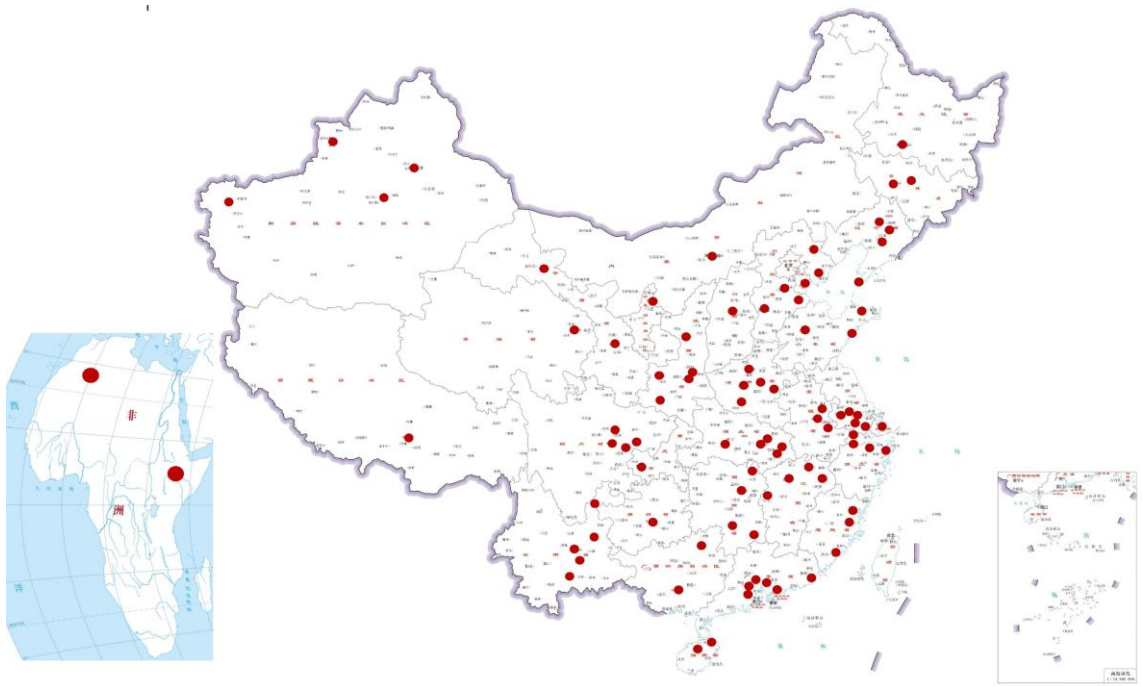
中建信息
CNBMTECH

中建信息
打破OpenAI 配额限制

中建信息
CNBMTECH

中建信息云业务介绍

中建材信息技术股份有限公司概况



诚信企业

- 创立于2005年
- 注册资本14936万元
- 中国建材集团有限公司所属企业
- **2023年实现营业收入约192.64亿元人民币**

全方位赋能

- 立足行业经验，依托自研产品，以数智化之力赋能政府、金融、互联网、教育、制造、能源、交通等众多领域
- 携手各行业合作伙伴打造良性生态圈，提供转型升级的方案和服务

规模与布局

- 现有员工超1500人
- 总部位于北京，在广州、上海、成都、沈阳、西安、武汉、南京等国内城市和埃塞俄比亚、阿尔及利亚等海外国家共设立近100个分公司和办事处
- 拥有全资子公司中建材信息科技有限公司、中建材信息技术(香港)有限公司、中建材信云智联科技有限公司，和控股子公司中建材数字科技(北京)有限公司、中建材信云智联科技(北京)有限公司

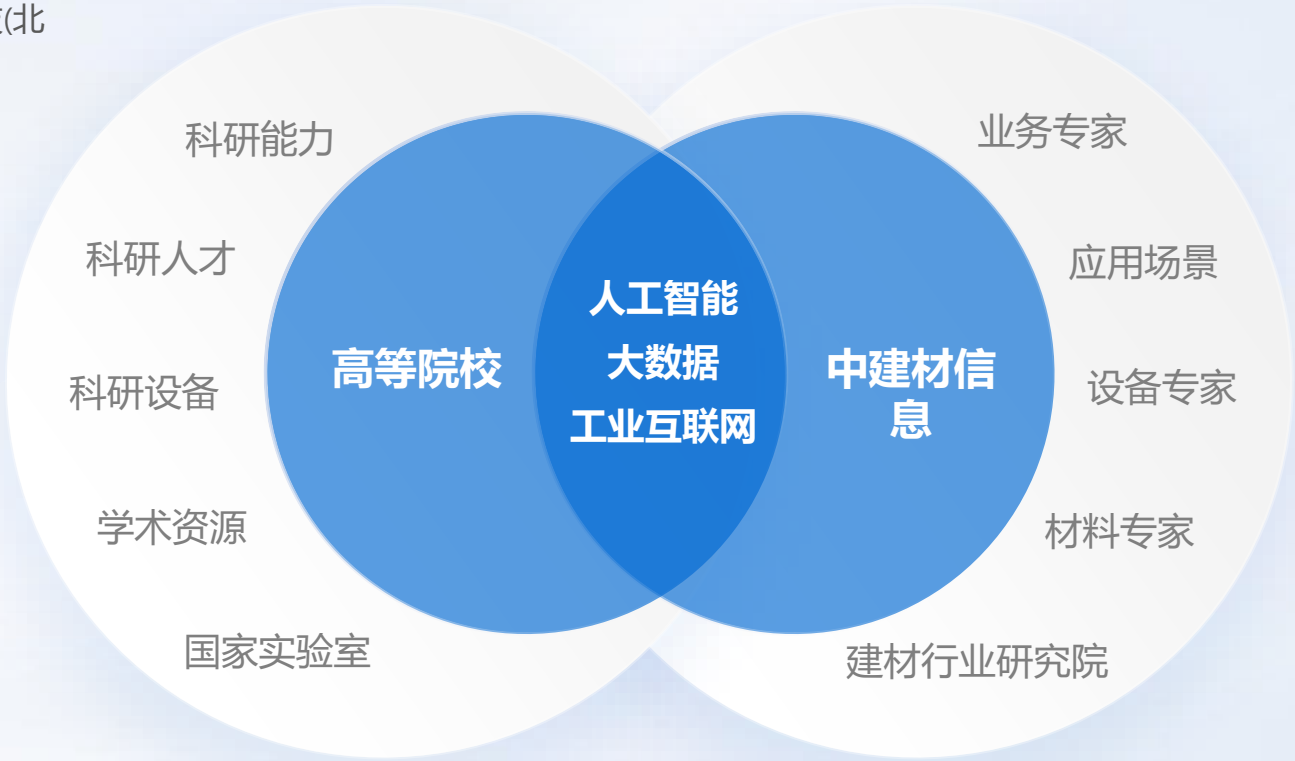
持续加强研发投入力度、重视知识产权累积

积极开展知识产权创造与保护工作

年研发投入逾亿元，目前已拥有专利72项，其中通过自主研发获得的已授权**发明专利12项**，实用新型专利55项，通过转让获取实用新型专利5项；正在申请专利170项（**发明专利122项**、实用新型48项）；新增软著22项，正在申请软著20项，已拥有软件著作权共399项〔含中建材数字科技(北京)有限公司、中建材信云智联科技(北京)有限公司〕。

聚焦产学研 推动高校人才培养和科研产出

通过加强与北京交通大学、北京航空航天大学合作，促进产教融合，构建产学研联盟创新体系，共同推进在科研创新、人才培养方面的合作，形成“校企合作、 价值创造、产学共赢”的良性循环。



资质荣誉

- 连续10年，荣获中国增值分销商十强
- 荣获2017-2021年高新技术领军企业
- 荣获2022-2023中国数字生态平台领军企业
- “人工智能助力企业实现智能化安全生产管理”获得 2023年（第五届）全国智慧企业建设创新案例典型案例

- 信云智联助力浙江水泥旗下山亚南方荣获世界水泥协会授予的2023年度“模范工厂奖”
- 信云智联与京能秦皇岛热电共同完成的“5G-AR-AI联合驱动的设备安全与人员违章综合管控系统研究与应用”项目成果被中国电力企业联合会成果监定委员会认为达到国际先进水平。

- 荣获《2023年数字转型赋能智慧发电企业建设技术专题交流会》优秀论文
- 荣获“2022-2023工业互联网建材领域领航企业”
- 数字化转型贯标试点企业
- 智能安全生产管理系统获得多项技术认证
- 荣获水泥行业供应商百强“匠心企业”奖

生态服务业务多元化布局

软/硬件

华为企业业务
总经销商

全线产品

超聚变总经销商

全线产品

绿盟科技
总经销商

软件/硬件

软件

麒麟软件
全国总经销商

软件

达梦数据库
总经销商

软件

东方通中国区
总经销商

软件

微软（中国）有限公司
总分销商

软件/云服务

SAP 金牌
合作伙伴

软件

硬件

无人机及机巢产品
平台商

硬件

华鲲振宇天宫系列
服务器全国总经销商

硬件

AMD总经销商

硬件

奔图自主可控打印机全
国区域行业总经销商

硬件

“智能+”赋能企业智慧化转型

基于在微服务、云计算、大数据和人工智能领域的技术优势，聚焦制造业，以自主研发结合集成创新的思路，为客户提供从行业专家咨询、定制开发、解决方案实施到集成与运维服务的全生命周期数字化服务，与客户和生态伙伴携手打造人工智能、工业互联网、智慧企业应用等行业解决方案，助力传统企业实现全面数字化转型。

中建信息
CNBMTECH

中建信息
打破OpenAI 配额限制

Azure OpenAI 配额限制

- Azure OpenAI 的配额限制策略主要基于订阅级别，按区域和模型类型分配 Tokens per Minute (TPM) 和 Requests per Minute (RPM) 限额
- 用户可将 TPM 灵活分配到多个部署中，总和不超过该区域的配额上限
- 不同的订阅类型和模型对应不同的配额

Model	Enterprise and MCA-E	Default	Monthly credit card-based subscriptions	MSDN subscriptions
`gpt-4.1`	5B	200M	50M	90K
`gpt-4.1 mini`	15B	1B	50M	90K
`gpt-4.1-nano`	15B	1B	50M	90K
`gpt-4o`	5B	200M	50M	90K
`gpt-4o-mini`	15B	1B	50M	90K
`gpt-4-turbo`	300M	80M	40M	90K
`gpt-4`	150M	30M	5M	100K
`gpt-35-turbo`	10B	1B	100M	2M
`o3-mini`	15B	1B	50M	90K
`o4-mini`	15B	1B	50M	90K
`gpt-5`	5B	200M	50M	90K

Microsoft Foundry 中的Azure OpenAI模型配额限制

由此引发的问题

- 面对大量并发时，调用端需要自己实现多订阅间的负载平衡
- 在单个订阅的TPM接近配额上限时，调用端会收到429错误
- 响应延迟增加和性能不稳定

解决方案

利用Azure API Management 组件实现多订阅之间的Azure OpenAI负载平衡

解决方案

- **扩展性和配额优化**：通过活跃-活跃负载均衡，将请求分布到多个订阅或部署（如预配吞吐量与标准部署），实现流量溢出和突发处理，突破单一订阅的 TPM/RPM 限额，支持多租户场景下的集中配额分配。
- **高可用性和可靠性**：支持故障转移、自动重试和断路器逻辑，当一个订阅因配额耗尽（429 错误）或故障时，智能路由到可用后端，确保服务连续性，并减少客户端代码负担。
- **性能提升**：采用优先级分组路由（如优先使用高优先级后端），避免传统轮询延迟，实现随机负载分布和无延迟切换，提高整体响应速度和吞吐量。
- **成本控制**：允许低配预配实例并用标准实例处理溢出，优化资源利用，避免过度配置导致的浪费，同时支持基于客户端的计费和展示模型。
- **安全与合规**：在网关层集中管理凭证终止、客户端识别路由和模型隔离，提供比 Azure OpenAI 实例级 IAM 更细粒度的访问控制，增强多订阅环境的安全性。
- **监控与可观测性**：统一收集跨订阅的遥测数据、日志和配额使用指标，便于仪表板可视化、警报设置和使用追踪，支持多区域冗余部署的单一控制平面管理。

负载平衡策略选择

- **Round-Robin（轮询）**：静态算法，按顺序均匀分配请求到多个后端实例（如不同订阅的 OpenAI 部署），适用于基本负载均衡场景。
- **Random（随机）**：随机选择后端，避免单一实例过载，常用于简单流量分散。
- **Priority-Based（基于优先级）**：优先路由到高优先级后端（如高配额订阅），当其不可用（如 429 限流）时降级到次级，支持分组和权重内部分配，提升资源利用率。
- **Weight-Based（基于权重）**：自定义权重（如根据 TPM 配额比例分配流量），通过策略表达式实现动态路由，优化多订阅配额使用。

API Management 配置选择

	Basic v2	Standard v2	Premium v2
适用场景	团队和项目的 API 管理	启动组织内部的 API 项目，并随着项目的发展逐步扩展。	适用于庞大请求量的企业级场景
基础价格	$150.01 per month8	$700/月	$2,801/月
扩展价格 (每增加一个单位)	$150.01/月	$500/月	$1,401/月
月请求量上限	基础价格内含10M次请求	基础价格内含50M次请求	无上限
	每增加1百万次请求$3	每增加1百万次请求$2.50	

利用生产环境数据确定 APIM SKU

类别：Azure OpenAI - HTTP 请求

展开表

指标	REST API 中的名称	单位	集合体	尺寸	时间粒度	DS 导出
Azure OpenAI AvailabilityRate 使用以下公式计算可用性百分比：(调用总数 - 服务器错误数)/调用总数。 服务器错误包括任何 >=500 的 HTTP 响应。	`AzureOpenAIAvailabilityRate`	百分比	最小值、最大值、平均值	`ApiName`、`OperationName`、`Region`、`StreamType`、`ModelDeploymentName`、`ModelName`、`ModelVersion`	PT1M	否
Azure OpenAI 请求 一段时间内对 Azure OpenAI API 的调用次数。 适用于 PTU、PTU 管理的部署以及即用即付部署。 若要细分 API 请求，可以按以下维度添加筛选器或应用拆分：ModelDeploymentName、ModelName、ModelVersion、StatusCode（成功、客户端程序、服务器错误）、IsSpillover 以获取溢出信息、StreamType（流式处理请求和非流式处理请求）和作。	`AzureOpenAIRequests`	计数	总计（总和）	`ApiName`、`OperationName`、`Region`、`StreamType`、`ModelDeploymentName`、`ModelName`、`ModelVersion`、`StatusCode`、`IsSpillover`	PT1M	是的

我们能做什么?

- **架构设计与配额规划**：评估客户流量，设计最优的多订阅/多区域/PTU+PayGo 混合架构APIM
- **统一网关一键部署**：通过 ARM/Bicep/Terraform 快速交付支持优先级+权重+自动故障转移的完整策略模板
- **智能负载均衡策略实施**：Round-Robin、权重路由、429 自动降级、断路器、突发流量溢出处理
- **企业级安全与治理**：APIM 层集中密钥管理、客户端证书/JWT 验证、IP 白名单、订阅级模型隔离、Azure Policy 合规
- **统一监控与告警**：构建跨所有订阅的配额使用、延迟、429 错误率仪表板 + 自动告警与配额预警
- **自动扩缩容与成本优化**：按月/季分析使用率，动态调整订阅数量
- **企业支持与 SLA**：7×24 监控与应急响应，保证 99.95%+ 可用性，
- **迁移与落地陪跑**：从单订阅迁移到多订阅高可用架构，全程陪跑与知识转移