# Teams Language Assistants (TLA) 需求说明书

## 概要

### 产品愿景

TLA（Teams Language Assistants）旨在彻底改变微软 Teams 中的跨语种沟通体验。现有 Teams 自带的翻译功能只能单向转换，缺乏可编辑和回贴能力，无法满足团队在全球化协作中的即时双向交流需求。本插件通过灵活接入主流或自建的大模型，实现语言自动检测、双向翻译、术语优先、语气模板等高级功能，并提供一键“翻译并回复”体验，使用户在不离开对话的情况下完成跨语种交流。

### 非目标

* **不复刻通用翻译器。** 插件不提供独立的文本翻译网页，也不承担离线文档批量翻译等非即时场景。
* **不改变 Teams 消息原文。** 原始消息始终保持不变，译文作为新回复或卡片追加呈现。
* **不替代专业人译。** 虽然支持术语库和语气模板，但不保证全部译文符合出版标准，重要场景仍需人工复核。
* **不在插件中集成外部聊天功能。** 本插件聚焦翻译与回复，不包含实时聊天或视频会议功能。

### 术语表

| 术语 | 定义 |
| --- | --- |
| Teams | Microsoft Teams 协作平台。 |
| Message Extension（ME） | Teams 中一种扩展，允许在撰写区或消息的“更多操作”中触发插件。 |
| Adaptive Card | 一种可渲染在 Teams 等宿主中的卡片格式，支持丰富交互。 |
| TLA | Teams Language Assistants 插件，为 Teams 提供即时双向翻译能力。 |
| LLM | Large Language Model，大型语言模型，可提供翻译、改写等能力。 |
| Glossary | 术语库，包含预定义词汇及其指定翻译，用于提升术语准确性[[1]](https://phrase.com/blog/posts/machine-translation-glossaries/#:~:text=What%20are%20machine%20translation%20glossaries%3F)。 |
| RAG | Retrieval Augmented Generation，通过检索外部知识库增强大模型响应准确性[[2]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=Retrieval%20augmented%20generation%20,accurate%20and%20contextually%20relevant%20responses)。 |
| OBO | On-behalf-of，代表用户在后端调用 API 的身份委托流程[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain)。 |

## 业务目标与度量

### 问题陈述

现代企业在 Teams 中频繁跨语种协作，现有翻译功能只能单向展示译文；无法基于译文直接回复、编辑；也无法优先适用组织术语或自定义语气。此外，企业希望自主选择公有或私有模型以满足合规要求。[[4]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=%23%20Do%3A%20Integrate%20with%20single,on)强调扩展不应将用户带离对话，因此需要在 Teams 内实现端到端的翻译与回复体验。

### KPI / 北极星指标

| 指标 | 定义 | 目标（MVP） |
| --- | --- | --- |
| 平均翻译回复时延（P95） | 用户点击“一键翻译并回复”至译文回贴完成的 95 分位延迟。 | ≤ 2 秒（短文本） |
| 翻译准确率提升率 | 译文满足用户意图的比例，与 Teams 默认翻译相比提升幅度。 | ≥ 10 % |
| 术语匹配命中率 | 翻译中正确应用术语库的次数 ÷ 术语出现次数。 | ≥ 95 % |
| 用户留存率 | 使用插件的月活用户 / 安装用户比例。 | ≥ 80 % |
| 内部成本控制 | 每日单租户翻译成本不超过 {DailyBudget}。 | 100 % 符合法规 |

### 成功判据

* 在私有预览阶段，90 %的试点团队认为翻译质量和工作流优于 Teams 原生翻译。
* 在 3 个月的公测期内，插件平均每日调用量达到 500 次/租户以上，且无重大安全事件。
* 合规评审通过，满足 {ComplianceRefs} 规定，数据主权遵循 {RegionPolicy}。

## 角色与权限

| 角色 | 权限与职责 |
| --- | --- |
| **租户管理员** | 安装与配置插件；设定模型与路由策略；管理租户级术语库与禁译库；配置成本上限与合规策略。 |
| **普通用户** | 在消息阅读或撰写时使用翻译功能；编辑译文并回复；配置个人偏好（目标语言、语气模板）。 |
| **审计员** | 访问审计日志及翻译留痕；查看原文、译文、模型、参数、操作者和时间线，确保合规。 |
| **Bot/消息扩展** | 后端身份，用于与 Teams 消息通信；通过 OBO 流程代用户调用翻译 API[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain)。 |

## 使用场景与用户旅程

1. **消息阅读内联翻译**：用户查看一条外语消息时，可点击“翻译”按钮或在选中文本后弹出的悬浮入口，快速查看译文并在原文下方展开译文。可切换不同目标语言，支持自动或手动检测源语言。
2. **即时回复翻译**：用户想用对方语言回复外语消息，可选择“一键翻译并回复”；系统打开编辑模式，显示原文及译文，允许用户修订译文或变更语气模板，点击“发送”后将译文作为新回复发送，并在回复卡片附带“查看原文/切换译文语言”的入口。
3. **群组多语广播**：发送者撰写群公告，可选择多个目标语言并一次性广播；插件将根据群成员首选语言生成对应译文，并回贴多条语言版本。支持术语库和语气模板选择。
4. **术语库约束翻译**：在医疗、金融等专业场景，发送者可选择特定术语库并启用禁译库，确保译文遵循专业规定。若术语与常用翻译冲突，系统提示并提供遵循术语/保留原意/折衷三选项。
5. **离线草稿**：在网络不稳定或模型响应较慢时，用户可选择“离线草稿”，插件将保存原文及翻译请求，待网络恢复后自动翻译并提醒用户回贴。

### 用户旅程示例 – 选中文本翻译并回复

1. **Given** 用户在 Teams 消息中选中了部分外语文本；
2. **When** 悬浮条显示翻译入口，用户点击“一键翻译并回复”；
3. **Then** 插件弹出对话框，自动检测语言并展示并列的原文和译文，附带语气模板和术语策略选项；
4. **Then** 用户可编辑译文或切换目标语言；点击“发送”后，插件通过 Bot 在原消息线程中回贴译文卡片，并附加“查看原文/切换语言”按钮。

## 需求一览（MVP vs vNext）

每条需求采用 MoSCoW 标注，便于项目优先级划分。MVP 表示私有预览必须具备的功能，vNext 为后续迭代。

| ID | 描述 | MoSCoW | 版本 | 可验收条件 |
| --- | --- | --- | --- | --- |
| R1 | 支持自动检测源语言并翻译为用户 Teams 首选语言；允许覆盖目标语言 | Must | MVP | Given 外语消息，当用户点击翻译时，系统正确检测源语言并翻译，译文语言与用户首选匹配，用户可手动切换目标语言。 |
| R2 | 支持双向翻译与编辑；用户可修改译文并回贴 | Must | MVP | When 翻译结果展示后，用户可以修改译文文本；点击发送后，修改后的译文被正确发送并显示。 |
| R3 | 提供术语库和禁译库；支持三层级（团队/频道/个人）合并规则 | Must | MVP | Given 指定术语条目，当译文与术语库冲突时，系统提示并提供遵循术语、保留原意或折衷的选项；按照层级优先级应用术语。 |
| R4 | 支持多模型路由（OpenAI、Azure OpenAI、Anthropic、Google、本地模型等） | Should | MVP | 管理员可配置 {ModelAllowList} 及优先级；当首选模型不可用或超出预算时，自动回退至次级模型；记录模型来源。 |
| R5 | 成本与延迟控制：支持字数阈值、缓存去重、并发与速率限制 | Must | MVP | 每次翻译请求不超过翻译服务的字符限制（例如 Azure Translator 50k 字符）[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)；配置每日租户预算 {DailyBudget}；系统记录请求成本和延迟并向用户提示。 |
| R6 | 支持质量增强：利用消息线程/频道上下文与可选 RAG 来提高准确度 | Should | vNext | 在翻译时引入上下文摘要或检索相关信息；用户可选择启用 RAG。 |
| R7 | 完整审计与可追溯：记录原文、译文、模型、参数、操作者及时间线 | Must | MVP | 审计员可查询任意翻译记录，查看译前文本、译文、使用的模型与参数、操作者标识及时间戳。 |
| R8 | 国际化与方言支持：支持简体/繁体/日语敬语等多种语言及语气模板 | Should | MVP | 在语言矩阵中支持至少 30 种语言；提供敬体、常体、商务、技术等语气模板供用户选择。 |
| R9 | 集成 Microsoft Teams 宿主能力：Message Extension、Bot、Compose Box 插件、Tab | Must | MVP | 在消息的“更多操作”和撰写区可触发翻译；Bot 能处理命令并回贴卡片；Tab 提供设置页面。 |
| R10 | 安全与合规：数据主权、密钥托管、PII/PHI 避免出境，遵循 | Must | MVP | 系统确保数据在 {RegionPolicy} 区域处理；支持将租户密钥保存在 Azure Key Vault；不将敏感信息发送到不符合合规的模型。 |
| R11 | MCP Server 与 Agent 扩展：对外暴露 translation 能力，符合 Model Context Protocol，提供工具定义 | Should | vNext | 外部 Agent 可通过 MCP 工具调用 TLA 功能，如 tla.translate、tla.detectLanguage 等，满足 JSON Schema 定义并处理错误码。 |
| R12 | 离线草稿与消息队列支持长文本分片 | Could | vNext | 对超出服务字符限制的长文本进行分片并异步翻译合并；在网络异常时保存草稿并重试。 |
| R13 | 支持多语广播与群组策略 | Could | vNext | 当用户选择多个目标语言时，插件分别生成译文并回贴；根据群成员首选语言自动推送对应版本。 |
| R14 | 提供术语与风格编辑界面；用户可上传术语库（来自 {GlossarySources}） | Could | vNext | 在 Tab 页面允许导入 CSV/TermBase 等格式的术语库，并配置优先级与回退策略。 |

## 关键功能规格

### 技术选型

开发语言：.NET Framework架构，C#优先。

本地存储（缓存）：SQLite优先

### 界面语言

界面语言以日文作为默认语言，由于大模型能力较强，可由用户自选界面语言。

文档语言为中文，代码注释为日文。

### 语言自动检测与双向翻译

1. **自动检测**：发送翻译请求前调用模型适配层 /detect 接口。采用统计方法或模型自带检测能力，需支持 100 种以上语言。若检测置信度 < 70%，向用户提供手动选择源语言。该功能应利用翻译服务的 Detect API，遵循字符数限制（每请求不超过 50,000 字符[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)）。
2. **双向翻译**：核心接口 /translate 接收 sourceText、sourceLang?、targetLang、tone?、glossaryPolicy?、modelHints? 等参数，返回 translatedText、modelId、cost、latency。支持自定义语气模板（敬体、常体、商务、技术等），通过提示工程控制模型输出。需保留原文不变，译文作为回复发送。
3. **编辑模式**：翻译完成后，用户可在 UI 中编辑译文。系统在回贴前重新调用 /rewrite 接口进行校正，保证风格统一。
4. **回贴按钮**：编辑完成后提供“发送/Reply”按钮，点击后通过 Bot 调用 Teams API 发送 Adaptive Card；卡片中包含译文、原文引用及切换语言按钮。根据 Teams 设计规范，消息扩展应在对话框内展现搜索框、内容区域和可选标签[[6]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Anatomy)。

#### 时序图（文字版）1 – 选中文本翻译并回复

User TeamsClient TLA ME Gateway/API ModelAdapter  
 | | | | |  
 |--select text-->|  
 | |--invoke ME------>|--detectLang---->|--/detect-------->|--Model.detect  
 | | |<----lang--------|<---------------| |  
 | |--translate----->|--/translate---->|--/translate------>|--Model.translate  
 | | |<--translated----|<---------------| |  
 |<-show editor---| | | |  
 |--edit+send---->|--send card----->|--replyInThread->|--Teams API-------| |  
 | | | | |

文本说明：用户在消息中选中外语文本；Teams 客户端调用消息扩展，插件通过网关请求语言检测和翻译，返回译文后在前端显示编辑器；用户修改译文并点击发送，插件通过 Bot 回复到原线程。

### 术语库/禁译库/语气模板

1. **术语库应用**：通过 /applyGlossary 接口将术语项应用于翻译文本。根据机器翻译专家建议，术语库是含有优先翻译的词汇集合，可提升术语准确性[[1]](https://phrase.com/blog/posts/machine-translation-glossaries/#:~:text=What%20are%20machine%20translation%20glossaries%3F)。在翻译前，模型适配层识别源文本中的术语并替换为预定义译文；当冲突发生时，提示用户选择遵循术语或保留原意等选项。
2. **禁译库**：包含敏感词或禁止翻译的词汇，翻译时应保留原文或替换为合法词汇。租户管理员可配置行业场景开关（金融/医疗/教育等）。
3. **语气模板**：提供敬体、常体、商务、技术四种预设模板，用户可选择在翻译时使用，以控制文体和口吻。例如敬体适合日文正式邮件，常体适合朋友聊天。模型提示词示例：请用{{tone}}语气翻译以下内容。未来支持自定义模板。
4. **层级合并规则**：术语库分为团队、频道、个人三级。优先级：个人 > 频道 > 团队。冲突时按优先级覆盖；禁译库可配置全局强制；团队管理员可设定回退策略（如未命中术语则使用默认翻译）。

#### 时序图（文字版）2 – 术语匹配流程

User TLA ME Gateway/API GlossaryService ModelAdapter  
 | | | | |  
 |--translate-->|--/translate------>|--fetchGlossary-->|--getEntries---->| |  
 | | |<--entries--------| | |  
 | |--applyGlossary-->|------------------>|--applyGlossary-->|--Model pre-process  
 | | | |<--glossApplied--| |  
 | |--call model----->|--/translate------>|--Model.translate|  
 | | |<--translated-----|<---------------| |

文字说明：翻译请求到达网关后，先从 GlossaryService 取回匹配到的术语；然后对源文本进行术语替换，再调用模型进行翻译；返回译文供前端显示。若术语和默认翻译冲突，前端弹窗供用户决策。

### 多模型路由

1. **模型适配层**：封装多个模型提供商，暴露统一接口，包括 /translate、/detect、/rewrite、/summarize。模型适配层根据租户配置和策略选择合适的模型进行调用；支持 OpenAI API、Azure OpenAI、Anthropic Claude、Google Translate、Ollama 本地模型等。
2. **路由策略**：管理员在租户配置中设置模型列表 {ModelAllowList}，可根据成本、延迟、质量、合规要求排序。当首选模型不可用或响应超时（超过 15 秒[[7]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=The%20Translator%20has%20a%20maximum,characteristics%20on%20sustained%20request%20rate)），自动回退至次级模型，并记录回退事件。
3. **成本控制**：网关根据模型调用成本估算每次请求的费用并与每日预算 {DailyBudget} 对比；若预计超过预算，则拒绝请求并提示用户“今日翻译预算已用尽”。Azure Translator 按字符计费[[8]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)，需实时统计翻译字符数。
4. **延迟优化**：对小于 100 字符的请求，期望在 300 毫秒内返回翻译[[7]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=The%20Translator%20has%20a%20maximum,characteristics%20on%20sustained%20request%20rate)；超过字符阈值将分片并行翻译；每个模型连接最大并发限制由管理员配置，默认不超过 5 并发请求。

#### 时序图（文字版）3 – 多模型路由与回退

User TLA ME Gateway/API ModelAdapter ProviderA ProviderB  
 | | | | | |  
 |--translate-->|--/translate---->|--selectModel---->|--call A------>|--OK? |  
 | | | |<--error/timeout---------|  
 | | |--retryModel----->|--call B------>|--translated  
 | | |<--translated-----|<---------------| |

文字说明：当首选模型（Provider A）响应错误或超时时，模型适配层根据路由策略选择 Provider B 继续翻译；网关记录耗时与成本并返回给前端。

### 成本/延迟控制

* **字数阈值与分片**：根据 Azure Translator 每请求 50k 字符限制[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)，接口预处理将长文本分成不超过 5k 字符的片段；对于超过 2 万字符的文本，排入异步队列进行批量翻译，前端以草稿形式存储，待完成后通知用户。
* **缓存与去重**：在模型适配层实现语义去重缓存，使用向量索引保存近期翻译结果；对重复句子直接返回缓存，避免重复费用。可配置 TTL（默认 24 小时）和租户隔离。
* **并发与速率限制**：遵守第三方翻译服务的字符/小时限额：Azure Translator F0 层每小时 2 M 字符[[9]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=Character%20limits%20per%20hour)；同时控制请求速率均匀分布。网关对每个租户和模型维度实施令牌桶限流，防止瞬时流量过高。

### 质量增强

1. **上下文引用**：在翻译时携带上下文摘要（如线程标题、前几条消息），帮助模型理解语境。通过调用 /summarize 接口生成上下文摘要，再拼接入翻译请求。
2. **可选 RAG**：在专业场景，用户可启用 RAG 模式，通过检索企业知识库或公开文档获取补充材料，结合大模型产生更准确的译文。RAG 可以从外部文档或向量数据库中检索相关内容[[2]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=Retrieval%20augmented%20generation%20,accurate%20and%20contextually%20relevant%20responses)，进而提高准确性；它能提供更高的命中率和准确性以及域专业化优势[[10]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=By%20definition%2C%20retrieval%20augmented%20generation,of%20AI%20in%20several%20ways)。
3. **风格锁定**：使用提示工程确保译文保持一致文风，结合用户历史译文训练的小模型微调实现；默认提供四种语气模板，未来支持自定义。若模型风格输出偏差较大，系统自动触发 /rewrite 调整风格。

### 审计与可追溯

* **日志内容**：审计系统存储每个翻译的原文、译文、时间戳、操作者、使用模型 ID、参数配置、成本、延迟、模型响应 ID 等信息。审计数据按租户隔离，审计员可用时间范围和过滤条件查询。
* **权限控制**：审计员身份由租户管理员授权，不可修改译文；日志只读，提供导出功能（CSV/JSON）。
* **不可逆哈希**：为避免泄露原文，在日志中存储原文指纹（例如 SHA256 哈希 + 盐）而非全文；只有持有审计密钥的管理员才能解密。
* **合规留痕**：记录是否启用了术语库、禁译库、RAG、语言检测等功能；便于审计时复盘与责任追溯。

## Microsoft Teams 宿主能力映射

| 宿主能力 | 落地方式 | 描述 |
| --- | --- | --- |
| **Message Extension（ME）** | 撰写区触发、选择消息触发 | 用户在撰写区点击插件图标或在消息“更多操作”选择“TLA”，弹出对话框显示翻译结果；支持搜索和列表视图[[6]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Anatomy)。 |
| **Bot** | 命令词/上下文卡片 | Bot 处理与外部 API 的通信；可支持命令如 /translate、/help，并通过 Adaptive Card 回贴译文。 |
| **Compose Box 插件** | 内联建议与替换 | 在撰写区实时监测输入的外语，提示用户是否翻译；提供一键替换文本为目标语言。 |
| **Tab** | 个人/团队设置页面 | 管理模型密钥、术语库、策略；提供文件上传以导入术语库 ({GlossarySources})；展示使用统计和预算。 |
| **Graph/身份与权限** | AAD/OBO、SSO、租户级密钥管理 | 使用 Single Sign On 简化登录[[11]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Set%20up%20a%20message%20extension)；Bot 通过 OBO 流程代表用户调用翻译 API，传递其身份和权限[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain)；租户密钥存储在 Azure Key Vault。 |

## UI/UX 明细

### 总体布局

由于未提供具体草图，本说明基于微软 Teams UI Kit 及通用设计模式进行推测性还原。所有尺寸和间距可按草图调整。

#### 桌面端

1. **顶部工具条**：高度 48 px，横向排列控制项，包含：
2. **模型选择下拉框**：宽 120 px，显示当前模型名称，可展开选择列表。
3. **语种切换**：两个下拉框，分别为“源语言”（默认为自动）和“目标语言”（默认为用户首选），各宽 100 px。
4. **术语策略开关**：Toggle 控件，启用/禁用术语库应用。
5. **成本/延迟提示**：文本标签，显示本次翻译预计成本（字符数×单价）与估算延迟，如“成本：0.03$ | 延迟：300ms”。
6. 控件间水平间距 8 px。
7. **主区**：占用宽度自适应，高度可滚动，分为左右两栏：
8. **原文块**：左栏，宽度 50%，背景浅灰，内含原文文本的只读框，支持复制；顶部显示源语言；底部显示字符数。
9. **译文块**：右栏，宽度 50%，背景白色或淡黄色，用于显示模型生成的译文；当进入编辑模式时变为可编辑文本域；其下方放置“回贴/发送”按钮（主按钮，宽 88 px，高 32 px，圆角 4 px）和“取消”按钮。
10. 两栏之间间距 16 px；整体外边距 20 px。
11. **悬浮/快捷入口**：当用户在消息中选中文本时，在选区附近显示浮动操作条（高度 32 px）；包含“翻译”按钮（图标 + 文本）和“翻译并回复”按钮。浮动条背景为 Fluent UI Neutral 轻色，阴影不遮挡原文。

#### 移动端

1. **布局改为上下排列**：由于屏幕宽度有限，原文块在上、译文块在下；顶部工具条滚动为水平可滑动列表，允许左右滑动查看所有控制项。
2. **按钮尺寸增加**：确保触控友好；按钮高度 ≥ 40 px，间距 12 px。
3. **回复卡片**：在消息线程中显示紧凑版本，提供切换语言按钮，点击后在全屏对话框中展示完整译文及原文。

#### 默认文案与国际化 Key

| 区域 | 默认文案 | i18n Key |
| --- | --- | --- |
| 模型选择下拉 | “选择模型” | tla.ui.modelSelector.placeholder |
| 源语言下拉 | “自动检测” | tla.ui.sourceLang.auto |
| 目标语言下拉 | “目标语言” | tla.ui.targetLang.placeholder |
| 术语开关 | “应用术语库” | tla.ui.glossary.toggle |
| 成本提示 | “预计成本：{cost}，延迟：{latency}” | tla.ui.costLatency |
| 原文标题 | “原文 ({lang})” | tla.ui.source.title |
| 译文标题 | “译文 ({lang})” | tla.ui.translation.title |
| 编辑提示 | “您可以修改译文后发送” | tla.ui.editable.hint |
| 发送按钮 | “发送” | tla.ui.send |
| 取消按钮 | “取消” | tla.ui.cancel |
| 悬浮条翻译 | “翻译” | tla.ui.hover.translate |
| 悬浮条翻译并回复 | “翻译并回复” | tla.ui.hover.translateReply |

### Traceability 表

下表展示界面控件、交互、调用 API 和验收用例之间的关系。

| 控件 | 用户交互 | 调用 API/工具 | 验收用例 |
| --- | --- | --- | --- |
| 模型选择下拉 | 选择不同模型 | 更新模型适配层的 modelHints; 调用 /translate 时传入 | Given 用户切换模型，When 再次翻译时使用新的模型并显示模型 ID；Then 翻译成本和延迟变化记录。 |
| 源语言下拉 | 手动指定源语言 | 将 sourceLang 传入 /translate，跳过检测 | Given 检测置信度低，When 用户手动选择源语言，Then 翻译结果应改变且不再提示自动检测。 |
| 目标语言下拉 | 切换目标语种 | 更新 targetLang 参数 | Given 用户切换目标语言，Then 翻译文本立即刷新为新语言。 |
| 术语库开关 | 开启或关闭术语应用 | 设置 glossaryPolicy；调用 /applyGlossary | Given 术语库开启，When 翻译包含术语，Then 显示术语识别和三选项提示。 |
| 成本/延迟提示 | 查看成本估算 | 调用 tla.getCostLatency(payloadSize, modelId) | Given 输入文本长度变化，Then 提示中的成本和延迟自动更新。 |
| 原文文本框 | 只读展示原文 | 无 API 调用 | Given 大文本时，文本框滚动正常，不允许编辑。 |
| 译文文本框 | 展示并允许编辑译文 | 初始调用 /translate; 编辑后调用 /rewrite | Given 用户编辑译文并发送，Then 回贴卡片内容应是编辑后的文本。 |
| 发送按钮 | 点击发送译文 | 调用 tla.replyInThread | Given 用户点击发送，Then Adaptive Card 显示在原消息线程并包含原文引用。 |
| 悬浮条翻译按钮 | 快速查看译文 | 调用 /translate，在卡片内展示 | Given 选中文本，When 点击翻译，Then 卡片内显示译文。 |
| 悬浮条翻译并回复 | 快速翻译并弹出编辑 | 调用 /translate + 进入编辑模式 | Given 选中文本，When 点击翻译并回复，Then 弹出完整编辑对话框可发送。 |
| Tab 上传按钮 | 上传术语库文件 | 调用 /applyGlossary 上传端点 | Given 上传 CSV 文件成功，Then 术语库条目可用于翻译并在系统管理界面显示。 |

## 架构设计

### 高层架构文字描述

Teams 客户端 ←→ TLA 插件（Message Extension/Bot/Tab） ←→ 网关（API）  
 ←→ 模型适配层（Provider SDK） ←→ 术语/RAG/缓存/审计存储

1. **Teams 客户端**：包括桌面端、移动端和 Web。用户通过消息扩展、Bot 或撰写区插件与 TLA 交互。
2. **TLA 插件**：前端部分由 Teams JS SDK 实现，实现 UI 渲染和用户交互逻辑；通过消息扩展调用后端 API；在 Bot 场景下接收命令并触发翻译；在 Tab 页面用于设置。
3. **网关（API）**：核心后端服务，提供 REST 或 Graph 类接口，负责身份验证、速率限制、日志记录、成本计费、请求路由和错误处理。网关通过 OBO 流程获取用户访问令牌并保护下游服务安全。
4. **模型适配层**：抽象不同提供商接口，统一实现 /translate、/detect、/rewrite、/summarize 方法。每个 Provider 插件实现 Provider SDK 的抽象类，负责调用相应的 API（如 OpenAI、Azure Translator 等），并返回标准化响应。
5. **术语/RAG/缓存/审计存储**：
6. **术语库服务**：存储和检索术语条目，根据用户租户和优先级返回匹配词汇；支持 CSV/TermBase 导入和词条管理。
7. **RAG 知识库**：可选组件，保存向量化的文档和消息历史，为翻译提供上下文检索能力。
8. **缓存服务**：保存近期翻译的结果，提供语义去重和 TTL 管理；按租户隔离。
9. **审计存储**：持久化审计日志；根据 {RegionPolicy} 部署在相应区域，满足数据主权要求。

### 组件职责、边界与接口

| 组件 | 职责 | 接口 |
| --- | --- | --- |
| **Gateway/API** | 调度请求；权限校验；租户隔离；负载均衡；成本计费；错误处理 | REST API /translate, /detect, /rewrite, /summarize, /applyGlossary, /replyInThread, /getCostLatency；WebSocket 用于异步通知。 |
| **ModelAdapter** | 根据租户配置路由请求到不同模型；封装不同模型协议；统一返回格式 | Provider SDK 抽象类，定义 translate(text, sourceLang, targetLang, options), detect(text), rewrite(text, tone), summarize(text) 等；针对 OpenAI/Azure/Anthropic/Ollama 等实现各自适配器。 |
| **GlossaryService** | 管理术语与禁译库；提供三层级优先级合并；支持批量上传和检索 | getEntries(text, tenantId, channelId, userId), applyGlossary(text, glossaryIds[], policy), uploadGlossary(file). |
| **RAGService** (可选) | 检索上下文和知识文档并生成参考片段 | retrieve(query, topK), embed(text), addDocument(doc)；返回匹配文本供模型提示。 |
| **CacheService** | 保存翻译结果和检测结果；提供向量索引实现语义匹配 | get(key), set(key, value, ttl), searchSimilar(embedding)；单租户隔离。 |
| **AuditService** | 记录翻译任务和响应；提供审计查询接口；支持不可逆哈希 | log(entry), query(filter), export(format)；仅审计员可访问。 |
| **QueueService** | 处理超长文本与离线任务；支持重试与重组 | enqueue(job), dequeue(), updateStatus(jobId, status)；保证任务执行顺序和幂等性。 |

### 缓存策略

* **语义去重**：使用向量索引（如 cosine similarity）比较新翻译请求与缓存中已有请求；相似度 > 0.95 则直接返回缓存结果，减少重复费用。
* **TTL 管理**：缓存条目默认 24 小时，管理员可配置；命中率统计用于监控。
* **租户隔离**：缓存按 TenantId 分区，防止数据串用。

### 异步队列与重试

* **分片规则**：当文本长度 > 5k 字符时，拆分为句子或段落片段，生成任务列表；每个任务含文本片段、顺序号和重试计数；入队后异步翻译。
* **并行翻译**：队列工作者根据模型并发限制拉取任务并处理；翻译完成后在任务合并器中按顺序拼接译文。
* **失败重试**：每个任务支持 3 次重试；失败原因记录在审计日志中；当重试用尽时，将任务标记为失败并通知用户。
* **长文本回调**：网关通过 WebSocket 或 Teams Notification 卡片向用户推送完成通知；用户可在草稿列表查看并发送译文。

### MCP Server 与 Agent 扩展预留

MCP（Model Context Protocol）允许外部智能体以结构化方式调用 TLA 能力，使得 Copilot 等 Agent 能够协同翻译任务。

#### MCP 高层设计

* **MCP Server**：托管在 TLA 网关内，暴露一组工具方法供 Agent 调用；根据智能体请求安全地代理到实际功能。
* **Agent 扩展**：在 Copilot 或第三方 Agent 中注册 TLA 工具，使智能体可以使用 translate、detectLanguage、applyGlossary、replyInThread 等能力。
* **安全模型**：每个工具调用包含 tenantId、channelId、userId 等标识，通过 token 交换和权限裁剪确保仅在授权范围内执行。外部 Agent 只能使用只读工具（如检测语言、获取成本）或在授权的消息线程中使用可写工具（如 replyInThread）。

#### MCP Tools 规范（示例 JSON Schema）

{  
 "$schema": "http://json-schema.org/draft-07/schema#",  
 "title": "tla.translate",  
 "type": "object",  
 "properties": {  
 "sourceText": { "type": "string", "description": "待翻译的原文" },  
 "sourceLang": { "type": "string", "description": "可选的源语言代码" },  
 "targetLang": { "type": "string", "description": "目标语言代码" },  
 "tone": { "type": "string", "enum": ["polite", "casual", "business", "technical"], "description": "语气模板" },  
 "glossaryPolicy": { "type": "string", "enum": ["apply", "bypass", "askOnConflict"], "description": "术语应用策略" },  
 "modelHints": { "type": "object", "description": "模型路由提示，如优先模型列表" }  
 },  
 "required": ["sourceText", "targetLang"]  
}

同样的结构可用于其他工具：

{  
 "title": "tla.detectLanguage",  
 "type": "object",  
 "properties": { "text": { "type": "string" } },  
 "required": ["text"]  
}  
  
{  
 "title": "tla.applyGlossary",  
 "type": "object",  
 "properties": {  
 "text": { "type": "string" },  
 "glossaryIds": { "type": "array", "items": { "type": "string" } },  
 "policy": { "type": "string", "enum": ["strict", "fallback"] }  
 },  
 "required": ["text", "glossaryIds"]  
}  
  
{  
 "title": "tla.replyInThread",  
 "type": "object",  
 "properties": {  
 "threadId": { "type": "string" },  
 "replyText": { "type": "string" },  
 "languagePolicy": { "type": "object", "description": "包含目标语言、语气模板等信息" }  
 },  
 "required": ["threadId", "replyText"]  
}  
  
{  
 "title": "tla.getCostLatency",  
 "type": "object",  
 "properties": {  
 "payloadSize": { "type": "integer" },  
 "modelId": { "type": "string" }  
 },  
 "required": ["payloadSize", "modelId"]  
}

#### 事件流

* **MessageSelected**：用户在消息或撰写区选中文本时触发；可供 Agent 获取上下文。
* **BeforeSend**：在用户发送译文前触发；可进行术语检查、合规校验。
* **AfterTranslate**：翻译完成事件；返回译文、成本、延迟等信息。
* **PolicyViolation**：当检测到敏感词或违背合规策略时触发；包含违规详情和建议处理方式。

#### Agent 协作

* **仲裁/投票翻译**：多个 Agent 可同时调用 tla.translate 提供备选译文；插件根据规则（投票或质量评分）选择最佳结果并呈现给用户。
* **链式优化**：Agent 可以按顺序调用 detectLanguage → translate → applyGlossary → rewrite → replyInThread，每一步获取前一阶段结果并进行改进。

### 数据模型

#### 核心实体

| 实体 | 说明 | 主键与索引 |
| --- | --- | --- |
| **TranslationJob** | 描述一次翻译任务，包括原文、译文、状态、成本、延迟、模型 ID、片段列表 | 主键 JobId；索引 TenantId + CreatedAt；状态字段 Pending/Processing/Completed/Failed；支持子表存储分片任务。 |
| **MessageContext** | 保存消息信息：ThreadId、ChannelId、UserId、Timestamp、首选语言等；用于翻译上下文和权限检查 | 主键 ContextId；索引 ThreadId；包含外键到用户与频道。 |
| **GlossaryEntry** | 术语库条目：源词、目标词、语言对、优先级、备注；可属于特定租户/频道/用户 | 主键 EntryId；复合键 TenantId + SourceTerm + TargetLang；索引 GlossaryId。 |
| **Policy** | 存储租户配置，包括模型路由、预算、禁译库、RAG 开关、语气模板等 | 主键 TenantId；包含 JSON 配置；支持版本控制。 |
| **ProviderConfig** | 每个模型提供商的配置：API Key、Endpoint、优先级、费用估算等 | 主键 ProviderId；索引 TenantId（多租户各自配置）。 |
| **AuditTrail** | 审计记录：JobId、UserId、Action、Timestamp、HashDigest、ModelId、Parameters、ResultStatus | 主键 AuditId；索引 TenantId + Timestamp；存储指纹和敏感字段加密。 |

#### 多租户隔离

* **TenantId + Partition Key**：所有表均以 TenantId 为分区键；数据按租户隔离，保障安全。
* **ChannelId & UserId**：细粒度标识上下文；在权限层面与 Teams 的授权模型对齐。

### 接口契约

#### 端点与方法

以 REST 形式示例，亦可通过 Graph 或 WebSocket 实现；所有请求需携带 OAuth Bearer Token。错误码按 HTTP 状态码和业务码组合。

1. **POST /translate**
2. **请求参数**：

* {  
   "sourceText": "string",  
   "sourceLang": "string(optional)",  
   "targetLang": "string",  
   "tone": "polite|casual|business|technical(optional)",  
   "glossaryPolicy": "apply|bypass|askOnConflict(optional)",  
   "modelHints": { "primary": "gpt-4o", "fallback": ["gpt-35-turbo", "ollama:mixtral"] }  
  }

1. **响应示例**：

* {  
   "jobId": "uuid",  
   "translatedText": "译文内容",  
   "modelId": "gpt-4o",  
   "sourceLang": "en",  
   "targetLang": "zh",  
   "toneApplied": "business",  
   "cost": 0.0023,  
   "latencyMs": 312  
  }

1. **错误码**：400（参数错误）、402（超出预算）、503（模型不可用）、504（超时）。
2. **幂等性**：通过 jobId 去重；同样的 sourceText + targetLang 若参数相同，则返回缓存结果。
3. **速率限制**：按租户限制请求频率，默认 60 QPS；大于限额返回 429。
4. **POST /detect**
5. **请求**：{ "text": "string" }
6. **响应**：{ "language": "code", "confidence": 0.93 }
7. **限额**：单次请求不得超过 50k 字符[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)；超出返回 413。
8. **POST /rewrite**
9. **请求**：{ "text": "译文", "tone": "business" }
10. **响应**：{ "rewrittenText": "调整后的译文" }
11. 用于用户编辑译文后，通过模型微调语气。
12. **POST /summarize**
13. **请求**：{ "context": "多条消息文本" }
14. **响应**：{ "summary": "简短摘要" }
15. 用于质量增强时提供上下文摘要。
16. **POST /applyGlossary**
17. **请求**：{ "text": "string", "glossaryIds": ["g1","g2"], "policy": "strict|fallback" }
18. **响应**：{ "processedText": "已替换术语的文本", "matches": [ {"term":"CPU","translation":"中央处理器"} ] }
19. 如果 policy 为 fallback，当未匹配到术语时直接返回原文。
20. **POST /replyInThread**
21. **请求**：{ "threadId": "id", "replyText": "译文内容", "languagePolicy": {"targetLang": "zh", "tone": "business"} }
22. **响应**：{ "messageId": "id", "status": "sent" }
23. 通过 Bot 调用 Teams API 发送 Adaptive Card。错误码 403 表示无权限写入线程。
24. **GET /getCostLatency**
25. **请求参数**：payloadSize（整数），modelId（字符串）。
26. **响应**：估算的成本和延迟，例如：{ "cost": 0.0015, "latencyMs": 250 }。
27. 根据模型定价和服务限制预估；用以 UI 提前提示用户。

#### 鉴权

* 使用 Azure AD OAuth2，采用 SSO 流程，当用户已在 Teams 中登录时无需再次登录[[11]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Set%20up%20a%20message%20extension)。
* 后端 API 使用 OBO 流程将用户令牌换取访问下游服务的令牌[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain)，确保调用采用用户的委托权限。
* 租户管理员可在 Tab 页面配置应用权限，指定哪些频道或用户可以调用写操作；超出范围时返回 403。

#### 错误码

| 代码 | 描述 | 处理建议 |
| --- | --- | --- |
| 400 | 请求参数无效，如缺失必填字段 | 检查参数格式并重试 |
| 401 | 未授权或 token 过期 | 重新登录或刷新 token |
| 402 | 超出成本预算 | 等待下一日或联系管理员增额 |
| 403 | 无权限执行操作，如在未授权的频道发送 | 检查租户策略或申请权限 |
| 413 | 请求体过大，超过字符限制 50k[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation) | 将文本分片或使用离线草稿 |
| 429 | 速率限制触发 | 稍后重试或降低并发 |
| 503 | 模型提供商不可用 | 系统将自动回退，建议稍后重试 |
| 504 | 模型调用超时（>15 秒）[[7]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=The%20Translator%20has%20a%20maximum,characteristics%20on%20sustained%20request%20rate) | 系统已自动回退，若仍失败则通知用户 |

## 安全与合规

* **数据主权**：所有数据处理均遵循 {RegionPolicy}。用户数据在所在地区处理，不跨境传输。模型提供商必须在相应地区部署。
* **PII/PHI 保护**：在翻译请求中识别个人身份信息和健康信息，除非用户显式授权，不将其发送到公有模型。敏感场景（金融、医疗）通过禁译库确保不翻译敏感词。
* **密钥托管**：租户管理员在 Tab 中配置模型 API 密钥，存储于 Azure Key Vault；网关通过托管身份访问密钥，杜绝明文传输。
* **日志与留痕**：审计日志包含指纹和时间线，审计员可追溯；使用不可逆哈希保护原文[[1]](https://phrase.com/blog/posts/machine-translation-glossaries/#:~:text=What%20are%20machine%20translation%20glossaries%3F)。
* **误译纠错与人工复核**：为高风险场景提供“人工复核模式”，译文先生成草稿，由专业译者复核后方可发送；管理员可配置触发场景（例如包含医疗术语或检测到低置信度）。

## 性能与容量

| 指标 | 目标 | 说明 |
| --- | --- | --- |
| **端到端 P95 延迟** | ≤ 2 秒（短文本）；≤ 5 秒（中等文本） | 包括网络、检测、翻译、回贴时延 |
| **端到端 P99 延迟** | ≤ 5 秒（短文本）；≤ 10 秒（中等文本） | 针对异常情况；超过阈值触发重试或回退 |
| **最大并发请求** | ≥ 1000 RPS | 通过横向扩展网关和模型适配层保证 |
| **长文本分片上限** | 每条翻译请求拆分的片段数 ≤ 10 | 避免合并过多片段造成性能下降 |
| **缓存命中率** | ≥ 40 %（MVP） | 通过向量索引提升命中率 |

## 监控与告警

* **可观测性指标**：成功率、超时率、平均/95/99 延迟、每模型成本、缓存命中率、术语冲突率、RAG 检索耗时等。
* **日志采集**：使用集中式日志服务收集网关、模型适配层、队列等组件日志；日志以结构化格式存储。
* **告警策略**：
* 当失败率 > 5 % 或延迟 > 3 秒持续 5 分钟触发告警。
* 当某个模型成本急剧上升（超出预算 50 %）时提醒管理员切换模型。
* 当缓存命中率低于 20 % 时提示优化术语库或查询策略。

## 验收标准与测试用例

测试用例按功能、兼容、无障碍、安全、恢复力、多租户和回归分类。以下示例：

### 功能测试

| Given | When | Then |
| --- | --- | --- |
| 用户查看外语消息 | 点击“翻译”按钮 | 系统自动检测语言并在原文下方显示译文；用户可切换目标语言 |
| 用户选中文本 | 点击“翻译并回复” | 弹出对话框显示原文与译文，允许编辑，并显示发送按钮 |
| 用户开启术语库 | 翻译文本包含术语 | 提示术语匹配并提供遵循/保留/折衷选项；译文中应用术语 |
| 模型 A 不可用 | 翻译请求发生超时 > 15 秒 | 系统自动回退到模型 B，记录回退原因并返回译文 |
| 翻译字符 > 50k | 发送翻译请求 | API 返回错误码 413，提示用户拆分或离线草稿 |
| 达到每日成本上限 | 用户继续发送翻译 | API 返回 402，提示预算耗尽，不再处理 |
| 审计员查询记录 | 提供时间范围过滤 | 返回符合条件的翻译记录，包含原文指纹、译文、操作者等；敏感内容可遮蔽 |

### 兼容性测试

* 在 Windows、macOS、iOS、Android 客户端测试插件界面和交互，确保布局响应式；
* 使用深色、高对比度主题检查 UI 是否遵循 Teams 颜色令牌[[12]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=,Teams%20color%20tokens)。

### 无障碍测试

* 所有控件支持键盘导航，提供清晰的焦点状态；
* 插件文本可被屏幕阅读器读出；
* 颜色对比符合 WCAG 标准；提供“Alt”文本和 ARIA 标签。

### 安全测试

* 检查 OAuth 流程，确保 access token 不被泄露；
* 注入测试：验证恶意输入不会被注入到下游模型或数据库；
* 权限测试：普通用户无权访问审计接口；租户之间数据隔离。

### 恢复力测试

* 模拟模型服务不可用，验证系统自动回退并提供提示；
* 模拟网络中断，验证离线草稿保存与恢复；
* 重试逻辑：对 transient error 重试 3 次并记录。

### 多租户测试

* 在两个租户分别配置不同模型和术语库，确保数据互不影响；
* 在租户 A 的频道发送翻译请求，租户 B 无法获取结果或日志。

### 回归测试

* 每次版本迭代后执行完整测试集，确保新增功能不影响原有流程。

## 里程碑与发布策略

| 阶段 | 时间 | 内容 |
| --- | --- | --- |
| **MVP 私有预览** | 第 0–2 个月 | 完成基础翻译、术语库、双向回复、成本控制、审计日志；在选定租户内试点；收集反馈和故障记录。 |
| **公测 Beta** | 第 3–5 个月 | 扩展到更多租户；加入多模型路由与回退；支持离线草稿和移动端优化；完善监控与告警；开放 MCP Server。 |
| **正式版 GA** | 第 6 个月 | 根据反馈优化 UI 和性能；支持 RAG 与风格锁定；发布术语上传管理；通过合规审核。 |
| **vNext** | 第 6 个月之后 | 开发多语广播、仲裁翻译、智能 Agent 协同、人机共译等高级功能；持续迭代。 |

发布策略采用灰度发布：先在小部分租户启用新功能，监控指标稳定后逐步扩大；支持一键回滚到前一版本，确保风险可控。

## 风险清单与对策

| 风险 | 影响 | 对策 |
| --- | --- | --- |
| **API 限流或停用** | 第三方模型限额导致翻译失败 | 提前设置多模型路由；监控各模型配额；在预算耗尽前切换到备用模型 |
| **模型不可用或质量波动** | 翻译质量下降 | 实施 A/B 对比和回退策略；允许管理员手动禁用模型；多提供商冗余 |
| **术语冲突** | 译文不符合规范或含有歧义 | 提示用户选择处理方式；记录冲突率；持续优化术语库管理 |
| **长文本卡顿** | 性能下降和用户体验差 | 实施分片和异步队列；提供离线草稿；限制单次翻译长度 |
| **合规审计要求突增** | 日志存储压力与响应性能 | 优化日志存储结构，增加索引；使用分区存储；扩展审计服务容量 |
| **数据泄露风险** | 可能违反隐私法规 | 加强密钥管理；使用加密传输和不可逆哈希；定期安全审计 |

## 附录

### 术语与缩写

| 缩写 | 含义 |
| --- | --- |
| PII | Personally Identifiable Information，个人身份信息 |
| PHI | Protected Health Information，受保护的健康信息 |
| RAG | Retrieval Augmented Generation[[2]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=Retrieval%20augmented%20generation%20,accurate%20and%20contextually%20relevant%20responses) |
| OBO | On-behalf-of 授权[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain) |
| LLM | Large Language Model |

### 示例术语库条目

源词,目标词,源语言,目标语言,备注  
CPU,中央处理器,en,zh,技术术语请保留完整翻译  
compliance,合规,en,zh,金融行业专用  
Connected,Connected,en,es,品牌名禁止翻译

### 消息卡片 JSON 示例

{  
 "$schema": "http://adaptivecards.io/schemas/adaptive-card.json",  
 "type": "AdaptiveCard",  
 "version": "1.5",  
 "body": [  
 {  
 "type": "TextBlock",  
 "text": "译文",  
 "wrap": true,  
 "size": "Medium",  
 "weight": "Bolder"  
 },  
 {  
 "type": "TextBlock",  
 "text": "原文: ...",  
 "wrap": true,  
 "isSubtle": true,  
 "spacing": "None"  
 }  
 ],  
 "actions": [  
 {  
 "type": "Action.ShowCard",  
 "title": "查看原文",  
 "card": {  
 "type": "AdaptiveCard",  
 "body": [  
 { "type": "TextBlock", "text": "原文内容", "wrap": true }  
 ],  
 "actions": [],  
 "version": "1.5"  
 }  
 },  
 {  
 "type": "Action.Submit",  
 "title": "切换译文语言",  
 "data": { "action": "changeLanguage" }  
 }  
 ]  
}

### 配置清单

| 配置项 | 说明 |
| --- | --- |
| {RegionPolicy} | 数据主权策略，如“欧盟优先”，“中国境内”等 |
| {ModelAllowList} | 允许使用的大模型及优先级列表，如 ["azureOpenAI:gpt-4o", "anthropic:claude-3", "ollama:llama3"] |
| {DailyBudget} | 每个租户每日成本上限，例如 10 USD |
| {ComplianceRefs} | 遵循的安全与合规框架，如 ISMAP、SOC2、ISO27001 |
| {GlossarySources} | 术语库来源，如 CSV 文件、TermBase 格式或外部词典 API |

本说明书结合 Teams 插件的设计原则和翻译服务的最佳实践编写，并使用最新文档提供的权威信息，如消息扩展应避免将用户带离对话、集成单点登录[[4]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=%23%20Do%3A%20Integrate%20with%20single,on)、翻译服务字符和延迟限制[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation)[[7]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=The%20Translator%20has%20a%20maximum,characteristics%20on%20sustained%20request%20rate)、术语库重要性[[1]](https://phrase.com/blog/posts/machine-translation-glossaries/#:~:text=What%20are%20machine%20translation%20glossaries%3F)、RAG 优势[[2]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=Retrieval%20augmented%20generation%20,accurate%20and%20contextually%20relevant%20responses)[[10]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=By%20definition%2C%20retrieval%20augmented%20generation,of%20AI%20in%20several%20ways)和 OBO 授权模型[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain)等，为研发团队提供可执行的方案。在执行过程中，请始终遵守用户数据保护和合规要求，同时根据实际草图调整 UI 尺寸与布局。

[[1]](https://phrase.com/blog/posts/machine-translation-glossaries/" \l ":~:text=What%20are%20machine%20translation%20glossaries%3F) A Detailed Guide to Machine Translation Glossaries | Phrase

<https://phrase.com/blog/posts/machine-translation-glossaries/>

[[2]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=Retrieval%20augmented%20generation%20,accurate%20and%20contextually%20relevant%20responses) [[10]](https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/#:~:text=By%20definition%2C%20retrieval%20augmented%20generation,of%20AI%20in%20several%20ways) Retrieval Augmented Generation (RAG): A Complete Guide - WEKA

<https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/>

[[3]](https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow#:~:text=The%20on,permissions%20through%20the%20request%20chain) Microsoft identity platform and OAuth2.0 On-Behalf-Of flow - Microsoft identity platform | Microsoft Learn

<https://learn.microsoft.com/en-us/entra/identity-platform/v2-oauth2-on-behalf-of-flow>

[[4]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=%23%20Do%3A%20Integrate%20with%20single,on) [[6]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Anatomy) [[11]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=Set%20up%20a%20message%20extension) [[12]](https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design#:~:text=,Teams%20color%20tokens) Designing your message extension - Teams | Microsoft Learn

<https://learn.microsoft.com/en-us/microsoftteams/platform/messaging-extensions/design/messaging-extension-design>

[[5]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation) [[7]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=The%20Translator%20has%20a%20maximum,characteristics%20on%20sustained%20request%20rate) [[8]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=AI%20Translator%20Service%20character%20and,for%20text%20and%20document%20translation) [[9]](https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits#:~:text=Character%20limits%20per%20hour) Service limits - Translator Service - Azure AI services | Microsoft Learn

<https://learn.microsoft.com/en-us/azure/ai-services/translator/service-limits>