议。
- 调用环境可访问所选接入点地址。

连接信息

信息	用途	获取方式
接入点 URL	将请求发送到指定 ContextTurbo 接入点	在接入点列表或详情中获取。
接入点 API Key	调用方访问接入点的鉴权凭证	在接入点详情中获取。
Model ID	指定本次请求使用的模型	平台提供模式查看快速接入；其他模式查看模型服务的模型列表。
模型 Base URL	ContextTurbo 转发请求的目标模型服务地址	由模型服务提供。
模型 API Key	ContextTurbo 调用模型服务的鉴权凭证	由模型服务提供。

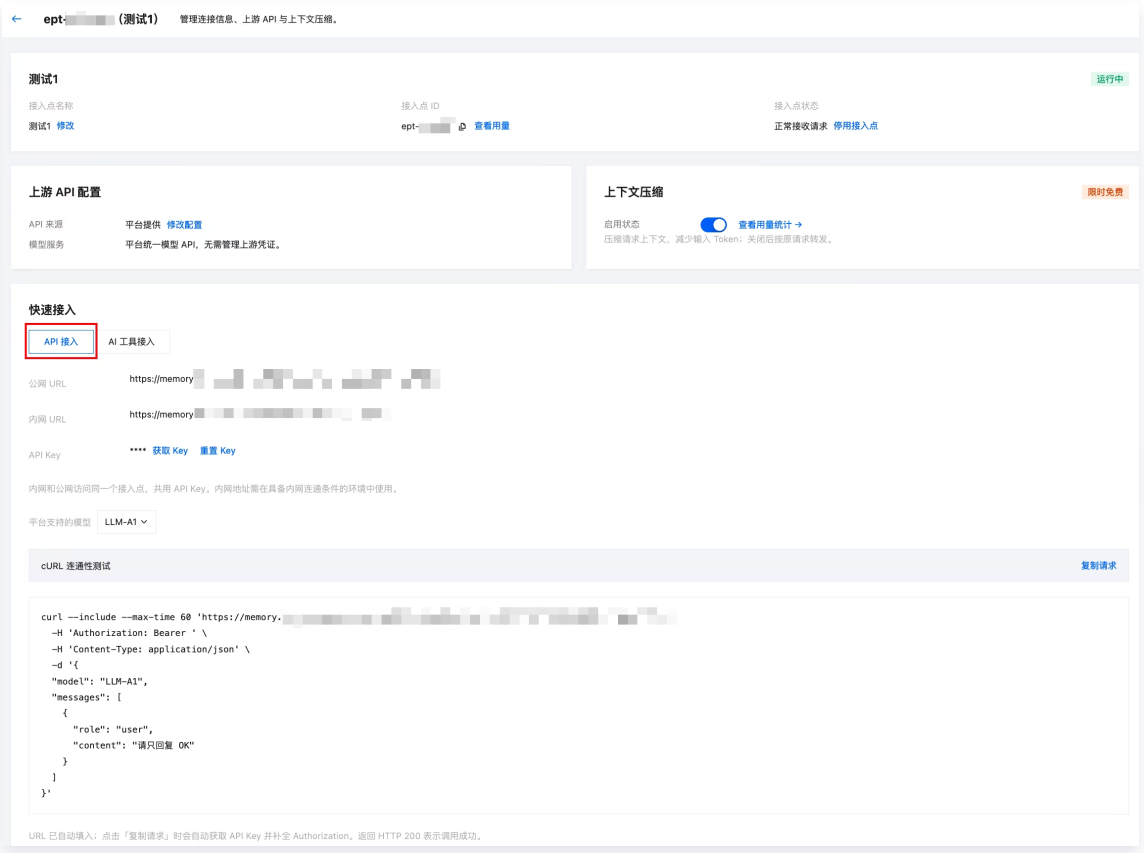
> **说明：**
> 接入点 URL 是接入地址。完整请求路径、协议版本和鉴权 Header 请使用该接入点快速接入中提供的示例，避免重复拼接接口路径。

业务应用接入

步骤一：获取请求示例

1. 登录 ContextTurbo 控制台 。

2. 在左侧导航栏中选择**接入点管理**。
3. 选择目标地域。
4. 单击目标接入点的**查看详情**。
5. 在快速接入区域选择 **API 接入**。
6. 选择适用的网络地址、模型和接口，获取请求示例。



步骤二：配置请求

按所选 API 来源配置以下内容。

API 来源	请求配置
平台提供	使用接入点 URL 和 API Key，在 `model` 中填写平台支持的模型标识。
第三方接入	使用接入点 URL 和 API Key，在 `model` 中填写已配置模型服务支持的模型标识。
透传模式	使用接入点 URL 和 API Key，同时在 Header 中传入本次调用的模型 Base URL 和模型 API Key。

ContextTurbo 将请求中的 `model` 原样传给模型服务。一个接入点可以通过不同 `model` 调用模型服务允许访问的模型，无需按每个模型分别创建接入点。

步骤三：验证调用

先按控制台提供的完整示例发送简单请求，确认返回正常，再将连接信息应用到业务代码。业务调用仍需遵循目标模型的参数、消息格式与工具调用约定。

Python SDK 示例：

```
from openai import OpenAI

client = OpenAI(
    api_key="{接入点API Key}",
    base_url="https://memory.ap-guangzhou.tencentdai.com/contextturbo/{接入点ID}"
)

response = client.chat.completions.create(
    model="{Model ID}",
    messages=[
        {"role": "user", "content": "请只回复 OK"}
    ]
)

print(response.choices[0].message.content)
```

> **说明：**
> 将 `{接入点 API Key}`、`{接入点 ID}`、`{Model ID}` 替换为实际值。ContextTurbo 兼容 OpenAI SDK，可直接使用 `openai` 库进行调用。

模型网关接入

步骤一：确定 API 来源

- 使用平台提供或第三方接入模式时，接入点使用预设模型服务，适合在模型网关中建立对应的模型服务连接。
- 使用透传模式时，模型服务由每次请求指定，适合由网关继续管理模型 Base URL 和模型 API Key。

步骤二：配置网关连接

1. 在模型网关中创建或调整模型服务连接。
2. 将调用地址配置为 ContextTurbo 接入点 URL。
3. 将访问凭证配置为接入点 API Key。
4. 保留目标模型的 Model ID，并确认网关发送的协议与模型服务兼容。

5. 透传模式下，按接入点指引配置模型连接 Header，并确认网关不会将其过滤或覆盖。

模型选择、用户权限和业务路由继续由调用方网关配置决定。

步骤三：验证完整链路

```
业务应用 → 模型网关 → ContextTurbo 接入点 → 模型 API
业务应用 ← 模型网关 ← ContextTurbo 接入点 ← 模型响应
```

通过网关发送测试请求，确认接入点鉴权、模型服务访问与响应返回均正常后，再切换业务流量。

凭证使用说明

凭证	使用方向	配置位置
接入点 API Key	调用方 → ContextTurbo	应用、模型网关或 AI 工具。
模型 API Key	ContextTurbo → 模型 API	第三方模式保存在模型连接配置中；透传模式由每次请求提供。

> **说明：**
> 透传模式需要同时满足这两层鉴权。仅提供模型 API Key，不能替代接入点访问凭证。

接入后检查

- 核对调用方使用的接入点 URL 与 API Key 是否对应同一资源。
- 检查 `model` 是否在模型服务可用范围内。
- 验证实际业务中的流式响应、工具调用和错误返回。
- 在用量统计中确认调用归属于预期接入点。

常见异常请参见 常见问题 。

通过 AI 工具接入

最近更新时间：2026-10-28 17:46:00

操作场景

ContextTurbo 可连接 Claude Code、Codex、OpenCode 等 AI 工具。工具通过接入点调用模型，ContextTurbo 在调用链路中处理请求上下文。

前提条件

- 已安装并能够正常使用目标 AI 工具。
- 已创建 ContextTurbo 接入点，并获取 URL 和 API Key。
- 已确认模型具备目标工具所需的接口及模型能力。

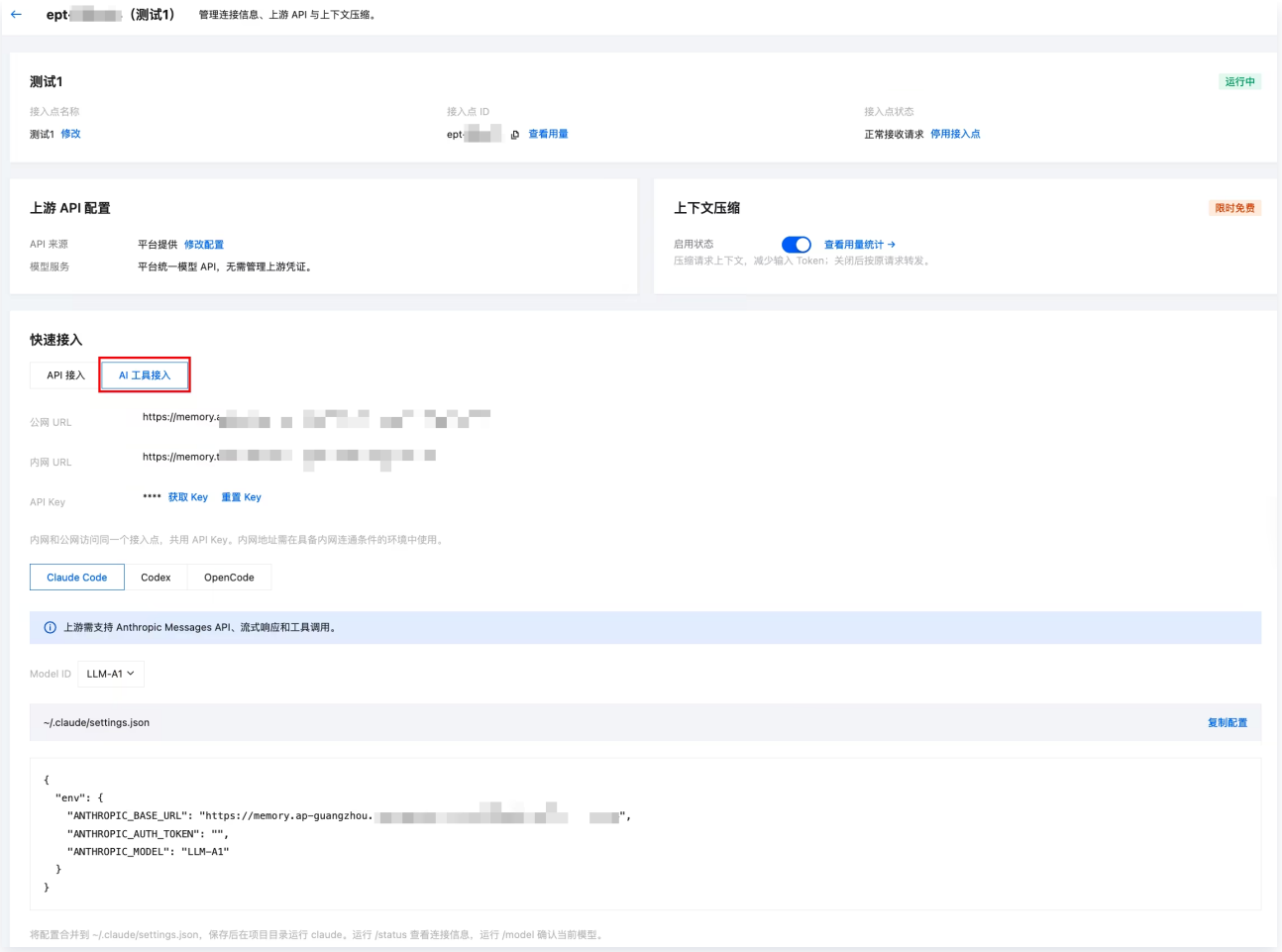
> **说明：**
> 工具可配置自定义模型地址，不代表任意模型都可直接使用。接入前需确认接口协议、流式响应、工具调用以及相关模型能力是否满足工具要求。

操作步骤

步骤一：选择工具接入指引

1. 登录 ContextTurbo 控制台 。
2. 在左侧导航栏中选择**接入点管理**。
3. 选择目标地域。
4. 单击目标接入点的**查看详情**。
5. 在快速接入区域选择 **AI 工具接入**页签。
6. 选择 Claude Code、Codex 或 OpenCode 等对应工具。

版权所有：腾讯云计算（北京）有限责任公司



步骤二：配置连接

按对应工具指引中的配置位置和格式，填写以下信息。

配置内容	说明
服务地址	使用目标接入点的公网或内网 URL。
访问凭证	使用接入点 API Key。
模型	使用模型服务支持的 Model ID。
模型连接 Header	仅透传模式需要，按控制台说明配置模型 Base URL 和模型 API Key。

各工具的配置文件、环境变量及鉴权字段可能不同，请直接采用控制台对应工具提供的配置内容，并按所安装工具版本核对。

步骤三：验证接入

1. 按工具指引重新加载配置或启动工具。
2. 发起简单对话，确认模型可以正常响应。

3. 发起包含文件读取、代码搜索或测试执行的任务，验证工具调用和连续对话。
4. 在 ContextTurbo 用量统计中选择对应接入点，确认产生调用数据。

步骤四：验证压缩效果

选取具有代表性的代码任务，观察压缩前后输入 Token、费用和任务完成情况。需要对照验证时，可在接入点详情中关闭压缩，再使用相同任务和模型配置测试。

接入建议

- 透传模式依赖工具或中间网关传递自定义 Header；若当前工具不能配置必要 Header，可使用第三方接入模式，将模型连接配置保存在接入点中。
- 初次验证时可使用独立接入点，便于区分不同工具和配置产生的用量。
- 工具已自带上下文整理能力时，仍应以完整任务的成本与完成情况评估组合效果。

相关文档

- 创建接入点
- 通过 API 或模型网关接入
- 查看用量与节省效果

查看用量与节省效果

最近更新时间：2026-10-28 17:46:00

操作场景

本文介绍如何通过用量统计查看上下文压缩对输入 Token 和费用的影响，以及如何评估压缩的净收益。

前提条件

已创建接入点并产生调用数据。

操作步骤

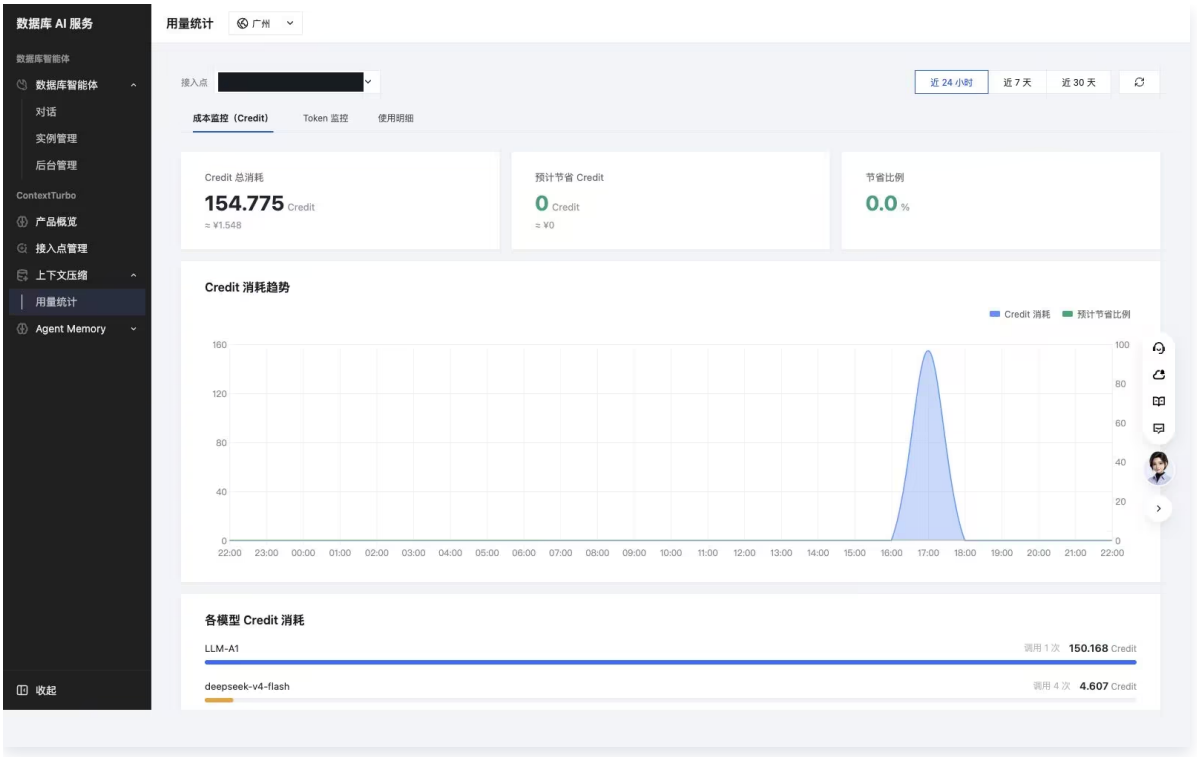
1. 登录 ContextTurbo 控制台 。
2. 在左侧导航栏选择**上下文压缩 > 用量统计**；也可在接入点列表中单击目标接入点的**查看用量**。
3. 选择目标接入点和统计时间范围（近24小时、近7天或近30天）。

成本监控

在**成本监控（Credit）**页签中，查看以下指标。

指标	说明
Credit 总消耗	统计范围内的 Credit 总消耗量，并换算为对应金额。
预计节省 Credit	压缩后预计节省的 Credit 数量。
节省比例	预计节省 Credit 占原始消耗的比例。

页面同时展示 Credit 消耗趋势图和各模型 Credit 消耗明细。

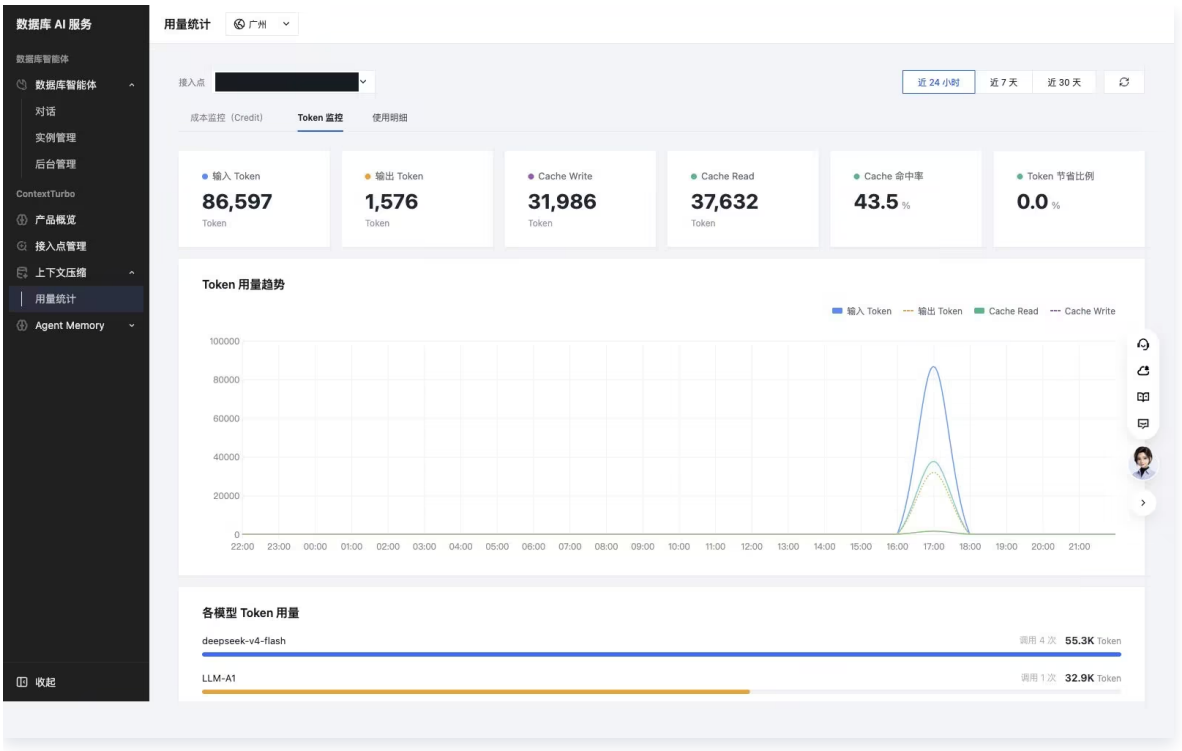


Token 监控

在 Token 监控页签中，查看以下指标。

指标	说明
输入 Token	统计范围内的总输入 Token 数量。
输出 Token	统计范围内的总输出 Token 数量。
Cache Write	写入缓存的 Token 数量。
Cache Read	从缓存读取的 Token 数量。
Cache 命中率	Cache Read 占总输入 Token 的比例。
Token 节省比例	压缩后输入 Token 的节省比例。

页面同时展示 Token 用量趋势图和各模型 Token 用量明细。



使用明细

切换至使用明细页签，查看各次调用的详细信息。

字段	说明
时间	调用发生时间。
模型	本次调用使用的模型。
输入	请求输入内容摘要。
输入 Token	本次调用的输入 Token 数量。
输出 Token	本次调用的输出 Token 数量。
Cache Write	本次写入缓存的 Token。
Cache Read	本次从缓存读取的 Token。
Credit	本次调用消耗的 Credit。
节省比例	本次调用的 Token 节省比例。

压缩效果指标的理解

指标	含义	阅读方式

压缩前输入 Token	请求在压缩处理前的输入规模	用于了解原始上下文大小。
压缩后输入 Token	压缩处理后发送给模型服务的输入规模	与压缩前用量比较，观察精简效果。
输入 Token 减少量	压缩前后输入 Token 的差额	反映输入规模变化，不直接等同于费用收益。
Token 费用节省	按模型计价比较压缩前后用量对应的费用差额	需结合缓存和其他计费项理解。
压缩模型开销	上下文处理产生的压缩模型开销	应与模型费用节省一起评估。
净收益	Token 费用节省减去压缩模型开销	用于判断压缩是否带来净节省。

计算关系

```
输入 Token 减少量 = 压缩前输入 Token - 压缩后输入 Token

输入 Token 减少比例 = 输入 Token 减少量 / 压缩前输入 Token

净收益 = Token 费用节省 - 压缩模型开销
```

> **说明：**
> 模型费用受输入、缓存输入和输出等计费规则影响，不能直接用输入 Token 减少比例替代费用节省比例。

如何判断效果

1. **查看输入是否减少**

   比较相同统计范围内压缩前后的输入 Token。输入较短或冗余较少时，精简空间可能有限；不同任务类型应分别观察。
2. **查看费用是否下降**

   在输入减少的基础上，继续核对模型计价和缓存变化。缓存输入与普通输入可能采用不同单价，相同的 Token 减少量可能对应不同的费用变化。
3. **查看是否形成净收益**

   将压缩模型开销纳入评估。压缩自身开销高于模型费用节省时，净收益可能为负，应结合请求内容与完整任务情况分析。
4. **查看完整任务结果**

对照同类任务的完成情况、重试次数、模型调用次数与总耗时。若压缩后的单次请求节省被后续额外调用抵消，需按完整任务重新评估收益。

数据理解

- **保持统计范围一致**：比较数据时使用相同时间范围、接入点和任务条件。
- **区分估算与账单**：未压缩费用属于比较基线，对应模型未实际处理原始请求时，该费用不能作为已发生账单。
- **避免重复累计**：各轮调用中历史上下文减少带来的节省，应计入对应请求，不应再额外累计一笔历史复用收益。
- **关注计量完整性**：缺少价格、缓存或用量信息时，不应将缺失值按零费用理解。费用结算以相应服务账单为准。

相关文档

- ContextTurbo 介绍
- 常见问题

常见问题

最近更新时间：2026-10-28 17:46:00

产品与能力

ContextTurbo 提供哪些上下文增强能力?

目前主要提供上下文压缩和 Agent Memory 两块能力。上下文压缩精简当前任务中的冗余输入；Agent Memory 沉淀和召回用户偏好、历史事实、任务经验与团队知识，为后续交互补充上下文。当前两块能力分别通过接入点和 Memory 实例使用，Memory 与接入点的联动即将支持。

ContextTurbo 适合哪些业务?

ContextTurbo 适合需要精简模型输入、延续跨会话信息或复用任务经验的 AI 应用，例如 AI Coding、多轮工作助手和团队协作 Agent。可根据业务需求使用上下文压缩和 Agent Memory，具体效果应使用实际业务任务验证。

一个接入点只能调用一个模型吗?

一个接入点可以调用所连接模型服务支持的多个模型，具体模型由请求中的 `model` 参数指定。平台提供模式使用平台支持的 Model ID；第三方与透传模式使用对应模型服务支持的 Model ID。

ContextTurbo 会自动更换请求中的模型吗?

当前接入点将 `model` 原样传给模型服务。用户应在应用、AI 工具或模型网关中完成模型选择。

关闭压缩后还能调用模型吗?

可以。关闭压缩后，接入点按原请求转发，调用方仍通过该接入点访问模型服务。需要停止接受调用时，应停用接入点。

压缩是否一定能降低费用或加快响应?

实际费用受可压缩内容、模型价格、缓存和压缩开销影响；时延还受压缩处理与模型服务推理影响。应同时验证完整任务的费用、耗时和完成情况。

是否需要先开通 Memory 实例?

使用接入点的上下文压缩能力无需先创建 Memory 实例。Agent Memory 沿用现有独立使用方式，相关文档见 Memory 应用指南 。

产品概览中的 Agent Memory 即将支持指什么?

指 Agent Memory 与 ContextTurbo 接入点的联动能力。现有独立 Memory 功能仍按官网文档使用。

版权所有：腾讯云计算（北京）有限责任公司

接入与配置

三种 API 来源如何选择?

直接使用平台统一模型 API 时，选择平台提供；沿用已有模型服务并固定模型连接配置时，选择第三方接入；希望每次请求指定模型 Base URL 与模型 API Key 时，选择透传模式。

接入点 API Key 和模型 API Key 有什么区别?

接入点 API Key 用于调用方访问 ContextTurbo；模型 API Key 用于 ContextTurbo 调用模型服务。透传模式需同时携带接入点鉴权信息，以及当前请求的模型 Base URL 和模型 API Key。

可以保留现有模型网关吗?

可以。将接入点作为网关中的模型服务连接，继续由网关管理业务权限、模型选择和路由。使用透传模式时，需确认网关可正确传递接入点要求的模型连接 Header。

可以直接使用内网 URL 吗?

需先确认调用环境具备相应内网连通条件。获得内网 URL 并不等同于已完成网络连接；接入前应验证域名解析、目标地址可达性和出站访问策略。

调用异常

请求提示鉴权失败如何排查?

1. 检查接入点 URL 与接入点 API Key 是否属于同一资源。
2. 对照快速接入示例，确认鉴权 Header 的名称和格式。
3. 第三方接入模式下，检查模型 API Key 是否有效、权限是否包含目标模型。
4. 透传模式下，检查携带模型 Base URL 和模型 API Key 的 Header 是否完整，是否被网关过滤或覆盖。

根据错误信息区分接入点鉴权失败与模型服务鉴权失败，再检查对应凭证。

请求提示模型不存在或无权使用如何排查?

核对 `model` 是否与模型服务公布的 Model ID 一致，并确认当前凭证具备该模型的调用权限。平台提供模式应使用平台模型列表中的标识。

请求超时或无法连接如何排查?

1. 确认接入点处于可用状态。
2. 检查调用方能否访问所选公网或内网地址。
3. 第三方接入模式下，检查模型 Base URL，并使用创建或配置页面中的连通性验证。
4. 检查调用方超时设置、网关出站策略与模型服务状态。

API 请求正常但 AI 工具调用异常如何排查?

检查所选模型服务是否支持该工具所需的接口、流式响应与工具调用。再对照控制台对应工具的配置示例，核对模型、请求路径和鉴权方式。

用量与费用

输入 Token 减少比例为什么与费用节省比例不同?

输入 Token 与费用不是同一口径。普通输入、缓存输入和输出可能采用不同价格，压缩也会产生开销，应结合模型计价查看净收益。

压缩限时免费是否代表模型调用免费?

不代表。上下文压缩与模型调用属于不同服务消耗，模型费用仍按对应模型服务规则执行。

用量统计能否替代正式账单?

不能。用量统计用于分析压缩效果，其中未压缩费用可能属于估算基线。费用结算应以对应服务账单为准。

版权所有：腾讯云计算（北京）有限责任公司