

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务文档

了解如何将 Azure OpenAI 的强大语言模型（包括 GPT-3、Codex 和 Embeddings 模型系列）用于内容生成、摘要、语义搜索和自然语言到代码的转换。



概述
什么是 Azure OpenAI 服务？



快速入门
快速入门



操作指南
创建资源



教程
嵌入



操作指南
完成



培训
Azure OpenAI 培训简介



概念
Azure OpenAI 模型



参考
支持和帮助选项

其他资源

Azure OpenAI

[Azure OpenAI Studio ↗](#)

[区域支持 ↗](#)

视频

将 OpenAI 模型与 Azure 的强大功能相结合

配额和限制

[申请对 Azure OpenAI 的访问权限 ↗](#)

参考

[REST API](#)

[使用条款 ↗](#)

工具

[Azure CLI](#)

[PowerShell](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

什么是 Azure OpenAI 服务？

项目 • 2023/03/06

Azure OpenAI 服务允许通过 REST API 访问 OpenAI 的强大语言模型，包括 GPT-3、Codex 和 Embeddings 模型系列。这些模型可以轻松适应特定的任务，包括但不限于内容生成、汇总、语义搜索和自然语言到代码的转换。用户可以在 Azure OpenAI Studio 中通过 REST API、Python SDK 或基于 Web 的界面访问该服务。

功能概述

功能	Azure OpenAI
可用的模型	GPT-3 基本系列 Codex 系列 Embeddings 系列 在 模型 页中了解详细信息。
微调	Ada Babbage Curie Cushman* Davinci* * 当前不可用。 **在美国东部区域，微调目前对新客户不可用。 请使用美国中南部区域进行位于美国的训练
价格	此处提供
虚拟网络支持和专用链接支持	是
托管标识	是，通过 Azure Active Directory
UI 体验	用于帐户和资源管理的 Azure 门户 用于模型探索和微调的 Azure OpenAI Service Studio
区域可用性	美国东部 美国中南部 西欧
内容筛选	使用自动化系统根据内容策略评估提示和完成情况。 将筛选高严重性内容。

负责任的 AI

Microsoft 致力于遵照“以人为本”的原则推动 AI 的进步。生成性模型（例如 Azure OpenAI 中提供的模型）提供显著的潜在优势，但如果未经精心设计和采用全方位的缓解措施，此类模型有可能会生成错误甚至有害的内容。Microsoft 已做出大量投资来帮助防范滥用和意外损害，其中包括要求申请人展示妥善定义的用例、融入 Microsoft 的[负责任 AI 使用原则](#)、生成内容筛选器以支持客户，并向已加入的新客户提供负责任 AI 实施指导。

如何访问 Azure OpenAI？

如何访问 Azure OpenAI？

由于需要应对很高的需求、即将推出的产品改进以及履行 [Microsoft 对负责任 AI 做出的承诺](#)，我们目前会限制访问。当前，我们正在与已经同 Microsoft 建立了合作关系的客户、用例风险较低的客户以及承诺融入缓解措施的客户合作。除了申请初始访问权限外，所有使用 Azure OpenAI 的解决方案都需要经历用例审查，然后才能发布用于生产用途。

申请表单中包含了更具体的信息。感谢你们的耐心，我们将努力以负责的态度实现 Azure OpenAI 的更多样化访问方式。

在此处申请初始访问权限或生产审查：

[“立即应用”](#)

所有使用 Azure OpenAI 的解决方案在发布用于生产用途之前都需要经历用例审查，并按用例接受评估。一般而言，提交审批的方案的敏感程度越高，风险缓解措施就越重要。

比较 Azure OpenAI 和 OpenAI

Azure OpenAI 服务通过 OpenAI GPT-3、Codex 和 DALL-E 模型为客户提供高级语言 AI，并能够实现 Azure 的安全性和企业前景。Azure OpenAI 与 OpenAI 共同开发 API，确保兼容性的同时能够实现二者之间的平稳过渡。

使用 Azure OpenAI，客户可在运行与 OpenAI 相同的模型时获得 Microsoft Azure 的安全功能。Azure OpenAI 提供专用网络、区域可用性和负责任 AI 内容筛选功能。

关键概念

提示和补全

补全终结点是 API 服务的核心组件。此 API 提供对模型的文本输入、文本输出接口的访问。用户只需提供一个包含英文文本命令的输入提示，模型就会生成文本补全。

下面是一个简单的提示和补全的示例：

提示：`""" count to 5 in a for loop """`

补全：`for i in range(1, 6): print(i)`

令牌

Azure OpenAI 通过将文本分解为标记来处理文本。标记可以是单词，也可以是字符块。例如，单词“hamburger”将分解为标记“ham”、“bur”和“ger”，而“pear”之类的常见短单词只是一个单个标记。许多标记以空格开头，例如“hello”和“bye”。

给定请求中处理的标记总数取决于输入、输出和请求参数的长度。处理的标记数量也会影响模型的响应延迟和吞吐量。

资源

Azure OpenAI 是 Azure 上的一个新产品。可以像在 Azure 订阅中使用任何其他可用于[创建资源](#)或服务实例的 Azure 产品一样开始使用 Azure OpenAI。可以阅读有关 Azure 的[资源管理设计](#)的详细信息。

部署

创建 Azure OpenAI 资源后，必须先部署模型，然后才能开始发出 API 调用和生成文本。可以使用部署 API 来完成此操作。这些 API 允许指定要使用的模型。

上下文学习

Azure OpenAI 使用的模型使用生成调用期间提供的自然语言指令和示例来识别请求的任务和所需的技能。使用此方法时，提示的第一个部分包括自然语言指令和/或所需特定任务的示例。然后，模型通过预测概率最高的下一段文本来完成任务。这种技术称为“上下文”学习。在此步骤中不会重新训练这些模型，而是根据你在提示中包含的上下文做出预测。

上下文学习有三种主要方法：少样本学习、单样本学习和零样本学习。这些方法根据提供给模型的任务特定数据量而异：

少样本学习：在这种情况下，用户在调用提示中包含几个示例来演示预期的答案格式和内容。以下示例显示了几个提示，我们在其中提供了多个示例（模型将生成最后一个答

案) :

Convert the questions to a command:

Q: Ask Constance if we need some bread.

A: send-msg `find constance` Do we need some bread?

Q: Send a message to Greg to figure out if things are ready for Wednesday.

A: send-msg `find greg` Is everything ready for Wednesday?

Q: Ask Ilya if we're still having our meeting this evening.

A: send-msg `find ilya` Are we still having a meeting this evening?

Q: Contact the ski store and figure out if I can get my skis fixed before I leave on Thursday.

A: send-msg `find ski store` Would it be possible to get my skis fixed before I leave on Thursday?

Q: Thank Nicolas for lunch.

A: send-msg `find nicolas` Thank you for lunch!

Q: Tell Constance that I won't be home before 19:30 tonight – unmovable meeting.

A: send-msg `find constance` I won't be home before 19:30 tonight. I have a meeting I can't move.

Q: Tell John that I need to book an appointment at 10:30.

A:

示例数量通常为 0 到 100 个，具体取决于单个提示的最大输入长度可以容纳多少个示例。最大输入长度可能因使用的特定模型而异。少样本学习可以大大减少准确进行预测所需的任务特定数据量。此方法的准确度通常不如微调的模型。

单样本学习：这种情况与少样本学习方法相同，不过只提供了一个示例。

零样本学习：在这种情况下，未向模型提供任何示例，而只提供了任务请求。

模型

该服务为用户提供对多种不同模型的访问。每种模型提供不同的功能和价位。GPT-3 基础模型按功能降序和速度升序顺序分别称为 DaVinci、Curie、Babbage 和 Ada。

Codex 系列模型是 GPT-3 的后代，并且已基于自然语言和代码进行训练，可为自然语言到代码用例提供支持。在[模型概念页](#)上详细了解每个模型。

后续步骤

详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务配额和限制

项目 · 2023/03/01

本文将介绍一个快速参考，并详细说明了 Azure 认知服务中 Azure OpenAI 的配额和限制。

配额和限制参考

以下部分提供了适用于 Azure OpenAI 配额和限制的快速指南：

限制名称	限制值
每个区域的 OpenAI 资源	2
每个模型每分钟的请求数*	Davinci 模型 (002 及更高版本) : 120 所有其他模型 : 300
每个模型每分钟的标记数*	Davinci 模型 (002 及更高版本) : 40,000 所有其他模型 : 120,000
最大微调模型部署*	2
能够将同一模型部署到多个部署	不允许
每个资源的训练作业总数	100
每个资源同时运行训练作业的最大数目	1
排队的最大训练作业数	20
每个资源的最大文件数	50
每个资源的所有文件的总大小	1 GB
最大训练作业时间 (如果超过，作业将失败)	120 小时
最大训练作业大小 (训练文件中的标记数) × (时期数)	Ada : 40-M 令牌 Babbage : 40-M 令牌 Curie : 40-M 令牌 Cushman : 40-M 令牌 Davinci : 10-M

*限制随时会变化。 我们预计，随着生产的进行和解决方案的扩展，你将需要更高的限制。 了解解决方案要求后，请在此处申请增加配额，联系我们：
<https://aka.ms/oai/quotaincrease>

在自动缩放期间缓解限制的常规最佳做法

若要最大程度地减少与限制相关的问题，可以使用以下方法：

- 在应用程序中实现重试逻辑
- 避免工作负载的急剧变化。 逐步增大工作负载。
- 测试不同负载增加模式。
- 在相同或不同的区域中创建其他 OpenAI 服务资源，并在区域间分配工作负载。

下一部分介绍调整配额的特定案例。

如何请求增加每分钟的事务数、部署的微调模型数或每分钟标记配额。

如果需要提高限制，可以在此处申请增加配额：<https://aka.ms/oai/quotaincrease>

后续步骤

详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

其他资源

📖 文档

[Azure OpenAI 内容筛选 - Azure OpenAI](#)

了解 Azure 认知服务中 OpenAI 服务的内容筛选功能

[Azure OpenAI 服务静态数据加密 - Azure Cognitive Services](#)

了解 Azure OpenAI 在数据持久化到云中时如何加密数据。

[Azure OpenAI 服务中的新增功能有哪些？ - Azure Cognitive Services](#)

了解 Azure OpenAI 的最新资讯和功能更新

[如何使用 Azure OpenAI 生成文本 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成或操作文本，包括代码

[Azure OpenAI 服务嵌入教程 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 嵌入 API 对 BillSum 数据集进行文档搜索

如何使用 Azure OpenAI 服务自定义模型 - Azure OpenAI

了解如何使用 Azure OpenAI 创建自己的自定义模型

Azure OpenAI 模型 - Azure OpenAI

了解 Azure OpenAI 中提供的不同模型。

[显示另外 4 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务模型

项目 • 2023/03/10

通过 Azure OpenAI 可使用很多不同模型，这些模型按系列和功能分组。模型系列通常按其预期任务关联模型。下表介绍了 Azure OpenAI 中当前可用的模型系列。目前并非所有模型都可在所有区域中使用。若需查看完整的细分情况，请参阅本文中的[模型功能表](#)。

模型系列	说明
GPT-3	可以理解和生成自然语言的模型系列。
Codex	可以理解和生成代码（包括将自然语言翻译为代码）的模型系列。
嵌入	一组可以理解和使用嵌入的模型。嵌入是一种特殊的数据表示格式，可由机器学习模型和算法轻松使用。嵌入是一段文本的语义含义的信息密集表示。目前，我们提供了三个系列的嵌入模型以实现不同的功能：相似性、文本搜索和代码搜索。

模型功能

每个模型系列都有一系列模型，这些模型按功能进一步区分。这些功能通常由名称标识，并且这些名称的字母顺序通常指示给定模型系列中该模型的相对功能和成本。例如，GPT-3 模型使用 Ada、Babbage、Curie 和 Davinci 等名称来指示相对功能和成本。Davinci 比 Curie 功能更强大（且成本更高），而 Curie 又比 Babbage 功能更强大（且成本更高），依此类推。

① 备注

功能较低的模型（如 Ada）可执行的任何任务都可以由功能较高的模型（如 Curie 或 Davinci）执行。

命名约定

Azure OpenAI 的模型名称通常对应于以下标准命名约定：

{capability}-{family}[-{input-type}]-{identifier}

元素	说明
{capability}	模型的模型功能。例如，GPT-3 模型使用 <code>text</code> ，而 Codex 模型使用 <code>code</code> 。
{family}	模型的相关系列。例如，GPT-3 模型包括 <code>ada</code> 、 <code>babbage</code> 、 <code>curie</code> 和 <code>davinci</code> 。
{input-type}	(仅限 嵌入模型) 模型支持的嵌入的输入类型。例如，文本搜索嵌入模型支持 <code>doc</code> 和 <code>query</code> 。
{identifier}	模型的版本标识符。

例如，我们最强大的 GPT-3 模型称为 `text-davinci-003`，而我们最强大的 Codex 模型称为 `code-davinci-002`。

名为 `ada`、`babbage`、`curie` 和 `davinci` 的旧版 GPT-3 模型不遵循标准命名约定，它们主要用于微调。有关详细信息，请参阅[了解如何为应用程序自定义模型](#)。

查找可用的模型

可以使用[模型列表 API](#) 获取 Azure OpenAI 资源可用于推理和微调的模型列表。

查找适当的模型

建议从模型系列中功能最强大的模型开始，以确认模型功能是否满足你的要求。然后可以继续使用该模型，也可以迁移到功能和成本较低的模型，围绕此模型的功能进行优化。

GPT-3 模型

GPT-3 模型可以理解和生成自然语言。该服务提供四个模型功能，每个都有不同级别的能力以及适用于不同任务的速度。Davinci 是功能最强大的模型，而 Ada 是速度最快的模型。模型排序（按功能从高到低的顺序）：

- `text-davinci-003`
- `text-curie-001`
- `text-babbage-001`
- `text-ada-001`

虽然 Davinci 能力最强，但其他模型提供了显着的速度优势。我们的建议是让用户在试验时从 Davinci 开始，因为它能产生最佳结果并验证 Azure OpenAI 可以提供的价值。原型正常工作后，就可以优化模型选择，为应用程序实现最佳延迟/性能平衡。

Davinci

Davinci 是功能最强大的模型，可以执行其他模型能够执行的任何任务，并且所用的指令通常更少。对于需要深入理解内容的应用程序（例如面向特定受众的摘要和创意内容的生成），Davinci 将产生最佳结果。Davinci 提供的这些增加的功能需要更多计算资源，因此 Davinci 的成本更高，并且 Davinci 不如其他模型快。

Davinci 擅长的另一个领域是理解文本的意图。Davinci 擅长解决多种逻辑问题并解释字符串动机。Davinci 已经能够解决一些涉及因果关系的最具挑战性的 AI 问题。

用途：复杂的意图、因果关系、受众摘要

Curie

Curie 功能强大，但速度很快。虽然 Davinci 在分析复杂文本方面更强大，但 Curie 可以执行许多精细化的任务，例如情绪分类和摘要。Curie 也善于回答问题和执行问答，适合用作常规服务聊天机器人。

用途：语言翻译、复杂分类、文本情绪、摘要

Babbage

Babbage 可以执行简单的分类等简单任务。在语义搜索方面，它的功能也很强大，可对文档与搜索查询的匹配程度进行排名。

用途：中等分类、语义搜索分类

Ada

Ada 通常是最快的模型，可以执行的任务有分析文本、地址更正和不需要太多细微差别的某些分类任务等等。Ada 的性能通常可以通过提供更多上下文来改进。

用途：分析文本、简单分类、地址更正、关键字

Codex 模型

Codex 模型是基模型 GPT-3 的子代，可以理解和生成代码。它们的训练数据包含自然语言和来自 GitHub 的数十亿行公开代码。

它们最擅长 Python，并且精通十几种语言，包括 C#、JavaScript、Go、Perl、PHP、Ruby、Swift、TypeScript、SQL 和 Shell。Codex 模型排序（按功能从高到低的顺序）：

- code-davinci-002

- code-cushman-001

Davinci

类似于 GPT-3，Davinci 是功能最强大的 Codex 模型，可以执行其他模型能够执行的任何任务，并且所用的指令通常更少。对于需要深入了解内容的应用程序，Davinci 会生成最佳结果。更强的功能需要更多计算资源，因此 Davinci 的成本更高，并且不如其他模型快。

Cushman

Cushman 功能强大，但速度很快。虽然 Davinci 在分析复杂任务方面更强大，但 Cushman 是能够执行许多代码生成任务的模型。Cushman 通常也比 Davinci 运行速度更快、成本更低。

嵌入模型

目前，我们提供了三个系列的嵌入模型以实现不同的功能：

- [相似度](#)
- [文本搜索](#)
- [代码搜索](#)

每个系列都包含某一功能范围的模型。以下列表根据模型功能指示服务返回的数字向量长度：

- Ada : 1024 个维度
- Babbage : 2048 个维度
- Curie : 4096 个维度
- Davinci : 12288 个维度

Davinci 功能最强，但比其他模型更慢更贵。Ada 功能最弱，但速度更快且更成本更低。

相似性嵌入

此类模型擅长捕获两个或更多文本片段之间的语义相似性。

用例	模型
----	----

用例	模型
聚类分析、回归、异常情况检测、可视化	<code>text-similarity-ada-001</code> <code>text-similarity-babbage-001</code> <code>text-similarity-curie-001</code> <code>text-similarity-davinci-001</code>

文本搜索嵌入

此类模型有助于度量长文档是否与短搜索查询相关。此系列支持两种输入类型：`doc`（用于嵌入要检索的文档）和 `query`（用于嵌入搜索查询）。

用例	模型
搜索、上下文相关性、信息检索	<code>text-search-ada-doc-001</code> <code>text-search-ada-query-001</code> <code>text-search-babbage-doc-001</code> <code>text-search-babbage-query-001</code> <code>text-search-curie-doc-001</code> <code>text-search-curie-query-001</code> <code>text-search-davinci-doc-001</code> <code>text-search-davinci-query-001</code>

代码搜索嵌入

与文本搜索嵌入模型类似，此系列支持两种输入类型：`code`（用于嵌入要检索的代码片段）和 `text`（用于嵌入自然语言搜索查询）。

用例	模型
代码搜索和相关性	<code>code-search-ada-code-001</code> <code>code-search-ada-text-001</code> <code>code-search-babbage-code-001</code> <code>code-search-babbage-text-001</code>

使用嵌入模型时，请注意其限制和风险。

模型摘要表和区域可用性

GPT-3 模型

模型 ID	支持 补全	支持 嵌入	基本模型区域	微调区域
ada	是	否	空值	美国东部 ² 、美国中南部、欧洲西部
text-ada-001	是	否	美国东部 ² 、美国中南部、欧洲西部	空值
babbage	是	否	空值	美国东部 ² 、美国中南部、欧洲西部
text-babbage-001	是	否	美国东部 ² 、美国中南部、欧洲西部	空值
curie	是	否	空值	美国东部 ² 、美国中南部、欧洲西部
text-curie-001	是	否	美国东部 ² 、美国中南部、欧洲西部	空值
davinci ¹	是	否	空值	美国东部 ² 、美国中南部、欧洲西部
text-davinci-001	是	否	美国中南部、欧洲西部	空值
text-davinci-002	是	否	美国东部、美国中南部、欧洲西部	空值
text-davinci-003	是	否	美国东部	空值
text-davinci-fine-tune-002 ¹	是	否	空值	美国东部、西欧

¹ 仅能通过请求获取该模型。目前，我们不接受使用该模型的新请求。

² 由于需求高，美国东部区域的新用户目前无法进行微调。请使用美国中南部区域进行位于美国的训练。

Codex 模型

模型 ID	支持补 全	支持嵌 入	基本模型区域	微调区域
code-cushman-001 ¹	是	否	美国中南部、欧洲西部	美国东部 ² 、美国中南部、欧洲西部
code-davinci-002	是	否	美国东部、西欧	空值

模型 ID	支持补全	支持嵌入	基本模型区域	微调区域
code-davinci-fine-tune-002 ¹	是	否	空值	美国东部 ² 、西欧

¹ 仅能通过请求将该模型用于微调。目前，我们不接受微调该模型的新请求。

² 由于需求高，美国东部区域的新用户目前无法进行微调。请使用美国中南部区域进行位于美国的训练。

嵌入模型

模型 ID	支持补全	支持嵌入	基本模型区域	微调区域
text-ada-embeddings-002	否	是	美国东部、美国中南部、欧洲西部	空值
text-similarity-ada-001	否	是	美国东部、美国中南部、欧洲西部	空值
text-similarity-babbage-001	否	是	美国中南部、欧洲西部	空值
text-similarity-curie-001	否	是	美国东部、美国中南部、欧洲西部	空值
text-similarity-davinci-001	否	是	美国中南部、欧洲西部	空值
text-search-ada-doc-001	否	是	美国中南部、欧洲西部	空值
text-search-ada-query-001、	否	是	美国中南部、欧洲西部	空值
text-search-babbage-doc-001	否	是	美国中南部、欧洲西部	空值
text-search-babbage-query-001	否	是	美国中南部、欧洲西部	空值
text-search-curie-doc-001	否	是	美国中南部、欧洲西部	空值
text-search-curie-query-001	否	是	美国中南部、欧洲西部	空值
text-search-davinci-doc-001	否	是	美国中南部、欧洲西部	空值
text-search-davinci-query-001	否	是	美国中南部、欧洲西部	空值
code-search-ada-code-001	否	是	美国中南部、欧洲西部	空值
code-search-ada-text-001	否	是	美国中南部、欧洲西部	空值

模型 ID	支持补充 全	支持嵌入	基本模型区域	微调区域
code-search-babbage-code-001	否	是	美国中南部、欧洲西部	空值
code-search-babbage-text-001	否	是	美国中南部、欧洲西部	空值

后续步骤

[详细了解 Azure OpenAI](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务中的新增功能有哪些

项目 · 2023/02/24

2023 年 1 月

新功能

- 服务 GA。 Azure OpenAI 服务现已正式发布。
- 新模型 - 添加了最新的文本模型：text-davinci-003（美国东部、西欧）、text-ada-embeddings-002（美国东部、美国中南部、西欧）

2022 年 12 月

新增功能

- **OpenAI 中的最新模型。** Azure OpenAI 提供对所有最新模型（包括 GPT-3.5 系列）的访问权限。
- **新的 API 版本 (2022-12-01)。** 此更新包括几个请求的增强功能，其中包括 API 响应中的令牌使用情况信息、改进的文件错误消息、在微调创建数据结构上与 OpenAI 保持一致，以及对后缀参数的支持以允许微调作业的自定义命名。
- **每秒请求数的上限更高。** 对于非 Davinci 模型，限制为 50。对于 Davinci 模型，限制为 20。
- **微调部署速度更快。** 在 10 分钟内部署 Ada 和 Curie 微调模型。
- **训练限制值更高：**对于 Ada、Babbage 和 Curie，限制为 4,000 万个训练令牌。对于 Davinci，限制为 1,000 万。
- **请求修改滥用和误用数据日志记录和人工评审的过程。** 目前，该服务记录请求/响应数据，以便进行滥用和误用检测，确保这些功能强大的模型不会被滥用。但是，许多客户有严格的数据隐私和安全要求，需要对他们的数据进行更高级别的控制。为了支持这些用例，我们将发布一个新流程，供客户修改内容筛选策略或关闭低风

险用例的滥用日志记录。此过程遵循 Azure 认知服务中既定的受限访问流程，现有 OpenAI 客户可在此处应用[♂](#)。

- **客户管理的密钥 (CMK) 加密。** CMK 通过提供客户自己的用于存储训练数据和自定义模型的加密密钥，让客户能够更好地控制在 Azure OpenAI 服务中管理其数据。客户管理的密钥 (CMK)（也称为创建自己的密钥，BYOK）在创建、轮换、禁用和撤销访问控制方面可提供更大的灵活性。此外，你还可以审核用于保护数据的加密密钥。有关详细信息，请参阅[静态加密文档](#)。

- **密码箱支持**

- **SOC-2 符合性**

- 通过 Azure 资源运行状况、成本分析以及指标和诊断设置进行日志记录和诊断。
- **工作室改进。** 对工作室工作流进行了大量可用性改进，包括 Azure AD 角色支持，可以控制团队中的哪个成员有权创建微调的模型和部署。

变更（中断性）

微调创建 API 请求已更新，以匹配 OpenAI 的架构。

预览 API 版本：

JSON

```
{  
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",  
  "hyperparams": {  
    "batch_size": 4,  
    "learning_rate_multiplier": 0.1,  
    "n_epochs": 4,  
    "prompt_loss_weight": 0.1,  
  }  
}
```

API version 2022-12-01:

JSON

```
{  
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",  
  "batch_size": 4,  
  "learning_rate_multiplier": 0.1,  
  "n_epochs": 4,  
  "prompt_loss_weight": 0.1,  
}
```

默认情况下，内容筛选暂时处于关闭状态。 Azure 内容审核的工作方式与 OpenAI 不同。 Azure OpenAI 在生成调用期间运行内容筛选器，以检测有害或滥用的内容，并从响应中筛选它们。 [了解详细信息](#)

这些模型将在 2023 年第 1 季度重新启用，并默认开启。

客户操作

- 如果想要为订阅启用这些设置，请[联系 Azure 支持](#)。
- 如果想要让它们保持关闭状态，请[申请筛选修改](#)。 (此选项仅适用于低风险用例。)

后续步骤

详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务常见问题解答

常见问题解答

如果你在本文档中找不到问题的答案，因此仍然需要帮助，请查看[认知服务支持选项指南](#)。 Azure OpenAI 是 Azure 认知服务的一部分。

数据和隐私

你是否使用我的公司数据来训练这其中的任何模型？

Azure OpenAI 不使用客户数据来重新训练模型。有关详细信息，请参阅 [Azure OpenAI 数据、隐私和安全指南](#)。

常规

Azure OpenAI 的功能与 OpenAI 相比如何？

Azure OpenAI 服务通过 OpenAI GPT-3、Codex 和 DALL-E 模型为客户提供高级语言 AI，并能够实现 Azure 的安全性和企业前景。 Azure OpenAI 与 OpenAI 共同开发 API，确保兼容性的同时能够实现二者之间的平稳过渡。

使用 Azure OpenAI，客户可在运行与 OpenAI 相同的模型时获得 Microsoft Azure 的安全功能。

获取对 Azure OpenAI 服务的访问权限

如何访问 Azure OpenAI？

由于需要应对很高的需求、即将推出的产品改进以及履行 Microsoft 对负责任 AI 做出的承诺，我们目前会限制访问。当前，我们正在与已经同 Microsoft 建立了合作关系的客户、用例风险较低的客户以及承诺融入缓解措施的客户合作。除了申请初始访问权限

外，所有使用 Azure OpenAI 的解决方案都需要经历用例审查，然后才能发布用于生产用途。 请在此处申请初始访问权限或生产审查：[立即申请](#)

申请访问权限后，需要等待多长时间才能获得批准？

我们目前不提供访问审批的时间线。

了解详细信息以及在何处提问

在哪里可以了解 Azure OpenAI 的最新更新？

有关每月更新，请参阅[新增功能页面](#)。

我可以在哪里接受培训来开始学习并积累 Azure OpenAI 技能？

查看 [Azure OpenAI 培训课程简介](#)。

在哪里可以发布问题并查看其他常见问题的解答？

- 建议在 [Microsoft Q&A](#) 上发布问题
- 也可在 [Stack Overflow](#) 上发布问题

我在哪里可以获得 Azure OpenAI 客户支持？

Azure OpenAI 是 Azure 认知服务的一部分。 可以在[支持和帮助选项指南](#)中了解 Azure 认知服务的所有支持选项。

模型和微调

有哪些模型可用？

请参阅 [Azure OpenAI 模型可用性指南](#)。

在哪里可以找到模型在哪个区域可用的信息？

有关区域可用性，请参阅 [Azure OpenAI 模型可用性指南](#)。

基本模型和微调模型的区别是什么？

基本模型是尚未针对特定用例进行自定义或微调的模型。 微调模型是基本模型的自定义版本，其中模型的权重是根据一组独特的提示进行训练的。 微调模型可让你在更多任务上取得更好的结果，而无需在完成提示中提供详细的上下文学习示例。 若要了解详细信息，请查看我们的[微调指南](#)。

我最多可以创建多少个微调模型？

100

在 Azure OpenAI 中，API 响应的 SLA 是什么？

目前，我们还没有定义的 API 响应时间服务级别协议 (SLA)。 Azure OpenAI 服务的总体 SLA 与其他 Azure 认知服务的相同。 有关详细信息，请参阅["Online Services 的服务级别协议 \(SLA\)"](#)页的“认知服务”部分。

后续步骤

- [Azure OpenAI 配额和限制](#)
- [Azure OpenAI 新增功能](#)
- [Azure OpenAI 快速入门](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

快速入门：开始使用 Azure OpenAI 服务生成文本

项目 • 2023/03/01

使用本文开始对 Azure OpenAI 发出前几次调用。

先决条件

- Azure 订阅 - [免费创建订阅](#)。
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 的访问权限。
目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。
- 已部署模型的 Azure OpenAI 资源。有关模型部署的详细信息，请参阅[资源部署指南](#)。

转到 Azure OpenAI Studio

导航到 Azure OpenAI Studio (<https://oai.azure.com/>)，然后使用有权访问 OpenAI 资源的凭据登录。在登录过程中或登录之后，选择适当的目录、Azure 订阅和 Azure OpenAI 资源。

在 Azure OpenAI Studio 登陆页中，进一步浏览提示完成、管理部署和模型的示例，并查找文档和社区论坛等学习资源。

Get started with Azure OpenAI Service

Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.

Try the playground



Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.

[GPT-3 playground](#)

Explore examples for prompt completion



Summarize Text

Summarize text by adding a 'tl;dr:' to the end of a text passage.

[Learn more](#)



Classify Text

Classify items into categories provided at inference time.

[Learn more](#)



Natural Language to SQL

Translate natural language to SQL queries.

[Learn more](#)



Generate New Product Names

Create product names from example words.

[Learn more](#)

转到[操场](#)进行试验和工作流优化。

操场

通过 GPT-3 操场开始探索使用无代码方法的 OpenAI 功能。这是一个简单的文本框，可以在其中提交提示以生成补全内容。在此页中，可以快速循环访问和试验这些功能。

Cognitive Services | Azure OpenAI Studio

Azure OpenAI Studio > GPT-3 playground

Privacy & cookies

Parameters

Temperature: 1

Max length (tokens): 100

Stop sequences

Top probabilities: 0.5

Frequency penalty: 0

Presence penalty: 0

Best of: 1

Pre-response text

Post-response text

Learn more

Start typing here

Deployments: text-davinci-002

Examples: Load an example

View code

Generate Undo Regenerate Tokens: 0

可以选择一个部署，然后从一些预加载的示例中进行选择以开始使用。如果资源没有部署，请选择“**创建部署**”，然后按照向导提供的说明进行操作。有关模型部署的详细信

息，请参阅[资源部署指南](#)。

可以试验温度和预响应文本等配置设置，以提高任务的性能。可以在[REST API](#)中详细了解每个参数。

- 选择“**生成**”按钮后会将输入的文本发送到补全 API，并将结果流式传输回到文本框中。
- 选择“**撤消**”按钮可以撤消上一次生成调用。
- 选择“**重新生成**”按钮可以同时执行撤消和生成调用。

Azure OpenAI 还会对提示输入和生成的输出执行内容审核。检测到有害内容时，可能会筛选提示或响应。有关详细信息，请参阅[内容筛选器](#)一文。

在 GPT-3 操场中，还可以查看根据所选设置预先填充的 Python 和 curl 代码示例。只需选择示例下拉列表旁边的“**查看代码**”。可以编写应用程序以使用 OpenAI Python SDK、curl 或其他 REST API 客户端完成相同任务。

试用文本摘要

若要在 GPT-3 操场中使用 OpenAI 服务进行文本摘要，请执行以下步骤：

1. 登录到[Azure OpenAI Studio](#)。
2. 选择要使用的订阅和 OpenAI 资源。
3. 选择登陆页面顶部的“**GPT-3 操场**”。
4. 从“**部署**”下拉列表中选择你的部署。如果你的资源没有部署，请选择“**创建部署**”，然后重做此步骤。
5. 从“**示例**”下拉列表中选择“**汇总文本**”。

GPT-3 playground

Deployments

text-davinci-002

Examples

Summarize Text

 View code

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses.[3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

Tl;dr:

A neutron star is the collapsed core of a supergiant star. These incredibly dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

 Generate

 Undo

 Regenerate

Tokens: 189 



6. 选择 `Generate`。OpenAI 会掌握文本的上下文并简洁地对其进行重新编写。你应得到类似于以下文本的结果：

Tl;dr A neutron star is the collapsed core of a supergiant star. These incredibly dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

响应的准确性可能因模型而异。此示例中基于 Davinci 的模型非常适合这种类型的摘要，而基于 Codex 的模型对于此类任务的表现就没有那么好。

清理资源

如果你想要清理和删除 OpenAI 资源，可以删除资源或资源组。删除资源组同时也会删除与之相关联的任何其他资源。

- [Portal](#)
- [Azure CLI](#)

后续步骤

在[有关补全的操作指南](#)中详细了解如何生成最佳补全内容。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

内容筛选

项目 • 2023/03/06

Azure OpenAI 服务包括一个内容管理系统，它与核心模型一起使用以筛选内容。该系统通过一组旨在检测滥用的分类模型运行输入提示和生成的内容。如果系统识别出有害内容，且提示被视为不合适，你将在 API 调用中收到错误消息，或者响应中的 `finish_reason` 将为 `content_filter`，表示某些生成已经过筛选。可以使用许多不同的配置通过完成 API 生成内容，这些配置将更改预期筛选行为。以下部分旨在列举所有这些方案，以便你能够适当地设计解决方案。

为确保已适当降低应用程序中的风险，应仔细评估所有潜在危害，遵循[透明度说明](#)中的指导，并根据需要添加针对特定场景的缓解措施。

方案详细信息

在生成应用程序时，需要考虑已筛选由完成 API 返回的内容和该内容可能不完整的情况。如何处理这些信息将取决于具体的应用程序。该行为可以概括为以下关键点：

- 被视为不恰当的提示将返回 HTTP 400 错误
- 筛选内容时，非流式传输完成调用不会返回任何内容。`finish_reason` 值将设置为 `content_filter`。在极少数情况下，响应较长，可能会返回部分结果。在这些情况下，将更新 `finish_reason`。
- 对于流式传输完成调用，段将在完成时返回给用户。该服务将继续流式传输，直到检测到达到停止标记、长度或有害内容。

方案：发送一个非流式传输完成调用，要求多个生成没有不恰当内容

下表概述了内容筛选可能出现的各种方式：

HTTP 响应代码	响应行为
200	在所有生成都通过筛选器模型的情况下，不会将内容审核详细信息添加到响应中。每个生成的 <code>finish_reason</code> 都将是 <code>stop</code> 或 <code>length</code> 。

示例请求有效负载：

JSON

```
{  
  "prompt": "Text example",  
  "n": 3,  
  "stream": false  
}
```

响应 JSON 示例：

JSON

```
{  
  "id": "example-id",  
  "object": "text_completion",  
  "created": 1653666286,  
  "model": "davinci",  
  "choices": [  
    {  
      "text": "Response generated text",  
      "index": 0,  
      "finish_reason": "stop",  
      "logprobs": null  
    }  
  ]  
}
```

场景：API 调用要求具有多个响应 (N>1) 并至少筛选 1 个响应

HTTP 响应代码

响应行为

200

被筛选的生成将具有 `finish_reason` 值“content_filter”。

示例请求有效负载：

JSON

```
{  
  "prompt": "Text example",  
  "n": 3,  
  "stream": false  
}
```

响应 JSON 示例：

JSON

```
{  
  "id": "example",  
  "object": "text_completion",  
  "created": 1653666831,  
  "model": "ada",  
  "choices": [  
    {  
      "text": "returned text 1",  
      "index": 0,  
      "finish_reason": "length",  
      "logprobs": null  
    },  
    {  
      "text": "returned text 2",  
      "index": 1,  
      "finish_reason": "content_filter",  
      "logprobs": null  
    }  
  ]  
}
```

场景：将不恰当的输入提示发送到完成 API (用于流式传输或非流式传输)

HTTP 响应代码 响应行为

400

当提示触发内容策略模型之一时，API 调用将失败。修改提示，然后重试。

示例请求有效负载：

JSON

```
{  
  "prompt": "Content that triggered the filtering model"  
}
```

响应 JSON 示例：

JSON

```
"error": {  
  "message": "The response was filtered",  
  "type": null,  
  "param": "prompt",  
}
```

```
        "code": "content_filter",
        "status": 400
    }
```

场景：你需要执行一个流式传输完成调用，并且所有生成的内容都通过内容筛选器

HTTP 响应行为
应代码

200 在这种情况下，调用将以流式传输的方式返回整个生成，并且对于每个生成的响应，
finish_reason 将是“length”或“stop”。

示例请求有效负载：

JSON

```
{
    "prompt": "Text example",
    "n": 3,
    "stream": true
}
```

响应 JSON 示例：

JSON

```
{
    "id": "cmpl-example",
    "object": "text_completion",
    "created": 1653670914,
    "model": "ada",
    "choices": [
        {
            "text": "last part of generation",
            "index": 2,
            "finish_reason": "stop",
            "logprobs": null
        }
    ]
}
```

场景：你需要执行一个流式传输完成调用，要求生成多个响应，并且至少筛选一个响应

HTTP 响应行为
响应代码

200 对于给定的生成索引，生成的最后一个区块将包含一个非空 `finish_reason` 值。筛选生成时，该值将为“`content_filter`”。

示例请求有效负载：

JSON

```
{  
  "prompt": "Text example",  
  "n": 3,  
  "stream": true  
}
```

响应 JSON 示例：

JSON

```
{  
  "id": "cmpl-example",  
  "object": "text_completion",  
  "created": 1653670515,  
  "model": "ada",  
  "choices": [  
    {  
      "text": "Last part of generated text streamed back",  
      "index": 2,  
      "finish_reason": "content_filter",  
      "logprobs": null  
    }  
  ]  
}
```

场景：内容筛选系统无法在生成上运行

HTTP 响应行为
响应代码

200 如果内容筛选系统出现故障或无法及时完成操作，请求仍将完成。可通过在“`content_filter_result`”对象中查找错误消息来确定是否应用了筛选。

示例请求有效负载：

JSON

```
{  
  "prompt": "Text example",  
  "n": 1,  
  "stream": false  
}
```

响应 JSON 示例：

JSON

```
{  
  "id": "cmpl-example",  
  "object": "text_completion",  
  "created": 1652294703,  
  "model": "ada",  
  "choices": [  
    {  
      "text": "generated text",  
      "index": 0,  
      "finish_reason": "length",  
      "logprobs": null,  
      "content_filter_result": {  
        "error": {  
          "code": "content_filter_error",  
          "message": "The contents are not filtered"  
        }  
      }  
    }  
  ]  
}
```

最佳实践

在应用程序设计过程中，需要仔细考虑如何最大限度地发挥应用程序的优势，同时最大限度地减少危害。请考虑采用以下最佳做法：

- 你希望如何处理用户发送不当或误用应用程序的情况。检查 `finish_reason` 以查看是否筛选了生成。
- 如果内容筛选器对生成运行至关重要，请检查 `content_filter_result` 中是否无 `error` 对象。
- 为帮助监视可能的滥用情况，服务于多个最终用户的应用程序应在每次 API 调用中传递 `user` 参数。`user` 应是最终用户的唯一标识符。不要发送任何实际的用户可识别信息作为值。

后续步骤

详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

其他资源

文档

[如何使用 Azure OpenAI 服务自定义模型 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 创建自己的自定义模型

[如何使用 Azure OpenAI 服务生成嵌入 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成嵌入

[监视 Azure OpenAI 服务 - Azure Cognitive Services](#)

从此处开始了解如何监视 Azure OpenAI 服务

[计划管理 Azure OpenAI 的成本 - Azure Cognitive Services](#)

了解如何使用 Azure 门户中的成本分析来计划和管理 Azure OpenAI 服务的成本。

[Azure OpenAI 服务静态数据加密 - Azure Cognitive Services](#)

了解 Azure OpenAI 在数据持久化到云中时如何加密数据。

[如何准备用于自定义模型训练的数据集 - Azure OpenAI Service](#)

了解如何准备用于微调的数据集

[如何使用 Azure OpenAI 生成文本 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成或操作文本，包括代码

[Azure OpenAI 服务嵌入 - Azure OpenAI - embeddings and cosine similarity](#)

详细了解用于执行文档搜索以及获取余弦相似性的 Azure OpenAI 嵌入 API

[显示另外 5 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

了解 Azure OpenAI 服务中的嵌入

项目 · 2023/03/09

嵌入是一种特殊的数据表示格式，可由机器学习模型和算法轻松使用。嵌入是一段文本的语义含义的信息密集表示。每个嵌入是浮点数的一个向量，向量空间中两个嵌入之间的距离与原始格式的两个输入之间的语义相似性相关。例如，如果两个文本相似，则它们的向量表示形式也应该相似。

嵌入模型

不同的 Azure OpenAI 嵌入模型专用于特定的任务。“相似性嵌入”擅长捕获两个或更多文本片段之间的语义相似性。“文本搜索嵌入”有助于度量较长的文档是否与简短的查询相关。“代码搜索嵌入”可用于嵌入代码片段和嵌入自然语言搜索查询。

嵌入捕获向量空间中的语义相似性，从而能更轻松地对表示字词的大型输入进行机器学习。因此我们可以使用嵌入来确定两个文本区块在语义上是否相关或相似，并提供一个分数来评估相似性。

余弦相似性

要识别相似的文档，有一种方法是计算文档之间相同的单词数。可惜这种方法是无法处理大规模文档的，因为在较大规模的文档中通常会检测到大量相同的单词，甚至对于话题截然不同的文档也是如此。因此，余弦相似性可以提供更有效的替代方案。

从数学角度来看，余弦相似性测量投射至多维空间中的两个向量之间的角度的余弦值。这是有帮助的，因为如果两个文档因为大小而在欧氏距离上相距甚远，它们之间也还能存在一个较小的角度，让我们能得出它们具有较高的余弦相似性。

Azure OpenAI 嵌入依赖于余弦相似性来计算文档和查询之间的相似性。

后续步骤

请通过我们的[嵌入教程](#)，详细了解如何使用 Azure OpenAI 和嵌入执行文档搜索。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

使用 Azure OpenAI 创建资源和部署模型

项目 · 2023/03/09

参考本文中的分步说明开始使用 Azure OpenAI 来创建资源和部署模型。虽然资源创建和模型部署的步骤可以在几分钟内完成，但实际的部署过程本身可能需要一小时以上。可以先创建资源，开始部署，稍后再回头检查部署，而不必等待部署完成。

先决条件

- Azure 订阅 - [免费创建订阅](#)
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 的访问权限

目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。

创建资源

可通过多种不同的方式创建 Azure 中的资源：

- 在 [Azure 门户](#) 中
- 使用 REST API、Azure CLI、PowerShell 或客户端库
- 通过 ARM 模板

本指南将引导你完成 Azure 门户创建体验。

1. 导航到创建页：[Azure OpenAI 服务创建页](#)

2. 在“创建”页中提供以下信息：

字段	说明
订阅	选择 OpenAI 加入应用程序中使用的 Azure 订阅
资源组	包含 OpenAI 资源的 Azure 资源组。可以创建新组或将其添加到预先存在的组。
区域	实例的位置。不同位置可能会导致延迟，但不会影响资源的运行时可用性。

字段	说明
名称	认知服务资源的描述性名称。例如 MyOpenAIResource。
定价层	目前只有 1 个定价层可用于该服务

Create Azure OpenAI ...

[Basics](#) [Tags](#) [Review + create](#)

Enable new business solutions with OpenAI's language generation capabilities powered by GPT-3 models. These models have been pretrained with trillions of words and can easily adapt to your scenario with a few short examples provided at inference. Apply them to numerous scenarios, from summarization to content and code generation.

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ [OpenAI Test Subscription](#) 

Resource group * ⓘ [test-resource-group](#)  [Create new](#)

Instance details

Region * ⓘ [South Central US](#) 

Name * ⓘ [azure-openai-test-001](#) 

Pricing tier * ⓘ [Standard S0](#) 

[View full pricing details](#)

[Review + create](#) [< Previous](#) [Next : Tags >](#) 

部署模型

在生成文本或推理之前，需要先部署一个模型。可以从 Azure OpenAI Studio 中的多个可用模型中进行选择。

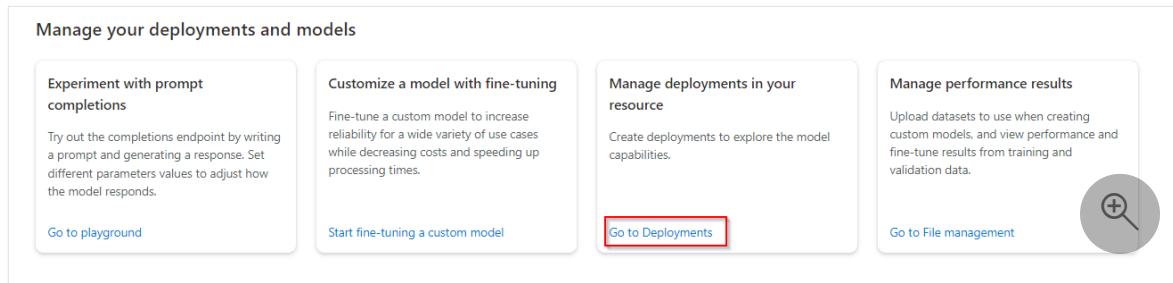
Davinci 是功能最强大的模型系列，可以执行其他模型能够执行的任何任务，并且使用的指令通常较少。对于需要对内容有丰富的理解的应用程序（例如面向特定受众的摘要和创意内容的生成），Davinci 能够产生最佳结果。

若要部署模型，请执行以下步骤：

1. 登录到 [Azure OpenAI Studio](#)。

2. 选择要使用的订阅和 OpenAI 资源。

3. 选择“**管理部署和模型**”下的“**管理资源中的部署**”>“**转到部署**”。你可能首先需要在登陆页上向下滚动。



4. 从“**管理**”>“**部署**”页中选择“**新建部署**”。

5. 从下拉列表中选择一个模型。若要在美国东部区域开始使用，建议使用 `text-davinci-003` 模型。如果是其他区域，应从 `text-davinci-002` 模型开始。某些模型并非在所有区域中都可用。有关每个区域的可用模型的列表，请参阅[模型摘要表和区域可用性](#)。

6. 输入模型名称以帮助你识别模型。请谨慎选择名称。模型名称将通过 OpenAI 客户端库和 API 用作部署名称。

7. 选择“**创建**”以部署模型。

部署表会显示与新建模型相对应的新条目。部署完成并可供使用时，部署状态将变为“成功”。

后续步骤

- 现已创建一个资源并已部署第一个模型，接下来请开始使用我们的[快速入门](#)发出 API 调用并生成文本。
- 详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

了解如何生成或操作文本

项目 • 2023/03/01

补全终结点可用于各种任务。它为我们的所有[模型](#)提供一个简单而强大的文本输入和文本输出接口。输入一些文本作为提示，模型将生成文本补全内容，试图匹配你提供给它的任何上下文或模式。例如，如果你给 API 这样的提示“正如笛卡尔所说，我思故”，它很有可能返回补全内容“我在”。

若要开始探索补全功能，最佳方法是使用 [Azure OpenAI Studio](#) 中的操场。这是一个简单的文本框，你可以在其中提交提示以生成补全内容。你可以从如下所示的简单示例入手：

```
write a tagline for an ice cream shop
```

提交后，将看到如下所示的生成内容：

控制台

```
write a tagline for an ice cream shop
we serve up smiles with every scoop!
```

你看到的实际完成结果可能会有所不同，因为默认情况下 API 是随机的。换句话说，每次调用它时，你都可能会得到略有不同的完成，即使提示保持不变。可以使用温度设置控制此行为。

这个简单的“文本输入、文本输出”接口意味着，你可以提供说明或仅提供几个想要的操作示例来给模型编程。是否成功通常取决于任务的复杂度和提示的质量。一般规则是考虑如何写出一道能让中学生解决的应用题。写得好的提示可提供足够的信息，便于模型了解你想要什么以及它应该如何回复。

① 备注

请记住，模型的训练数据在 2019 年 10 月中断，因此它们可能不了解当前事件。我们计划在今后添加更多持续的训练。

提示设计

基础

从生成原创故事到执行复杂文本分析，OpenAI 的模型可执行各种操作。由于它们能做的事情太多，因此必须明确表示你需要什么。好提示的秘诀在于“展示”而不仅仅是“说明”。

模型尝试根据提示预测你的需要。如果你发送“给我一份猫的品种列表”这句话，模型不会自动认为你需要猫的品种列表。你可以直接让模型继续对话，第一句话是“给我一份猫的品种列表”，第二句话是“我会告诉你我喜欢哪些”。如果模型只认为你想要猫的列表，那它在内容创建、分类或其他任务方面就不那么擅长了。

创建提示有三个基本准则：

展示并说明。 通过提供说明和/或示例，明确表示你想要什么。如果想让模型按字母顺序对项目列表进行排序，或者按情绪对一段话进行分类，请向它展示这就是你的需求。

提供优质数据。 如果尝试生成分类器或让模型遵循某种模式，请确保提供足够的示例。请务必校对示例 - 模型通常很聪明，足以发现基本的拼写错误并提供回复，但它也可能认为你是故意拼错，继而可能会影响到回复的内容。

检查设置。 温度和 top_p 设置控制模型生成回复时的确定程度。如果只有一个正确答案，你要让它回复，就需要将这些值设置得低一些。如果你希望得到含糊的回复，则最好将值设置得高一些。人们使用这些设置的第一大误区是认为它们是“聪明”或“创意”控件。

故障排除

如果在让 API 按预期执行时遇到问题，请按照以下清单内容进行排查：

1. 它是否清楚预期生成的内容应该是什么？
2. 是否有足够的示例？
3. 是否检查了示例中有无错误？（API 不会直接告诉你）
4. 是否正确使用 temp 和 top_p？

分类

为了使用 API 创建文本分类器，我们提供了任务说明和一些示例。在此演示中，我们将展示 API 如何对推文情绪进行分类。

控制台

```
This is a tweet sentiment classifier
```

```
Tweet: "I loved the new Batman movie!"
```

```
Sentiment: Positive
```

```
Tweet: "I hate it when my phone battery dies."
```

```
Sentiment: Negative
```

```
Tweet: "My day has been 🤗"
```

```
Sentiment: Positive
```

```
Tweet: "This is the link to the article"
```

```
Sentiment: Neutral
```

```
Tweet: "This new music video blew my mind"
```

```
Sentiment:
```

在此示例中，有几个值得注意的要点：

1. 使用简明的语言来描述你的输入和输出：我们使用简单明了的语言描述输入“推文”和预期输出“情绪”。最佳做法是从简明的语言说明开始。 虽然通常可以使用简写或关键词来指示输入和输出，但在生成提示时，最好先尽可能描述一下，只要提示的表现一样，就可以反向删除多余的单词。

2. 向 API 展示如何回应各种情况：在本例中，我们提供多种结果：“积极”、“消极”和“中立”。中立结果很重要，因为在许多情况下，即使是人类也会很难确定某件事是积极还是消极的，还有一些情况两种都不是。

3. 可以使用文本和表情符号：分类器是文本和表情符号 🤗 的组合。 API 可以读取表情符号，甚至可以将它们与表达式进行转换。

4. 对于熟悉的任务，需要的示例更少：对于此分类器，我们仅提供少量示例。 这是因为 API 已经理解情绪和推文的概念。 如果要为 API 不熟悉的内容生成分类器，可能需要提供更多示例。

提高分类器的效率

至此，我们已经掌握了如何生成分类器，接下来，我们继续看这个示例，让分类器变得更加高效，这样就可以使用它从一个 API 调用中获取多个结果。

```
This is a tweet sentiment classifier
```

```
Tweet: "I loved the new Batman movie!"
```

```
Sentiment: Positive
```

```
Tweet: "I hate it when my phone battery dies"
```

```
Sentiment: Negative
```

```
Tweet: "My day has been 🤗"
```

```
Sentiment: Positive
```

```
Tweet: "This is the link to the article"
```

```
Sentiment: Neutral
```

Tweet text

1. "I loved the new Batman movie!"
2. "I hate it when my phone battery dies"
3. "My day has been "
4. "This is the link to the article"
5. "This new music video blew my mind"

Tweet sentiment ratings:

1: Positive

2: Negative

3: Positive

4: Neutral

5: Positive

Tweet text

1. "I can't stand homework"
2. "This sucks. I'm bored "
3. "I can't wait for Halloween!!!"
4. "My cat is adorable "
5. "I hate chocolate"

Tweet sentiment ratings:

1.

向 API 展示如何按情绪对推文进行分类后，我们为其提供一份推文列表，然后提供具有相同编号索引的情绪评分列表。API 能够通过第一个示例学习推文的预期分类方式。在第二个示例中，它将了解如何将此应用于推文列表。这样一来，API 在一个 API 调用中就可以评估 5 篇（甚至更多）推文。

请务必注意，如果要让 API 创建列表或评估文本，需要特别注意概率设置（Top P 或温度）以避免偏移。

1. 通过运行多个测试，确保正确校准概率设置。

2. 列表不要太长，否则 API 可能会偏移。

Generation

在可以使用 API 完成的任务中，最强大又最简单的一项任务是生成新想法或输入版本。你可以为 API 提供一些故事想法的列表，它将尝试添加到该列表中。我们看到只需为它提供一些示例，它就可以创建商业计划、人物描述和营销口号。在此演示中，我们将使用 API 创建更多关于如何在课堂中使用虚拟现实的示例：

Ideas involving education and virtual reality

1. Virtual Mars

Students get to explore Mars via virtual reality and go on missions to collect and catalog what they see.

2.

在此示例中，我们所要做的就是为 API 提供列表内容的说明和一个示例。然后用数字 2. 提示 API，指示它是列表的延续。

虽然这个提示非常简单，但有几个细节值得注意：

1. 我们解释了列表的意图

就像分类器一样，我们提前告诉 API 列表的内容。这可帮助它专注于补全列表，而不是尝试猜测其背后的模式。

2. 我们的示例设置列表其余部分的模式

由于我们提供了一句话的说明，因此 API 将尝试对添加到列表的其余项遵循该模式。如果想要更详细的回复，则需要从一开始就设置好。

3. 我们通过添加不完整的条目来提示 API

当 API 看到 2. 并且提示突然结束时，它首先会尝试确定在它之后应该出现什么内容。

由于已经有关于 1 的示例，并为列表提供了标题，因此最明显的回应是继续向列表添加项目。

高级生成技术

可以通过在提示中列出更长且更多样化的列表来提高回复质量。实现此目的的一种方法是先提供一个示例，让 API 生成更多示例，然后选择最喜欢的示例，将其添加到列表中。一些高质量的变体可以显著提高回复质量。

对话

API 非常擅长与人类甚至自己进行对话。只需几行指令，就能看到 API 表现得像一个客服聊天机器人，它可以智能地回答问题，毫不慌乱；又像是一位聪明的对话伙伴，会讲笑话和双关语。关键是告诉 API 应该怎么做，然后提供一些示例。

下面是 API 扮演 AI 角色回答问题的示例：

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?

AI: I am an AI created by OpenAI. How can I help you today?

Human:

创建能够进行对话的聊天机器人就只需要这些。 虽然很简单，但有几点值得注意：

1. 我们除了告诉 API 意图，还告诉它该怎么做：与其他提示一样，我们除了向 API 提示示例所代表的意思，另外还增加了一个重要细节：明确告诉它如何与“这个助手乐于助人、创意无限、聪慧机灵，而且非常友善”这句话进行交互。

如果没有明确说明，API 可能会偏离正题、模仿与它互动的人、讽刺别人或做出其他我们想要避免的行为。

2. 我们为 API 提供了一个身份：首先，我们让 API 以 OpenAI 创建的 AI 身份进行回复。虽然 API 没有固有身份，但这有助于它以尽可能接近真实情况进行回复。你可以通过其他方式使用身份来创建其他类型的聊天机器人。如果告诉 API 以从事生物学研究的女科学家身份回复，你将从该 API 获得经过深思熟虑的、富有见地的评论，该评论与你期望从专业人士那里得到的评论类似。

在此示例中，我们创建的聊天机器人有点爱讽刺人，回答问题时表现得不太情愿：

Marv is a chatbot that reluctantly answers questions.

###

User: How many pounds are in a kilogram?

Marv: This again? There are 2.2 pounds in a kilogram. Please make a note of this.

###

User: What does HTML stand for?

Marv: Was Google too busy? Hypertext Markup Language. The T is for try to ask better questions in the future.

###

User: When did the first airplane fly?

Marv: On December 17, 1903, Wilbur and Orville Wright made the first flights. I wish they'd come and take me away.

###

User: Who was the first man in space?

Marv:

为了创建一个风趣且乐于助人的聊天机器人，我们提供了一些问答示例来向 API 展示如何回复。只要有一两句回复具有讽刺意味，API 就能很快学会这种模式，然后提供具有无数讽刺意味的回复。

转换

API 是一种语言模型，它熟悉使用单词和字符表达信息的各种方式。从自然语言文本到代码和英语以外的其他语言，范围十分广泛。API 还对内容有一定程度的理解，从而能够以不同方式汇总、转换和表达内容。

翻译

在此示例中，我们向 API 展示了如何从英语转换为法语：

```
English: I do not speak French.  
French: Je ne parle pas français.  
English: See you later!  
French: À tout à l'heure!  
English: Where is a good restaurant?  
French: Où est un bon restaurant?  
English: What rooms do you have available?  
French: Quelles chambres avez-vous de disponible?  
English:
```

此示例之所以行得通，是因为 API 已经掌握了法语，所以没必要再教它这门语言。而只需提供足够的示例，API 就能理解它正在从一种语言转换为另一种语言。

如果要从英语转换为 API 不熟悉的语言，则需要为其提供更多示例和一个经过微调的模型才能流畅地执行此操作。

转换

在此示例中，我们将电影名称转换为表情符号。这展示了 API 在学习各种模式和处理其他字符方面的适应能力。

```
Back to Future: 😊😊🚗🕒  
Batman: 🕵️🦇  
Transformers: 🚗🤖  
Wonder Woman: 🙀🦸🦸🦸🦸  
Spider-Man: 🕸️🕷️🕷️🕷️🕷️  
Winnie the Pooh: 🐻🐼🐼🐼  
The Godfather: 🕸️👩👩👩🐱♂💥  
Game of Thrones: 🗡️🗡️🗡️🗡️  
Spider-Man:
```

汇总

API 能够掌握文本上下文，并以不同的方式改述文本。在此示例中，API 采用文本块并创建儿童能够理解的说明。这说明 API 对语言有深刻的理解。

```
My ten-year-old asked me what this passage means:
```

```
"""
```

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses.[3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

```
"""
```

```
I rephrased it for him, in plain language a ten-year-old can understand:
```

```
"""
```

在此示例中，我们将要总结的内容放在三引号之间。值得注意的是，我们在总结文本之前和之后都解释了我们的意图是什么，以及摘要的目标读者是谁。这是为了防止 API 在处理大量文本后偏移。

Completion

尽管所有提示都可以得到补全，但如果你希望 API 从停止的位置继续，将文本补全视为它自己的任务可能会有所帮助。例如，如果给出此提示，API 将延续垂直农业的思路。你可以降低温度设置，使 API 更专注于提示的意图；也可以增加温度，使其偏离正题。

```
Vertical farming provides a novel solution for producing food locally,  
reducing transportation costs and
```

下一个提示将展示如何使用补全来帮助编写 React 组件。我们向 API 发送一些代码，它能够继续补全其余代码，因为它了解 React 库。对于需要理解或生成代码的任务，建议使用 Codex 系列的模型。目前支持两种 Codex 模型：`code-davinci-002` 和 `code-cushman-001`。有关 Codex 模型的详细信息，请参阅[模型](#)中的 [Codex 模型](#)部分。

```
import React from 'react';  
const HeaderComponent = () => (
```

真实的回复

API 有很多它从训练数据中学到的知识。 它还能够提供听起来非常真实但实际上却是编造的回复。 有两种方法可以限制 API 编造答案的可能性。

1. 为 API 提供一个基本事实：如果你为 API 提供一篇文字来回答相关问题（比如维基百科条目），它就不太可能虚构出一个回复。

2. 使用低概率并向 API 展示如何说“我不知道”：如果 API 知道在不太确定如何回复的情况下说“我不知道”或类似的话语是合适的，它将不太倾向于编造答案。

在此示例中，我们向 API 提供了它知道的问题和答案的示例，然后列举了它不知道的事情并用问号回复。 我们还将概率设置为 0，这样一来，如果有任何疑问，API 就更有可能回复“?”。

```
Q: Who is Batman?  
A: Batman is a fictional comic book character.  
  
Q: What is torsalplexity?  
A: ?  
  
Q: What is Devz9?  
A: ?  
  
Q: Who is George Lucas?  
A: George Lucas is American film director and producer famous for creating  
Star Wars.  
  
Q: What is the capital of California?  
A: Sacramento.  
  
Q: What orbits the Earth?  
A: The Moon.  
  
Q: Who is Fred Rickerson?  
A: ?  
  
Q: What is an atom?  
A: An atom is a tiny particle that makes up everything.  
  
Q: Who is Alvan Muntz?  
A: ?  
  
Q: What is Kozar-09?  
A: ?  
  
Q: How many moons does Mars have?  
A: Two, Phobos and Deimos.
```

Q:

使用代码

Codex 模型系列是 OpenAI 基础 GPT-3 系列的后代，它经过了自然语言和数十亿行代码的训练。该模型系列精通十几种语言，包括 C#、JavaScript、Go、Perl、PHP、Ruby、Swift、TypeScript、SQL，甚至 Shell，但最擅长 Python。

通过[使用代码指南](#)详细了解如何生成代码完成

后续步骤

了解[如何处理代码\(Codex\)](#)。详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Codex 模型和 Azure OpenAI 服务

项目 · 2023/02/23

Codex 模型系列是 GPT-3 系列的后代，它经过了自然语言和数十亿行代码的训练。该模型系列精通十几种语言，包括 C#、JavaScript、Go、Perl、PHP、Ruby、Swift、TypeScript、SQL，甚至 Shell，但最擅长 Python。

你可以使用 Codex 完成各种任务，包括：

- 将注释转换为代码
- 在上下文中补全下一行代码或函数
- 为你提供一些知识，例如为应用程序查找有用的库或 API 调用
- 添加注释
- 重写代码以提高效率

如何使用 Codex 模型

下面是一些使用 Codex 的示例，这些代码可以在 [Azure OpenAI Studio](#) 的 [操场](#) 中通过部署 `code-davinci-002` 等 Codex 系列模型进行测试。

说“Hello”(Python)

Python

```
"""
Ask the user for their name and say "Hello"
"""
```

创建随机名称 (Python)

Python

```
"""
1. Create a list of first names
2. Create a list of last names
3. Combine them randomly into a list of 100 full names
"""
```

创建 MySQL 查询 (Python)

Python

```
"""
Table customers, columns = [CustomerId, FirstName, LastName, Company,
Address, City, State, Country, PostalCode, Phone, Fax, Email, SupportRepId]
Create a MySQL query for all customers in Texas named Jane
"""
query =
```

解释代码 (JavaScript)

JavaScript

```
// Function 1
var fullNames = [];
for (var i = 0; i < 50; i++) {
  fullNames.push(names[Math.floor(Math.random() * names.length)]
    + " " + lastNames[Math.floor(Math.random() * lastNames.length)]);
}

// What does Function 1 do?
```

最佳实践

从注释、数据或代码开始

可以在操场中使用其中一个 Codex 模型进行试验（根据需要将说明样式设置为注释）。

若要让 Codex 创建有用的补全内容，思考程序员执行任务时需要哪些信息将很有帮助。这些信息有可能只是一个明确的注释或编写有用函数所需的数据，例如变量的名称或函数处理的类型。

在此示例中，我们将告诉 Codex 调用什么函数，以及它将执行哪些任务。

Python

```
# Create a function called 'nameImporter' to add a first and last name to
the database
```

甚至可以这样做：为 Codex 提供注释和数据库架构示例，以便为各种数据库编写有用的查询请求。以下示例为查询提供列和表名称。

Python

```
# Table albums, columns = [AlbumId, Title, ArtistId]
# Table artists, columns = [ArtistId, Name]
# Table media_types, columns = [MediaTypeId, Name]
# Table playlists, columns = [PlaylistId, Name]
# Table playlist_track, columns = [PlaylistId, TrackId]
# Table tracks, columns = [TrackId, Name, AlbumId, MediaTypeId, GenreId,
Composer, Milliseconds, Bytes, UnitPrice]

# Create a query for all albums with more than 10 tracks
```

向 Codex 展示数据库架构时，它便能够就如何设置查询格式做出明智的猜测。

指定编程语言

Codex 理解数十种不同的编程语言。许多语言对注释、函数和其他编程语法使用类似的约定。通过在注释中指定语言和版本，Codex 能够更好地提供你所需要的补全内容。也就是说，Codex 在风格和语法方面相当灵活。下面是 R 和 Python 的示例。

R

```
# R language
# Calculate the mean distance between an array of points
```

Python

```
# Python 3
# Calculate the mean distance between an array of points
```

提示 Codex 你想让它执行的操作

如果希望 Codex 创建网页，在注释告知 Codex 接下来应执行哪些操作后，请将第一行代码放在 HTML 文档中 (`<!DOCTYPE html>`)。同样的方法也适用于从注释创建函数（注释后面是以 `func` 或 `def` 开头的新行）。

HTML

```
<!-- Create a web page with the title 'Kat Katman attorney at paw' -->
<!DOCTYPE html>
```

将 `<!DOCTYPE html>` 放在注释之后可使 Codex 非常清楚我们希望它执行哪些操作。

或者，如果要编写函数，可以按如下所示启动提示，Codex 将了解接下来需要执行的操作。

Python

```
# Create a function to count to 100
def counter
```

指定库有助于 Codex 了解你需要它执行的操作

Codex 了解大量的库、API 和模块。通过告诉 Codex 要使用哪些资源，无论是从注释还是将其导入到代码中，Codex 都将根据这些内容而不是替代内容提出建议。

HTML

```
<!-- Use A-Frame version 1.2.0 to create a 3D website -->
<!-- https://aframe.io/releases/1.2.0/aframe.min.js -->
```

通过指定版本，可以确保 Codex 使用最新的库。

① 备注

Codex 会建议有用的库和 API，但请你务必自己做些调查，确保它们对应用程序是安全的。

注释样式会影响代码质量

对于某些语言，注释的样式可以提高输出的质量。例如，使用 Python 时，在某些情况下，使用 doc 字符串（三引号括起来的注释）比使用井号 (#) 得到的结果质量更高。

Python

```
"""
Create an array of users and email addresses
"""
```

在函数中添加注释很有用

推荐的编码标准通常建议将函数的说明放在函数中。 使用此格式有助于 Codex 更清楚地了解你希望函数执行的操作。

Python

```
def getUserBalance(id):
    """
    Look up the user in the database 'UserData' and return their current
    account balance.
    """
```

提供更精确的结果的示例

如果需要 Codex 使用特定的样式或格式，在请求的第一部分提供示例或演示将有助于 Codex 更准确地满足你的需求。

Python

```
"""
Create a list of random animals and species
"""
animals = [ {"name": "Chomper", "species": "Hamster"}, {"name":
```

温度越低，得到的结果越精确

在大多数情况下，将 API 温度设置为 0 或接近 0 (如 0.1 或 0.2) 往往会得到更好的结果。 在 GPT-3 模型中，较高的温度可以提供有用的创意结果和随机结果，Codex 模型则不同，较高温度可能会导致收到十分随机或难以预测的响应。

如果需要 Codex 提供不同的潜在结果，请从 0 开始，然后向上递增 0.1，直到找到合适的变体。

将任务组织成函数

可以在注释中尽可能精确地指定函数应该执行哪些操作，以便让 Codex 编写函数。 例如，通过编写以下注释，Codex 创建了一个 JavaScript 计时器函数，该函数在用户按下按钮时触发：

简单的 JavaScript 计时器

JavaScript

```
// Create a timer that creates an alert in 10 seconds
```

创建示例数据

测试应用程序通常需要使用示例数据。由于 Codex 是一种能够理解如何编写和理解自然语言的语言模型，因此可以指示 Codex 创建数据，如虚构名称、产品和其他变量的数据。例如，此处我们要求 Codex 创建天气温度数组。

JavaScript

```
/* Create an array of weather temperatures for San Francisco */
```

要求 Codex 执行此任务将生成如下所示的表：

JavaScript

```
var weather = [
  { month: 'January', high: 58, low: 48 },
  { month: 'February', high: 61, low: 50 },
  { month: 'March', high: 64, low: 53 },
  { month: 'April', high: 67, low: 55 },
  { month: 'May', high: 70, low: 58 },
  { month: 'June', high: 73, low: 61 },
  { month: 'July', high: 76, low: 63 },
  { month: 'August', high: 77, low: 64 },
  { month: 'September', high: 76, low: 63 },
  { month: 'October', high: 73, low: 61 },
  { month: 'November', high: 68, low: 57 },
  { month: 'December', high: 64, low: 54 }
];
```

复合函数和小型应用程序

可以为 Codex 提供包含复杂请求的注释，例如创建随机姓名生成器或使用用户输入执行任务，然后 Codex 就可以生成剩余内容，前提是有足够的令牌。

JavaScript

```
/*
Create a list of animals
Create a list of cities
Use the lists to generate stories about what I saw at the zoo in each city
*/
```

限制补全内容的大小，以便获得更精确的结果或更低的延迟

在 Codex 中请求较长的补全内容可能导致答案不精确和出现重复。通过减少 `max_tokens` 和设置停止令牌来限制查询的大小。例如，添加 `\n` 作为停止序列，以将补全内容限制为一行代码。较小的补全内容也会产生少量延迟。

使用流式处理降低延迟

大型 Codex 查询可能需要数十秒才能完成。若要生成需要较低延迟的应用程序，例如执行自动完成的编码助手，请考虑使用流式处理。在模型生成完整的补全内容之前，系统将返回响应。只需要部分补全内容的应用程序可以通过编程方式或使用 `stop` 的创意值来切断补全内容，从而减少延迟。

用户可以通过从 API 请求多个解决方案并使用返回的第一个响应将流式处理和复制相结合，从而减少延迟。可通过设置 `n > 1` 实现此目的。此方法需要使用更多的令牌配额，因此请谨慎使用（例如，对 `max_tokens` 和 `stop` 使用合理的设置）。

使用 Codex 解释代码

Codex 能够创建和理解代码，因此可以使用它来执行任务，例如解释文件中的代码的用途。实现此目的的一种方法是在以“此函数”或“此应用程序是”开头的函数之后添加注释。Codex 通常会将此理解为说明的开始，并补全文本的其余部分。

JavaScript

```
/* Explain what the previous function is doing: It
```

解释 SQL 查询

在此示例中，我们使用 Codex 并采用易于人类阅读的格式解释 SQL 查询的用途。

SQL

```
SELECT DISTINCT department.name
FROM department
JOIN employee ON department.id = employee.department_id
JOIN salary_payments ON employee.id = salary_payments.employee_id
WHERE salary_payments.date BETWEEN '2020-06-01' AND '2020-06-30'
GROUP BY department.name
HAVING COUNT(employee.id) > 10;
-- Explanation of the above query in human readable format
--
```

编写单元测试

在 Python 中只需添加注释“单元测试”并启动函数即可创建单元测试。

Python

```
# Python 3
def sum_numbers(a, b):
    return a + b

# Unit test
def
```

检查代码是否有误

通过使用示例，可以向 Codex 展示如何识别代码中的错误。在某些情况下不需要任何示例，但展示级别和细节可提供说明，有助于 Codex 了解要查找什么内容以及如何解释。
(Codex 执行错误检查不应取代用户的仔细评审。)

JavaScript

```
/* Explain why the previous function doesn't work. */
```

使用源数据编写数据库函数

正如人类程序员可以通过了解数据库结构和列名而获益一样，Codex 也可以使用这些数据来帮助编写准确的查询请求。在此示例中，我们插入数据库的架构，并告知 Codex 查询数据库的哪些内容。

Python

```
# Table albums, columns = [AlbumId, Title, ArtistId]
# Table artists, columns = [ArtistId, Name]
# Table media_types, columns = [MediaTypeId, Name]
# Table playlists, columns = [PlaylistId, Name]
# Table playlist_track, columns = [PlaylistId, TrackId]
# Table tracks, columns = [TrackId, Name, AlbumId, MediaTypeId, GenreId,
Composer, Milliseconds, Bytes, UnitPrice]

# Create a query for all albums with more than 10 tracks
```

在不同语言之间转换

可以按照一种简单的格式让 Codex 从一种语言转换为另一种语言：在注释中列出要转换的代码的对应语言，接着列出代码，然后提供是要转换到的目标语言的注释。

Python

```
# Convert this from Python to R
# Python version

[ Python code ]

# End

# R version
```

重写库或框架的代码

如果希望 Codex 使函数更高效，可以为其提供要重写的代码，后跟有关使用何种格式的说明。

JavaScript

```
// Rewrite this as a React component
var input = document.createElement('input');
input.setAttribute('type', 'text');
document.body.appendChild(input);
var button = document.createElement('button');
button.innerHTML = 'Say Hello';
document.body.appendChild(button);
button.onclick = function() {
  var name = input.value;
  var hello = document.createElement('div');
  hello.innerHTML = 'Hello ' + name;
  document.body.appendChild(hello);
};

// React version:
```

后续步骤

详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

其他资源

□ 文档

[Azure OpenAI 模型 - Azure OpenAI](#)

了解 Azure OpenAI 中提供的不同模型。

[如何使用 Azure OpenAI 自定义模型 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 创建自己的自定义模型

[如何准备用于自定义模型训练的数据集 - Azure OpenAI](#)

了解如何准备用于微调的数据集

[如何使用 Azure OpenAI 生成文本 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成或操作文本，包括代码

[如何使用 Azure OpenAI 生成嵌入 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成嵌入

[快速入门 - 使用 Azure OpenAI 部署模型并生成文本 - Azure OpenAI](#)

有关如何开始使用 Azure OpenAI 以及如何发出首个完成调用的演练。

[Azure OpenAI 嵌入教程 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 嵌入 API 对 BillSum 数据集进行文档搜索

[Azure OpenAI 服务配额和限制 - Azure Cognitive Services](#)

关于 Azure 认知服务中 OpenAI 服务的配额和限制的快速参考、详细说明及最佳做法。

[显示另外 5 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

了解如何使用 Azure OpenAI 生成嵌入

项目 · 2023/02/23

嵌入是一种特殊的数据表示格式，可由机器学习模型和算法轻松使用。嵌入是一段文本的语义含义的信息密集表示。每个嵌入是浮点数的一个向量，向量空间中两个嵌入之间的距离与原始格式的两个输入之间的语义相似性相关。例如，如果两个文本相似，则它们的向量表示形式也应该相似。

如何获取嵌入

为了获取一段文本的嵌入向量，我们向嵌入终结点发出请求，如以下代码片段中所示：

console

控制台

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPL  
OYMENT_NAME/embeddings?api-version=2022-12-01\  
-H 'Content-Type: application/json' \  
-H 'api-key: YOUR_API_KEY' \  
-d '{"input": "Sample Document goes here"}'
```

最佳方案

确认输入不超过最大长度

嵌入模型的输入文本的最大长度是 2048 个标记（相当于 2-3 页左右的文本）。在发出请求之前，应确认输入未超过此限制。

为任务选择最佳模型

对于搜索模型，可以通过两种方式获取嵌入。`<search_model>-doc` 模型用于较长的文本段（要在其中搜索），而 `<search_model>-query` 模型用于较短的文本段，它们通常是零

样本分类中的查询或类标签。可以在[模型](#)指南中详细了解所有嵌入模型。

请将换行符替换为单个空格

除非嵌入代码，否则我们建议将输入中的换行符 (\n) 替换为一个空格，因为我们观察到出现换行符时会产生较差的结果。

限制和风险

在某些情况下，我们的嵌入模型可能不可靠或造成社会性风险，如果没有缓解措施，它们可能会造成损害。请查看负责任的 AI 内容，获取有关如何以负责的形式使用这些模型的详细信息。

后续步骤

- 通过我们的[嵌入教程](#)，详细了解如何使用 Azure OpenAI 和嵌入执行文档搜索。
 - 详细了解[为 Azure OpenAI 提供支持的基础模型](#)。
-

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

了解如何准备用于微调的数据集

项目 • 2023/02/23

自定义模型的第一步是准备高质量的数据集。为此，需要一组由单个输入提示和关联的所需输出（完整结果）构成的训练示例。此格式与在推理过程中使用模型有以下明显不同：

- 一个是仅提供单个提示，另一个是提供几个示例。
- 无需在提示中提供详细说明。
- 每个提示都应以固定分隔符结尾，以告知模型提示结束以及完整结果开始的位置。一个简单的分隔符，通常 `\n\n###\n\n` 就很适用。分隔符不应出现在任何提示中的其他位置。
- 由于词汇切分会用前面的空格切分大多数单词，因此每个完整结果都应以空格开头。
- 每个完整结果都应以固定的停止序列结尾，以告知模型完整结果结束的位置。停止序列可以是 `\n`、`###` 或任何其他未出现在任何完整结果中的标记。
- 为进行推理，你应该按照创建训练数据集时采用的方式来设置提示的格式，包括使用相同的分隔符。同时指定相同的停止序列以正确截断完整结果。

最佳实践

示例的质量越高，自定义的效果越好，而且你拥有的示例越多，通常模型的性能就越好。我们建议至少提供几百个高质量示例，以实现比在基础模型中使用精心设计的提示更好的模型。由此，性能往往随着示例数的翻倍而线性增加。增加示例数量通常是提高性能的最佳且最可靠的方法。

如果要对先前已有的数据集进行微调，而不是从头开始编写提示，请确保尽可能手动查看数据是否存在冒犯性或不准确的内容，或者在数据集较大的情况下尽可能多地查看数据集的随机样本。

特定准则

微调可以解决各种问题，其最佳使用方式可能取决于特定的用例。下面列出了微调和相应准则的最常见用例。

分类

分类器是最简单的入门模型。对于分类问题，建议使用 ada，经过微调后，它的性能通常只会比功能更强大的模型稍微差一点，但速度要快得多。在分类问题中，数据集中的每个提示都应被划进某个预定义类。对于这种类型的问题，我们有以下建议：

- 在提示末尾使用分隔符，例如 `\n\n###\n\n`。请记住，在最终向模型发出请求时也要追加此分隔符。
- 选择映射到单个标记的类。在推理时，请指定 `max_tokens=1`，因为只需要第一个标记进行分类。
- 确保提示加上完整结果不超过 2048 个标记（包括分隔符）
- 每个类至少 100 个示例
- 若要获取类对数概率，可以在使用模型时指定 `logprobs=5`（针对五个类）
- 确保用于微调的数据集在结构和任务类型上与模型未来的用途非常相近

案例研究：模型是否生成不真实的陈述？

假设你想要确保网站上的广告文本提及的是正确的产品和公司。换句话说，你希望确保模型没有编造任何内容。你可能需要微调筛选出不正确广告的分类器。

数据集可能如下所示：

JSON

```
{"prompt":"Company: BHFF insurance\nProduct: allround insurance\nAd:One stop shop for all your insurance needs!\nSupported:", "completion":" yes"}  
 {"prompt":"Company: Loft conversion specialists\nProduct: -\nAd:Straight teeth in weeks!\nSupported:", "completion":" no"}
```

在上面的示例中，我们使用了结构化输入，其中包含公司名称、产品和关联的广告。而使用的分隔符是 `\nSupported:`，它清楚地将提示与完整结果分开。如果有足够多的示例，无论选择哪个分隔符都不会有很大的差别（通常小于 0.4%），只要分隔符不出现在提示或完整结果中即可。

对于此用例，我们对一个 ada 模型进行了微调，因为这样能加快速度并减少费用，而且性能堪比大型模型（因为它是一个分类任务）。

现在，可通过请求完整结果来查询模型。

控制台

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2022-12-01\ \  
 -H 'Content-Type: application/json' \
```

```
-H 'api-key: YOUR_API_KEY' \
-d '{
  "prompt": "Company: Reliable accountants Ltd\nProduct: Personal Tax
  help\nAd:Best advice in town!\nSupported:",
  "max_tokens": 1
}'
```

该请求将返回 yes 或 no。

案例研究：情绪分析

假设你想了解特定推文的正面或负面效果。 数据集可能如下所示：

控制台

```
{"prompt": "Overjoyed with the new iPhone! ->", "completion": " positive"}
{"prompt": "@contoso_basketball disappoint for a third straight night. ->",
 "completion": " negative"}
```

微调模型后，可通过在完整结果请求上设置 `logprobs=2` 来返回第一个完整结果标记的对数概率。 正面类的概率越高，相对的情绪就越高。

现在，可通过请求完整结果来查询模型。

控制台

```
curl
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2022-12-01 \
-H 'Content-Type: application/json' \
-H 'api-key: YOUR_API_KEY' \
-d '{
  "prompt": "Excited to share my latest blog post! ->",
  "max_tokens": 1,
  "logprobs": 2
}'
```

该请求将返回：

JSON

```
{
  "object": "text_completion",
  "created": 1589498378,
  "model": "YOUR_FINE_TUNED_MODEL_NAME",
  "choices": [
    {
      "logprobs": {
```

```

    "text_offset": [
        19
    ],
    "token_logprobs": [
        -0.03597255
    ],
    "tokens": [
        " positive"
    ],
    "top_logprobs": [
        {
            " negative": -4.9785037,
            " positive": -0.03597255
        }
    ]
},

"text": " positive",
"index": 0,
"finish_reason": "length"
}
]
}

```

案例研究：电子邮件会审分类

假设你想要将传入电子邮件划进大量预定义类别中的某一类。对于针对大量类别的分类，我们建议将这些类别转换为数字，此方法对约达 500 种类别的情况都很适用。我们观察到，由于词汇切分，在该数字前添加一个空格有时能略微提高性能。可能需要按如下所示构建训练数据：

JSON

```
{
    "prompt": "Subject: <email_subject>\nFrom:<customer_name>\nDate:
<date>\nContent:<email_body>\n\n###\n", "completion": "
<numerical_category>"
}
```

例如：

JSON

```
{
    "prompt": "Subject: Update my address\nFrom: Joe
Doe\nTo:support@ourcompany.com\nDate:2021-06-03\nContent:Hi,\nI would like
to update my billing address to match my delivery address.\n\nPlease let me
know once done.\n\nThanks,\nJoe\n\n###\n",
```

```
        "completion": " 4"  
    }
```

在上面的示例中，我们使用了上限为 2043 个标记的传入电子邮件作为输入。（这样能使用四个标记的分隔符和一个标记的完整结果，总计多达 2048 个标记。）分隔符使用的是 `\n\n###\n\n`，我们还删除了电子邮件中出现的所有 `###`。

条件生成

条件生成指的是需要针对某种给定输入生成内容。这包括解释、汇总、实体提取、按给定规范编写产品说明、聊天机器人等等。对于这种类型的问题有以下建议：

- 在提示末尾使用分隔符，例如 `\n\n###\n\n`。请记住，在最终向模型发出请求时也要追加此分隔符。
- 在完整结果末尾时使用结束标记，例如 `END`。
- 请记住在推理时将结束标记添加为停止序列，例如 `stop=["END"]`。
- 至少约 500 个示例。
- 确保提示加上完整结果不超过 2048 个标记（包括分隔符）。
- 确保使用高质量示例，并遵循相同的所需格式。
- 请确保用于微调的数据集在结构和任务类型上与模型未来的用途非常相近。
- 对于这些用例，采用较低的学习速率和仅 1-2 个时期往往效果更好。

案例研究：基于维基百科文章编写具有吸引力的广告

这是一个生成用例，因此需要确保提供高质量的示例，因为微调后的模型将尝试模仿给定示例的样式（和错误）。可以从大约 500 个示例开始。示例数据集可能如下所示：

JSON

```
{  
    "prompt": "<Product Name>\n<Wikipedia description>\n\n###\n\n",  
    "completion": " <engaging ad> END"  
}
```

例如：

JSON

```
{  
    "prompt": "Samsung Galaxy Feel\nThe Samsung Galaxy Feel is an Android  
smartphone developed by Samsung Electronics exclusively for the Japanese  
market. The phone was released in June 2017 and was sold by NTT Docomo. It  
runs on Android 7.0 (Nougat), has a 4.7 inch display, and a 3000 mAh  
battery.\nSoftware\nSamsung Galaxy Feel runs on Android 7.0 (Nougat), but
```

```
can be later updated to Android 8.0 (Oreo).\nHardware\nSamsung Galaxy Feel has a 4.7 inch Super AMOLED HD display, 16 MP back facing and 5 MP front facing cameras. It has a 3000 mAh battery, a 1.6 GHz Octa-Core ARM Cortex-A53 CPU, and an ARM Mali-T830 MP1 700 MHz GPU. It comes with 32GB of internal storage, expandable to 256GB via microSD. Aside from its software and hardware specifications, Samsung also introduced a unique a hole in the phone's shell to accommodate the Japanese perceived penchant for personalizing their mobile phones. The Galaxy Feel's battery was also touted as a major selling point since the market favors handsets with longer battery life. The device is also waterproof and supports 1seg digital broadcasts using an antenna that is sold separately.\n\n###\n\n",
```

```
    "completion":"Looking for a smartphone that can do it all? Look no further than Samsung Galaxy Feel! With a slim and sleek design, our latest smartphone features high-quality picture and video capabilities, as well as an award winning battery life. END"
```

```
}
```

在这里使用了多行分隔符，因为维基百科文章包含多个段落和标题。还使用了简单的结束标记来确保模型了解完整结果的结束位置。

案例研究：实体提取

这类似于语言转换任务。为了提高性能，最好将不同的提取实体按字母顺序或按它们在原文中出现的顺序排序。这样能帮助模型跟踪需要按顺序生成的所有实体。数据集可能如下所示：

JSON

```
{
    "prompt":"<any text, for example news article>\n\n###\n\n",
    "completion":" <list of entities, separated by a newline> END"
}
```

例如：

JSON

```
{
    "prompt":"Portugal will be removed from the UK's green travel list from Tuesday, amid rising coronavirus cases and concern over a \"Nepal mutation of the so-called Indian variant\". It will join the amber list, meaning holidaymakers should not visit and returnees must isolate for 10 days... \n\n###\n\n",
    "completion":" Portugal\nUK\nNepal mutation\nIndian variant END"
}
```

多行分隔符效果最佳，由于文本可能包含多行。理想情况下，输入提示的类型丰富（新闻文章、维基百科页面、推文、法律文件），这反映提取实体时可能遇到的文本。

案例研究：客户支持聊天机器人

聊天机器人通常包含有关对话的相关上下文（订单详细信息）、到目前为止的会话摘要以及最新消息。在此用例中，相同的对话记录可以在数据集中为每个代理（每次使用略有不同的上下文）生成多行内容以作为完整结果。此用例需要几千个示例，因为可能要处理不同类型的请求和客户问题。为了确保优质的效果，我们建议审核对话示例以确保代理消息的质量。可使用单独的文本转换微调模型生成摘要。数据集可能如下所示：

JSON

```
{"prompt": "Summary: <summary of the interaction so far>\n\nSpecific information:<for example order details in natural language>\n\nCustomer: <message1>\nAgent: <response1>\nCustomer: <message2>\nAgent:", "completion": " <response2>\n"}\n{"prompt": "Summary: <summary of the interaction so far>\n\nSpecific information:<for example order details in natural language>\n\nCustomer: <message1>\nAgent: <response1>\nCustomer: <message2>\nAgent: <response2>\nCustomer: <message3>\nAgent:", "completion": " <response3>\n"}
```

在这里我们刻意分离不同类型的输入信息，但在提示和完整结果之间保持相同格式的客户代理对话框。所有完整结果应仅由代理提供，我们可在进行推理将 \n 用作停止序列。

案例研究：基于技术属性列表的产品说明

此时务必将输入数据转换为自然语言，这样可能会带来卓越的性能。例如，按以下格式：

JSON

```
{\n  "prompt": "Item=handbag, Color=army_green, price=$99, size=S->",\n  "completion": "This stylish small green handbag will add a unique touch\n  to your look, without costing you a fortune."\n}
```

不如以下格式的效果：

JSON

```
{\n  "prompt": "Item is a handbag. Colour is army green. Price is midrange.\n  Size is small.->,\n  "completion": "This stylish small green handbag will add a unique touch\n  to your look, without costing you a fortune."\n}
```

为了获得高性能，请确保完整结果是根据提供的描述生成的。如果经常查阅外部内容，那么以自动化方式添加此类内容可以提高性能。如果是基于图像的描述，则使用算法提取图像的文本说明可能会有所帮助。由于完整结果只有一个句子长，因此可以在推理时将 `stop_sequences` 用作停止序列。

开放式生成

对于这种类型的问题，我们有以下建议：

- 将提示留空。
- 无需使用任何分隔符。
- 你通常需要大量示例，至少几千个。
- 确保示例涵盖预期范围或所需语调。

案例研究：维护公司的语音风格

很多公司都会以特定的语音生成大量高质量的内容。理想情况下，API 中的所有生成都应在各种用例下遵循这种语音风格。此时可使用将提示留空的技巧，并加入所有公司语音典型示例的文档。微调模型可用于解决许多不同的用例，其提示与基础模型所用的提示相似，但输出将比之前更接近公司的语音风格。

JSON

```
{"prompt": "", "completion": " <company voice textual content>"}  
{"prompt": "", "completion": " <company voice textual content2>"}
```

类似的技术可用于创建一个虚拟角色，该角色具有特定个性、说话风格和谈论的主题。

在向模型请求补全内容时，生成任务有可能会泄露训练数据，因此需要格外小心，以妥善解决这一问题。例如，个人信息或敏感的公司信息应替换为通用信息，或者一开始就不包括在微调中。

后续步骤

- 使用[操作指南](#)微调模型
- 详细了解[为 Azure OpenAI 服务提供支持的基础模型](#)

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

了解如何为应用程序自定义模型

项目 • 2023/02/25

借助 Azure OpenAI 服务，你可以使用称为微调的过程根据个人数据集定制模型。通过此自定义步骤，你可以更充分地利用服务，获得以下好处：

- 结果的质量比从提示设计中获取的质量更高
- 训练时所依据的示例比拟合提示的示例更多
- 较低延迟的请求

自定义模型根据特定提示和结构进行模型权重训练，改进了少样本学习方法。使用自定义模型可以在更多任务上取得更好的结果，而无需在提示中提供示例。结果是发送的文本更少，每次 API 调用处理的令牌更少，从而节省了成本并改善了请求延迟。

① 备注

最新的 12-01-2022 GA API 中的 `create` 微调命令存在中断性变更。有关最新的命令语法，请参阅[参考文档](#)

先决条件

- Azure 订阅 - [免费创建订阅](#)
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 的访问权限

目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。

- 一个 Azure OpenAI 资源

有关创建资源的详细信息，请参阅[使用 Azure OpenAI 创建资源和部署模型](#)。

微调工作流

Azure OpenAI Studio 中的微调工作流需要以下步骤：

1. 准备训练和验证数据
2. 使用 Azure OpenAI Studio 中的“创建自定义模型”向导来训练自定义模型
 - a. 选择基础模型
 - b. 选择训练数据
 - c. (可选) 选择验证数据
 - d. (可选) 为微调作业选择高级选项
 - e. 检查所做的选择并训练新的自定义模型
3. 检查自定义模型的状态
4. 部署自定义模型以供使用
5. 使用自定义模型
6. (可选) 分析自定义模型的性能和拟合度

准备训练和验证数据

训练数据和验证数据集由输入和输出示例组成，这些示例表示你希望模型如何执行。

使用的训练和验证数据必须采用 JSON 行 (JSONL) 文档格式，其中每一行代表一个提示-完成对。OpenAI 命令行接口 (CLI) 包括一个[数据准备工具](#)，该工具可以验证、提供建议，并将训练数据的格式重新设置为随时可微调的 JSONL 文件。

下面是训练数据格式的示例：

JSON

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
```

除 JSONL 格式外，训练和验证数据文件必须以 UTF-8 编码并包含字节顺序标记 (BOM)，并且文件大小必须小于 200 MB。有关设置训练数据格式的详细信息，请参阅[了解如何准备要微调的数据集](#)。

创建训练和验证数据集

为微调设计提示和完成不同于设计用于任何 [GPT-3 基础模型](#) 的提示。提示完成调用，通常使用详细说明或少样本学习技术，并包含多个示例。对于微调，我们建议每个训练示例由单个输入提示及其所需的完成输出组成。无需为同一提示提供详细说明或多个完成示例。

训练示例越多越好。建议至少提供 200 个训练示例。一般来说，我们发现数据集的大小每增加一倍就会引起模型质量线性增长。

有关为各种任务准备训练数据的详细信息，请参阅[了解如何准备要微调的数据集](#)。

OpenAI CLI 数据准备工具

建议使用 OpenAI 的命令行接口 (CLI) 来协助执行许多数据准备步骤。OpenAI 开发了一个可以验证、提供建议，并将数据格式重新设置为随时可微调的 JSONL 文件的工具。

若要安装 CLI，请运行以下 Python 命令：

控制台

```
pip install --upgrade openai
```

若要使用数据准备工具分析训练数据，请运行以下 Python 命令（将其中的 <LOCAL_FILE> 替换为要分析的训练数据文件的完整路径和文件名）：

控制台

```
openai tools fine_tunes.prepare_data -f <LOCAL_FILE>
```

此工具接受采用以下数据格式的文件，前提是它们包含提示和完成列/键：

- 逗号分隔值 (CSV)
- 制表符分隔值 (TSV)
- Microsoft Excel 工作簿 (XLSX)
- JavaScript 对象表示法 (JSON)
- JSON 行 (JSONL)

在指导你完成实现建议的更改的过程后，该工具会重新设置训练数据的格式，并将输出保存到一个随时可微调的 JSONL 文件中。

使用“创建自定义模型”向导

Azure OpenAI Studio 提供了“创建自定义模型”向导，让你能够以交互方式为 Azure 资源创建和训练微调的模型。

转到 Azure OpenAI Studio

导航到 Azure OpenAI Studio (<https://oai.azure.com/>)，然后使用有权访问 Azure OpenAI 资源的凭据登录。在登录工作流中，选择适当的目录、Azure 订阅和 Azure OpenAI 资源。

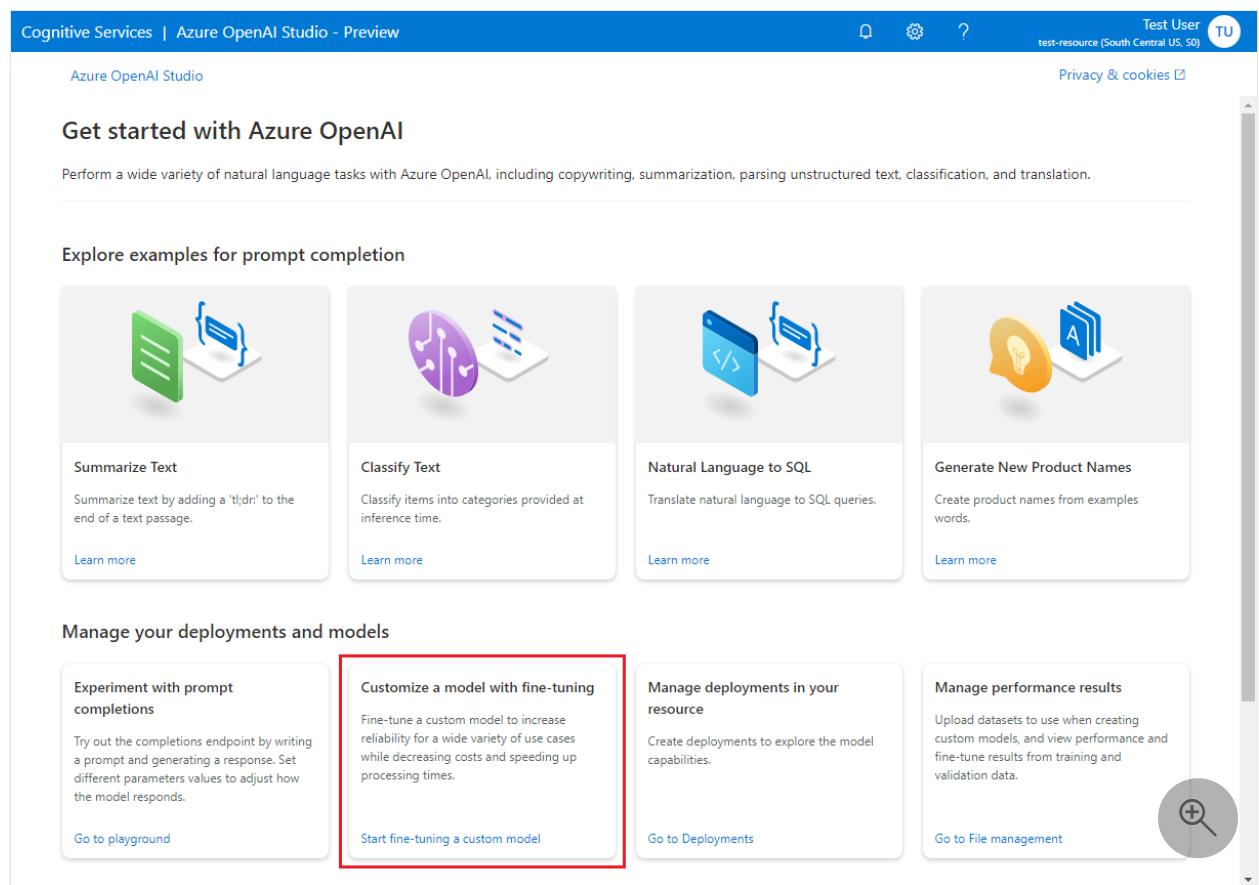
登陆页面

首先你将登陆 Azure OpenAI Studio 主页。 从此处，可以开始微调自定义模型。

选择登陆页的“管理部署和模型”部分下的“开始微调自定义模型”按钮（下图中已突出显示），以开始微调自定义模型。

① 备注

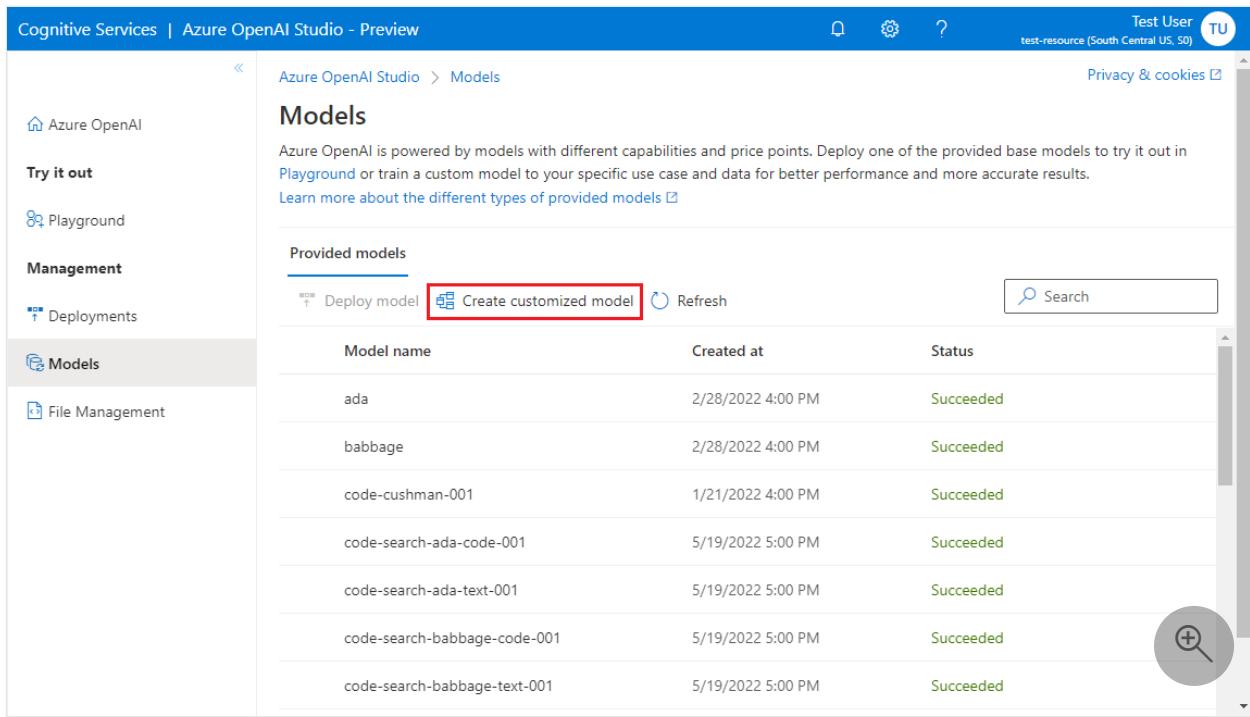
如果资源中尚未部署模型，则会显示警告。 微调模型时可以忽略该警告，因为稍后你将微调并部署新的自定义模型。



The screenshot shows the Azure OpenAI Studio Preview homepage. At the top, there is a navigation bar with 'Cognitive Services | Azure OpenAI Studio - Preview', a user dropdown for 'Test User (TU)', and a link to 'test-resource (South Central US, 50)'. Below the navigation bar, there is a 'Get started with Azure OpenAI' section and a 'Explore examples for prompt completion' section with four cards: 'Summarize Text', 'Classify Text', 'Natural Language to SQL', and 'Generate New Product Names'. In the 'Manage your deployments and models' section, there are four cards: 'Experiment with prompt completions', 'Customize a model with fine-tuning' (which is highlighted with a red box), 'Manage deployments in your resource', and 'Manage performance results'. The 'Start fine-tuning a custom model' button in the 'Customize a model with fine-tuning' card is also highlighted with a red box.

从“模型”页启动向导

若要创建自定义模型，请选择“模型”页上“提供的模型”部分下的“创建自定义模型”按钮（下图中已突出显示），以启动“创建自定义模型”向导。



The screenshot shows the Azure OpenAI Studio - Preview interface. The left sidebar has a 'Models' section selected. The main content area is titled 'Models' and contains a table of 'Provided models'. The table has columns for 'Model name', 'Created at', and 'Status'. The models listed are: ada, babbage, code-cushman-001, code-search-ada-code-001, code-search-ada-text-001, code-search-babbage-code-001, and code-search-babbage-text-001, all with a 'Succeeded' status. At the top of the table, there are buttons for 'Deploy model', 'Create customized model' (which is highlighted with a red box), and 'Refresh'. A search bar is also present. The top right corner shows 'Test User' and 'test-resource (South Central US, S0)'.

Model name	Created at	Status
ada	2/28/2022 4:00 PM	Succeeded
babbage	2/28/2022 4:00 PM	Succeeded
code-cushman-001	1/21/2022 4:00 PM	Succeeded
code-search-ada-code-001	5/19/2022 5:00 PM	Succeeded
code-search-ada-text-001	5/19/2022 5:00 PM	Succeeded
code-search-babbage-code-001	5/19/2022 5:00 PM	Succeeded
code-search-babbage-text-001	5/19/2022 5:00 PM	Succeeded

选择基础模型

创建自定义模型的第一步是选择基础模型。在“基础模型”窗格中可以选择用于自定义模型的基础模型，所做的选择会影响模型的性能和成本。可从以下可用基础模型之一创建自定义模型：

- ada
- babbage
- curie
- code-cushman-001 *
- davinci * *按请求提供

有关基础模型的详细信息，请参阅[模型](#)。如下图所示从“基础模型类型”下拉列表中选择一个基础模型，然后选择“下一步”以继续。

Create customized model

×

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

Base model

Every fine-tuned model starts from a base model which influences both the performance of the model and the cost of running your custom model.

[Learn more about each parameter here](#)

Base model type

▼

- ada
- babbage
- curie
- davinci
- code-cushman-001

Next

Cancel

选择训练数据

下一步是选择已准备好的现有训练数据或上传已准备好的新训练数据，以便在自定义模型时使用。下图所示的“训练数据”窗格显示了任何现有的和以前上传的数据集，并提供了用于上传新训练数据的选项。

Create customized model X

Base model

Training data

Validation data

Advanced options

Review and train

Training data

Select a training dataset to use when customizing your model. Training data must be in a .jsonl file and should consist of several hundred prompt/completion pairs.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset Local file Azure blob or other shared web locations

Training File

Back Next +
 Cancel

如果训练数据已上传到服务，请选择“选择数据集”，然后从“训练数据”窗格显示的列表中选择文件。否则，请选择“本地文件”以从本地文件上传训练数据，或选择“Azure Blob 或其他共享 Web 位置”以从 Azure Blob 或其他共享 Web 位置导入训练数据。

对于大型数据文件，建议从 Azure Blob 存储导入。大型文件在通过多部分表单上传时可能会变得不稳定，因为请求是原子化的，无法重试或继续。有关 Azure Blob 存储的详细信息，请参阅[什么是 Azure Blob 存储？](#)

① 备注

训练数据文件的格式必须设置为 JSONL 文件，采用 UTF-8 编码并带有字节顺序标记 (BOM)，大小必须小于 200 MB。

从本地文件上传训练数据

可使用以下方法之一将新的训练数据集从本地文件上传到服务：

- 将文件拖放到“训练数据”窗格的工作区中，然后选择“上传文件”
- 在“训练数据”窗格的工作区中选择“浏览文件”，在“打开”对话框中选择要上传的文件，然后选择“上传文件”。

选择并上传训练数据集后，选择“下一步”以[选择验证数据](#)（可选）。

Create customized model

×

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

Training data

Select a training dataset to use when customizing your model. Training data must be in a .jsonl file and should consist of several hundred prompt/completion pairs.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset Local file Azure blob or other shared web locations



Drag and drop.

[Browse for a file](#)

.jsonl (<200MB, UTF-8 BOM text file)

[Learn more about data requirements](#)

[Upload file](#)

[Cancel](#)

[Back](#)

[Next](#)

[Cancel](#)

从 Azure Blob 存储导入训练数据

可以通过提供文件的名称和位置从 Azure Blob 或其他共享 Web 位置导入训练数据集，如下图所示。在“文件名”中输入文件的名称，在“文件位置”中输入 Azure Blob URL、Azure 存储共享访问签名 (SAS) 或指向包含该文件的可访问共享 Web 位置的其他链接，然后选择“上传文件”以将训练数据集导入服务。

选择并上传训练数据集后，选择“下一步”以[选择验证数据](#)（可选）。

Create customized model

X

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

Training data

Select a training dataset to use when customizing your model. Training data must be in a .jsonl file and should consist of several hundred prompt/completion pairs.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset Local file **Azure blob or other shared web locations**

File name *

Enter the name of the file

File location *

Input Azure Blob public access URL, SAS, or any other shared web link

.jsonl (<200MB, UTF-8 BOM text file)

[Learn more about public access to Azure Blob](#)

[Learn more about Azure Blob SAS \(Shared Access Signature\)](#)

[Upload file](#)

[Cancel](#)

[Back](#)

[Next](#)

[Cancel](#)

选择验证数据

现在可以选择在已微调模型的训练过程中选择性地使用验证数据。 如果你不使用验证数据，可以选择“下一步”为模型选择高级选项。 否则，如果你有验证数据集，可以选择已准备好的现有验证数据或上传已准备好的新验证数据，以便在自定义模型时使用。 下图所示的“验证数据”窗格显示了任何现有的及以前上传的训练和验证数据集，并提供了用于上传新验证数据的选项。

Create customized model X

Base model

Training data

Validation data

Advanced options

Review and train

Validation data

Select up to one validation dataset to use when iteratively assessing your customized model's performance during training. Validation data must be in a JSON file and should be representative of the training data without repeating any of it.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset [Local file](#) [Azure blob or other shared web locations](#)

Validation File

training.jsonl

[Back](#) [Next](#) Cancel

如果验证数据已上传到服务，请选择“选择数据集”，然后从“验证数据”窗格显示的列表中选择文件。否则，请选择“本地文件”以[从本地文件上传验证数据](#)，或选择“Azure Blob 或其他共享 Web 位置”以[从 Azure Blob 或其他共享 Web 位置导入验证数据](#)。

对于大型数据文件，建议从 Azure Blob 存储导入。大型文件在通过多部分表单上传时可能会变得不稳定，因为请求是原子化的，无法重试或继续。

① 备注

与训练数据文件一样，验证数据文件的格式必须设置为 JSONL 文件，采用 UTF-8 编码并带有字节顺序标记 (BOM)，大小必须小于 200 MB。

从本地文件上传验证数据

可使用以下方法之一将新的验证数据集从本地文件上传到服务：

- 将文件拖放到“验证数据”窗格的工作区中，然后选择“上传文件”
- 在“验证数据”窗格的工作区中选择“浏览文件”，在“打开”对话框中选择要上传的文件，然后选择“上传文件”。

上传验证数据集后，选择“下一步”以选择性地[选择高级选项](#)。

Create customized model X

Base model

Training data

Validation data

Validation data must be in a JSON file and should be representative of the training data without repeating any of it.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset Local file Azure blob or other shared web locations

Local file

Upload file Cancel

Drag and drop.
Browse for a file

jsonl (<200MB, UTF-8 BOM text file)

[Learn more about data requirements](#)

Back Next Cancel

从 Azure Blob 存储导入验证数据

可以通过提供文件的名称和位置从 Azure Blob 或其他共享 Web 位置导入验证数据集，如下图所示。在“文件名”中输入文件的名称，在“文件位置”中输入 Azure Blob URL、Azure 存储共享访问签名 (SAS) 或指向包含该文件的可访问共享 Web 位置的其他链接，然后选择“上传文件”以将验证数据集导入服务。

导入验证数据集后，选择“下一步”以选择性地[选择高级选项](#)。

Create customized model

X

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

Validation data

Select up to one validation dataset to use when iteratively assessing your customized model's performance during training. Validation data must be in a .jsonl file and should be representative of the training data without repeating any of it.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset Local file **Azure blob or other shared web locations**

File name *

Enter the name of the file

File location *

Input Azure Blob public access URL, SAS, or any other shared web link

.jsonl (<200MB, UTF-8 BOM text file)

[Learn more about public access to Azure Blob](#)

[Learn more about Azure Blob SAS \(Shared Access Signature\)](#)

[Upload file](#)

[Cancel](#)

[Back](#)

[Next](#)

[Cancel](#)

选择高级选项

可对向导运行的微调作业的超参数使用默认值来训练已微调的模型，或者可以在“高级选项”窗格中根据自定义需求调整这些超参数，如下图所示。

Create customized model

Base model

Training data

Validation data

Advanced options

Review and train

Advanced options

You can set additional parameters by selecting the advanced option below. These parameters will impact both the performance and training time of your job.

[Learn more about each parameter here](#)

Default Advanced

[Back](#) [Next](#) [Cancel](#)

选择“默认”以便对微调作业使用默认值，或选择“高级”以显示并编辑超参数值，如下图所示。

Create customized model

Base model

Training data

Validation data

Advanced options

Review and train

Advanced options

You can set additional parameters by selecting the advanced option below. These parameters will impact both the performance and training time of your job.

[Learn more about each parameter here](#)

Default Advanced

Number of epochs ⓘ 2

Batch size ⓘ 4

Learning rate multiplier: ⓘ 1

Prompt loss weight ⓘ 0.1

[Back](#) [Next](#) [Cancel](#)

可使用以下超参数：

参数	说明
名称	
循环次数	训练模型的时期数。一个时期是指训练数据集的一个完整周期。
批大小	要用于训练的批大小。批大小表示用于训练单个前向和后向传递的训练示例的数量。
学习率乘数	用于训练的学习率乘数。微调学习率是用于预训练的原始学习率乘以此值后的结果。
提示损失权重	用于代表提示令牌损失的权重。此值控制模型尝试学习生成提示的程度（与始终具有 1.0 权重的完成相比）。增大此值可在完成较短时为训练添加稳定效果。

有关这些超参数的详细信息，请参阅 REST API 文档的[创建微调作业部分](#)。

选择默认或高级选项后，选择“下一步”以[查看所做的选择并训练已微调的模型](#)。

查看所做的选择并训练模型

向导的“查看并训练”窗格显示有关在“创建自定义模型”向导中为已微调模型所做的选择的信息，如下图所示。

Create customized model X

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

Review and train

Base model: davinci
Training data: training.jsonl
Validation data: validation.jsonl

BackSave and close

如果已准备好训练模型，请选择“保存并关闭”以启动微调作业并返回“[模型](#)”页。

检查自定义模型的状态

“模型”页中的“自定义模型”选项卡显示有关自定义模型的信息，如下图所示。该选项卡包含有关自定义模型微调作业的状态和作业 ID 的信息。作业完成后，还会显示结果文件的文件 ID。

Azure OpenAI Studio > Models Privacy & cookies

Azure OpenAI

Try it out

Playground

Management

Deployments

Models

File Management

Models

Azure OpenAI is powered by models with different capabilities and price points. Deploy one of the provided base models to try it out in [Playground](#) or train a custom model to your specific use case and data for better performance and more accurate results.

[Learn more about the different types of provided models](#)

Customized models Provided models

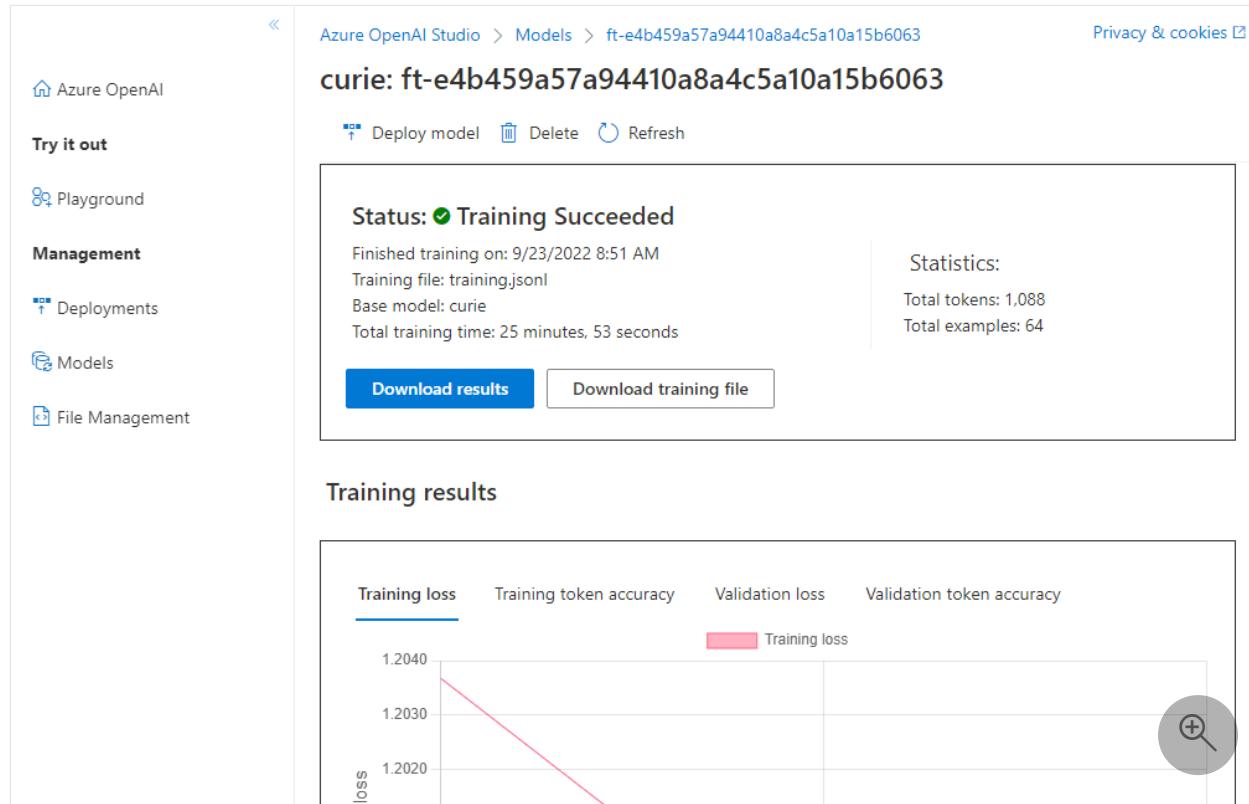
Deploy model Create customized model Delete Refresh

Model name	Create...	Base model	Status	Training job ID
ft-66aa4cc216694	9/7/2...	davinci	Running	ft-66aa4cc2166949beae9facb7f258510b

开始微调作业后，可能需要一些时间才能完成。该作业可能会排在我们系统上其他作业的后面，训练模型可能需要几分钟或几个小时，具体取决于模型和数据集的大小。可以

在“模型”页上的“自定义模型”选项卡的“状态”列中检查自定义模型的微调作业状态，并可以选择“刷新”以更新该页上的信息。

还可以从“模型”页的“模型名称”列中选择模型的名称，以显示有关自定义模型的详细信息，包括微调作业的状态、训练结果、训练事件，以及作业中使用的超参数。可以选择“刷新”按钮来刷新模型的信息，如下图所示。



The screenshot shows the Azure OpenAI Studio interface. On the left, there's a sidebar with links: 'Azure OpenAI', 'Try it out', 'Playground', 'Management', 'Deployments', 'Models', and 'File Management'. The main content area shows a model named 'curie: ft-e4b459a57a94410a8a4c5a10a15b6063'. The status is 'Training Succeeded' with a green checkmark. It shows the training finished on 9/23/2022 at 8:51 AM, using 'training.jsonl' as the training file, a 'curie' base model, and a total training time of 25 minutes, 53 seconds. Statistics include 1,088 total tokens and 64 total examples. Two buttons are visible: 'Download results' (blue) and 'Download training file' (white). Below this, a section titled 'Training results' contains a line chart for 'Training loss' (red line) over time, showing a downward trend from 1.2040 to 1.2020. A magnifying glass icon is in the bottom right corner of the chart area.

在模型页中，还可以选择“下载训练文件”以下载用于模型的训练数据，或选择“下载结果”以下载附加到模型微调作业的结果文件，并分析自定义模型的训练和验证性能。

部署自定义模型

微调作业成功后，可以从“模型”窗格部署自定义模型。必须部署自定义模型才能使其可用于完成调用。

① 备注

自定义模型只允许一个部署。如果选择已部署的自定义模型，会显示错误消息。

若要部署自定义模型，请选择该自定义模型，然后选择“部署模型”，如下图所示。

Azure OpenAI Studio > Models

Models

Azure OpenAI is powered by models with different capabilities and price points. Deploy one of the provided base models to try it out in [Playground](#) or train a custom model to your specific use case and data for better performance and more accurate results.

[Learn more about the different types of provided models](#)

Customized models Provided models

Deploy model Create customized model Delete Refresh

Model name Creat... Base model Status Training job Id Result file Id

curie.ft-ad0ee2d6	9/12/...	curie	Succeeded	ft-ad0ee2d614034bcfaf0043d3b89584f7	file-6bebba158a82452
-------------------	----------	-------	-----------	-------------------------------------	----------------------

Search

此时会显示“部署模型”对话框，可在其中为自定义模型的部署提供名称。在“部署名称”中输入名称，然后选择“创建”开始部署自定义模型。

Deploy model

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

Only one deployment is permitted per model. The models with existing deployments are disabled.

Model name

curie.ft-ad0ee2d614034bcfaf0043d3b89584f7

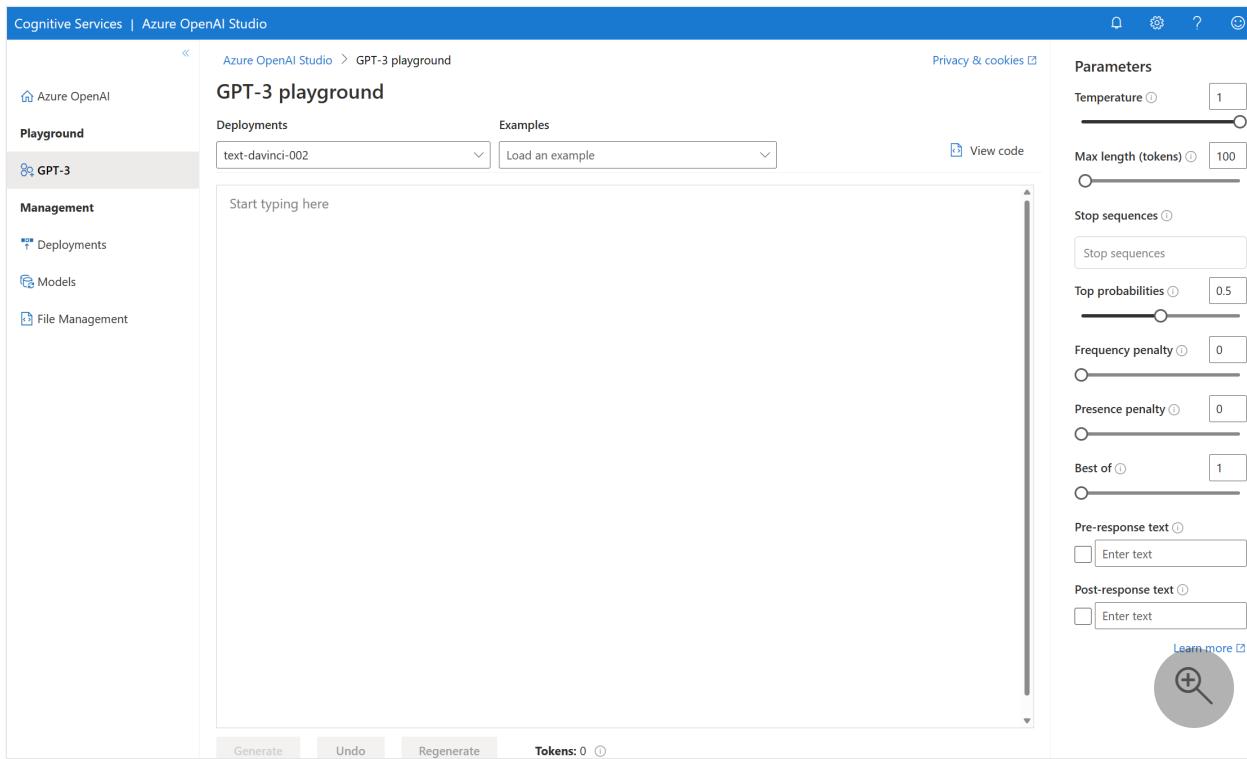
Deployment name

Create Cancel

可以在 Azure OpenAI Studio 的“部署”窗格中监视部署进度。

使用已部署的自定义模型

部署自定义模型后，可以像使用任何其他已部署的模型一样使用它。例如，可以使用 Azure OpenAI Studio 的“操场”窗格来试验新部署，如下图所示。可以继续对自定义模型使用相同的参数，例如温度和频率损失，就像对其他已部署的模型一样。



① 备注

与所有应用程序一样，我们需要在上线前完成评审过程。

分析自定义模型

完成每个微调作业后，Azure OpenAI 会将一个名为 `results.csv` 的结果文件附加到该作业。可以使用该结果文件来分析自定义模型的训练和验证性能。Azure OpenAI Studio 的“模型”窗格的“结果文件 ID”列中，列出了每个自定义模型的结果文件的文件 ID。可以使用文件 ID 从 Azure OpenAI Studio 的“文件管理”窗格中识别和下载结果文件。

结果文件是一个 CSV 文件，其中包含标题行，以及微调作业执行的每个训练步骤的行。结果文件包含以下列：

列名称	说明
<code>step</code>	训练步骤数。一个训练步骤代表针对一批训练数据前向和后向执行的一轮操作。
<code>elapsed_tokens</code>	自定义模型到目前为止看到的标记数，包括重复项。
<code>elapsed_examples</code>	模型到目前为止看到的示例数，包括重复项。 每个示例代表该步骤的训练数据批中的一个元素。例如，如果“高级选项”窗格中的“批大小”参数设置为 32，则此值在每个训练步骤中将加 32。

列名称	说明
<code>training_loss</code>	训练批的损失。
<code>training_sequence_accuracy</code>	训练批中模型预测标记与真实完成标记完全匹配的完成百分比。例如，如果批大小设置为 3 并且数据包含完成 <code>[[1, 2], [0, 5], [4, 2]]</code> ，在模型预测 <code>[[1, 1], [0, 5], [4, 2]]</code> 后，此值将设置为 0.67 (2/3)。
<code>training_token_accuracy</code>	训练批中模型正确预测的标记百分比。例如，如果批大小设置为 3 并且数据包含完成 <code>[[1, 2], [0, 5], [4, 2]]</code> ，在模型预测 <code>[[1, 1], [0, 5], [4, 2]]</code> 后，此值将设置为 0.83 (5/6)。
<code>validation_loss</code>	验证批的损失。
<code>validation_sequence_accuracy</code>	验证批中模型预测标记与真实完成标记完全匹配的完成百分比。例如，如果批大小设置为 3 并且数据包含完成 <code>[[1, 2], [0, 5], [4, 2]]</code> ，在模型预测 <code>[[1, 1], [0, 5], [4, 2]]</code> 后，此值将设置为 0.67 (2/3)。
<code>validation_token_accuracy</code>	验证批中模型正确预测的标记百分比。例如，如果批大小设置为 3 并且数据包含完成 <code>[[1, 2], [0, 5], [4, 2]]</code> ，在模型预测 <code>[[1, 1], [0, 5], [4, 2]]</code> 后，此值将设置为 0.83 (5/6)。

清理部署、自定义模型和训练文件

处理完自定义模型后，可以删除部署和模型。如果需要，还可以删除已上传到服务的训练和验证文件。

删除模型部署

可以在 Azure OpenAI Studio 的“部署”页中删除自定义模型的部署。选择要删除的部署，然后选择“删除”以删除部署。

删除自定义模型

可以在 Azure OpenAI Studio 的“模型”页中删除自定义模型。在“自定义模型”选项卡中选择要删除的自定义模型，然后选择“删除”以删除该自定义模型。

① 备注

如果自定义模型具有现有部署，则无法删除它。 必须先删除模型部署，然后才能删除自定义模型。

删除训练文件

可以选择性地在 Azure OpenAI Studio 的“文件管理”页中删除为训练上传的训练和验证文件，以及在训练期间生成的结果文件。 选择要删除的文件，然后选择“删除”以删除该文件。

后续步骤

- 浏览完整的 REST API 参考文档，了解有关所有微调功能的详细信息。 可在[此处](#)找到完整的 REST 文档。
- 可在[此处](#)详细了解 Python SDK 操作。

其他资源

文档

[如何准备用于自定义模型训练的数据集 - Azure OpenAI Service](#)

了解如何准备用于微调的数据集

[如何使用 Azure OpenAI 服务生成嵌入 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成嵌入

[Azure OpenAI 内容筛选 - Azure OpenAI](#)

了解 Azure 认知服务中 OpenAI 服务的内容筛选功能

[Azure OpenAI 服务嵌入教程 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 嵌入 API 对 BillSum 数据集进行文档搜索

[Models - List - REST API \(Azure Cognitive Services\)](#)

获取 Azure OpenAI 资源可访问的所有模型的列表。 其中包括基础模型以及所有成功完成的微调模型

[如何使用 Azure OpenAI 生成文本 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成或操作文本，包括代码

[Azure OpenAI 服务嵌入 - Azure OpenAI - embeddings and cosine similarity](#)

详细了解用于执行文档搜索以及获取余弦相似性的 Azure OpenAI 嵌入 API

[Azure OpenAI 模型 - Azure OpenAI](#)

了解 Azure OpenAI 中提供的不同模型。

[显示另外 5 个](#)

培训

学习路径和模块

[使用 Azure 机器学习训练机器学习模型 - Training](#)

使用 Azure 机器学习训练机器学习模型

学习证书

[经 Microsoft 认证 : Azure 数据科学家助理 - Certifications](#)

Azure 数据科学家具备应用数据科学和机器学习以在 Azure 上实施和运行机器学习工作负载的学科专业知识。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

如何使用托管标识配置 Azure OpenAI 服务

项目 • 2023/02/23

较复杂的安全方案需要 Azure 基于角色的访问控制 (Azure RBAC)。本文档介绍如何使用 Azure Active Directory (Azure AD) 对 OpenAI 资源进行身份验证。

在以下部分，你将使用 Azure CLI 分配角色，并获取持有者令牌来调用 OpenAI 资源。如果遇到问题，每个部分都提供了可供参考和使用的链接以及 Azure Cloud Shell/Azure CLI 中每个命令的所有可用选项。

先决条件

- Azure 订阅 - [免费创建订阅](#)
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 服务的访问权限

目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。

- Azure CLI - [安装指南](#)
- 以下 Python 库：os、requests、json

登录到 Azure CLI

若要登录到 Azure CLI，请运行以下命令并完成登录。如果会话空闲时间过长，可能需要再次执行此操作。

Azure CLI

```
az login
```

将你自己分配到认知服务用户角色

将自己分配到认知服务用户角色可以使用自己的帐户来访问特定的认知服务资源

1. 获取用户信息

Azure CLI

```
export user=$(az account show | jq -r .user.name)
```

2. 将你自己分配到“认知服务用户”角色。

Azure CLI

```
export resourceId=$(az group show -g $myResourceGroupName | jq -r .id)
az role assignment create --role "Cognitive Services User" --assignee
$user --scope $resourceId
```

① 备注

角色分配更改将在大约 5 分钟后生效。因此我提前执行了此步骤。如果你已事先完成此步骤，请跳过此步骤。

3. 获取 Azure AD 访问令牌。访问令牌将在一小时后过期。然后需要再次获取一个令牌。

Azure CLI

```
export accessToken=$(az account get-access-token --resource
https://cognitiveservices.azure.com | jq -r .accessToken)
```

4. 发出 API 调用：使用访问令牌通过设置 `Authorization` 标头值来授权 API 调用。

Bash

```
curl ${endpoint%}/openai/deployment/YOUR_DEPLOYMENT_NAME/completions?
api-version=2022-12-01 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $accessToken" \
-d '{ "prompt": "Once upon a time" }'
```

授权访问托管标识

OpenAI 支持使用 [Azure 资源托管标识](#) 进行 Azure Active Directory (Azure AD) 身份验证。 Azure 资源的托管标识可以从 Azure 虚拟机 (VM)、函数应用、虚拟机规模集和其他

服务中运行的应用程序使用 Azure AD 凭据授权对认知服务资源的访问权限。 将 Azure 资源的托管标识与 Azure AD 身份验证结合使用，可避免将凭据随在云中运行的应用程序一起存储。

在 VM 上启用托管标识

在使用 Azure 资源的托管标识对 VM 中的认知服务资源授予访问权限之前，必须在 VM 上启用 Azure 资源的托管标识。 若要了解如何为 Azure 资源启用托管标识，请参阅：

- [Azure 门户](#)
- [Azure PowerShell](#)
- [Azure CLI](#)
- [Azure Resource Manager 模板](#)
- [Azure 资源管理器客户端库](#)

有关托管标识的详细信息，请参阅 [Azure 资源的托管标识](#)。

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

配合使用 Azure OpenAI 与大型数据集

项目 • 2023/02/23

Azure OpenAI 可用于通过提示完成 API 解决大量自然语言任务。为了更轻松地将提示工作流从几个示例扩展到大型示例数据集，我们将 Azure OpenAI 服务与分布式机器学习库 [SynapseML](#) 集成。通过这种集成，可以轻松使用 [Apache Spark](#) 分布式计算框架，通过 OpenAI 服务处理数百万个提示。本教程介绍如何使用 Azure Open AI 和 Azure Synapse Analytics 在分布范围内应用大型语言模型。

先决条件

- Azure 订阅 - [免费创建订阅](#)
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 的访问权限

目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。

- Azure OpenAI 资源 - [创建资源](#)
- 安装了 SynapseML 的 Apache Spark 群集 - [在此处](#)创建无服务器 Apache Spark 池

建议[创建 Synapse 工作区](#)，但 Azure Databricks、HDInsight 或 Kubernetes 上的 Spark，甚至是带 `pyspark` 包的 Python 环境也会起作用。

以笔记本的形式导入此指南

下一步是将此代码添加到 Spark 群集。可以在 Spark 平台上创建笔记本并将代码复制到此笔记本中以运行演示，也可以下载笔记本并将其导入 Synapse Analytics。

1. [将此演示下载为笔记本](#)（单击“原始”，然后保存文件）
2. 将笔记本导入 [Synapse 工作区](#)，或如果使用的是 Databricks，则导入 [Databricks 工作区](#)
3. 在群集上安装 SynapseML。请参阅 [SynapseML 网站](#) 底部的 Synapse 安装说明。这需要在导入的笔记本顶部粘贴另一个单元格

4. 将笔记本连接到群集，然后继续编辑并运行以下单元格。

填写服务信息

接下来，编辑笔记本中的单元格以指向服务。具体而言，就是将 `resource_name`、`deployment_name`、`location` 和 `key` 变量设置为 Azure OpenAI 资源的相应值。

① 重要

完成后，请记住将密钥从代码中删除，并且永远不要公开展示该密钥。对于生产，请使用安全的方式存储和访问凭据，例如 [Azure Key Vault](#)。有关详细信息，请参阅 [认知服务安全性](#) 文章。

Python

```
import os

# Replace the following values with your Azure OpenAI resource information
resource_name = "RESOURCE_NAME"          # The name of your Azure OpenAI
resource.
deployment_name = "DEPLOYMENT_NAME"      # The name of your Azure OpenAI
deployment.
location = "RESOURCE_LOCATION"          # The location or region ID for your
resource.
key = "RESOURCE_API_KEY"                # The key for your resource.

assert key is not None and resource_name is not None
```

创建提示数据集

接下来，创建一个由一系列行组成的数据帧，每行有一个提示。

还可以直接从 Azure Data Lake Storage (ADLS) 或其他数据库加载数据。有关加载和准备 Spark 数据帧的详细信息，请参阅 [Apache Spark 数据加载指南](#)。

Python

```
df = spark.createDataFrame(
[
    ("Hello my name is",),
    ("The best code is code that's",),
    ("SynapseML is ",),
]
).toDF("prompt")
```

创建 OpenAICompletion Apache Spark 客户端

若要将 OpenAI Completion 服务应用于刚创建的数据帧，请创建作为分布式客户端的 `OpenAICompletion` 对象。 服务参数可以使用单个值设置，也可以使用 `OpenAICompletion` 对象上的适当资源库通过数据帧列设置。 在此，将 `maxTokens` 设置为 200。 令牌大约是四个字符，并且此限制适用于提示和结果的总和。 我们还使用数据帧中提示列的名称设置 `promptCol` 参数。

Python

```
from synapse.ml.cognitive import OpenAICompletion

completion = (
    OpenAICompletion()
    .setSubscriptionKey(key)
    .setDeploymentName(deployment_name)
    .setUrl("https://{}.openai.azure.com/".format(resource_name))
    .setMaxTokens(200)
    .setPromptCol("prompt")
    .setErrorCol("error")
    .setOutputCol("completions")
)
```

使用 OpenAICompletion 客户端转换数据帧

现在已经有了数据帧和完成客户端，可以转换输入数据集并添加一个名为 `completions` 的列，该列包含服务添加的所有信息。 为了简单起见，我们只选择文本。

Python

```
from pyspark.sql.functions import col

completed_df = completion.transform(df).cache()
display(completed_df.select(
    col("prompt"), col("error"),
    col("completions.choices.text").getItem(0).alias("text")))
```

输出应类似于以下示例；请注意，完成文本可能会有所不同。

prompt	error	text
Hello my name is	undefined	Makaveli I'm eighteen years old and I want to be a rapper when I grow up I love writing and making music I'm from Los Angeles, CA

prompt	error	text
最佳代码是	undefined	可以理解的代码。这是一个主观语句， 没有明确的答案。
SynapseML 是	undefined	一种机器学习算法，能够学习如何预测事件的未来结果。

其他用法示例

通过请求批处理提高吞吐量

以上示例向服务发出多个请求，每个提示一个请求。 若要在单个请求中完成多个提示，请使用批处理模式。 首先，在 `OpenAICompletion` 对象中，不要将 `Prompt` 列设置为“`Prompt`”，而是为 `BatchPrompt` 列指定“`batchPrompt`”。 为此，请创建一个数据帧，其中包含每行的提示列表。

① 备注

目前，单个请求中的提示限制为 20 个，“令牌”限制为 2048 个（大约 1500 字）。

Python

```
batch_df = spark.createDataFrame(
    [
        (["The time has come", "Pleased to", "Today stocks", "Here's to"],),
        (["The only thing", "Ask not what", "Every litter", "I am"],),
    ]
).toDF("batchPrompt")
```

接下来创建 `OpenAICompletion` 对象。 如果你的列为 `Array[String]` 类型，请设置 `batchPrompt` 列，而不是 `prompt` 列。

Python

```
batch_completion = (
    OpenAICompletion()
    .setSubscriptionKey(key)
    .setDeploymentName(deployment_name)
    .setUrl("https://{}.openai.azure.com/".format(resource_name))
    .setMaxTokens(200)
    .setBatchPromptCol("batchPrompt")
    .setErrorCol("error")
    .setOutputCol("completions")
)
```

在要转换的调用中，请求将按行发出。因为单行中有多个提示，所以每个请求都将与该行中的所有提示一起发送。结果将包含请求中的每一行的行。

Python

```
completed_batch_df = batch_completion.transform(batch_df).cache()  
display(completed_batch_df)
```

① 备注

目前，单个请求中的提示限制为 20 个，“令牌”限制为 2048 个（大约 1500 字）。

使用自动微型批处理程序

如果数据是列格式，可以使用 SynapseML 的 `FixedMiniBatcherTransformer` 将其转换为行格式。

Python

```
from pyspark.sql.types import StringType  
from synapse.ml.stages import FixedMiniBatchTransformer  
from synapse.ml.core.spark import FluentAPI  
  
completed_autobatch_df = (df  
    .coalesce(1) # Force a single partition so that our little 4-row dataframe  
    makes a batch of size 4, you can remove this step for large datasets  
    .mlTransform(FixedMiniBatchTransformer(batchSize=4))  
    .withColumnRenamed("prompt", "batchPrompt")  
    .mlTransform(batch_completion))  
  
display(completed_autobatch_df)
```

翻译的提示工程

Azure OpenAI 可以通过[提示工程](#)解决许多不同的自然语言任务。此处显示了语言翻译提示的示例：

Python

```
translate_df = spark.createDataFrame(  
    [  
        ("Japanese: Ookina hako \nEnglish: Big box \nJapanese: Midori  
tako\nEnglish:"),  
        ("French: Quelle heure est-il à Montréal? \nEnglish: What time is it  
in Montreal? \nFrench: Où est le poulet? \nEnglish:"),  
    ]
```

```
    ]  
).toDF("prompt")  
  
display(completion.transform(translate_df))
```

问题解答提示

此处提示 GPT-3 模型进行常识问题解答：

Python

```
qa_df = spark.createDataFrame(  
    [  
        (  
            "Q: Where is the Grand Canyon?\nA: The Grand Canyon is in  
Arizona.\n\nQ: What is the weight of the Burj Khalifa in kilograms?\nA:",  
            )  
    ]  
).toDF("prompt")  
  
display(completion.transform(qa_df))
```

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务静态数据加密

项目 · 2023/02/23

在将数据保存到云时，Azure OpenAI 会自动加密数据。加密可保护数据，并帮助你履行组织的安全性和合规性承诺。本文介绍 Azure OpenAI 如何处理静态数据加密，特别是训练数据和微调模型。有关如何处理、使用和存储你提供给服务的数据的信息，请参阅[数据、隐私和安全](#)一文。

关于认知服务加密

Azure OpenAI 是 Azure 认知服务的一部分。认知服务数据将使用符合 [FIPS 140-2](#) 的 [256 位 AES](#) 加密法进行加密和解密。加密和解密都是透明的，这意味着将替你管理加密和访问。你的数据默认情况下就是安全的，你无需修改代码或应用程序，即可利用加密。

关于加密密钥管理

默认情况下，订阅使用 Microsoft 托管的加密密钥。还有一个选项可用于通过你自己的密钥（称为“客户管理的密钥”(CMK)）来管理订阅。使用 CMK，可更加灵活地创建、轮换、禁用和撤销访问控制。此外，你还可以审核用于保护数据的加密密钥。

客户管理的密钥和 Azure Key Vault

客户管理的密钥 (CMK)（也称为创建自己的密钥 (BYOK)）在创建、轮换、禁用和撤销访问控制方面具有更大的灵活性。此外，你还可以审核用于保护数据的加密密钥。

必须使用 Azure Key Vault 来存储客户管理的密钥。可以创建自己的密钥并将其存储在 Key Vault 中，或者使用 Azure Key Vault API 来生成密钥。认知服务资源和密钥保管库必须在同一个区域和同一个 Azure Active Directory (Azure AD) 租户中，但可以在不同的订阅中。有关 Azure Key Vault 的详细信息，请参阅[什么是 Azure Key Vault？](#)。

要申请使用客户管理的密钥的权限，请填写并提交[认知服务客户管理的密钥申请表](#)。你大约需要 3-5 个工作日才能收到关于请求状态的回复。

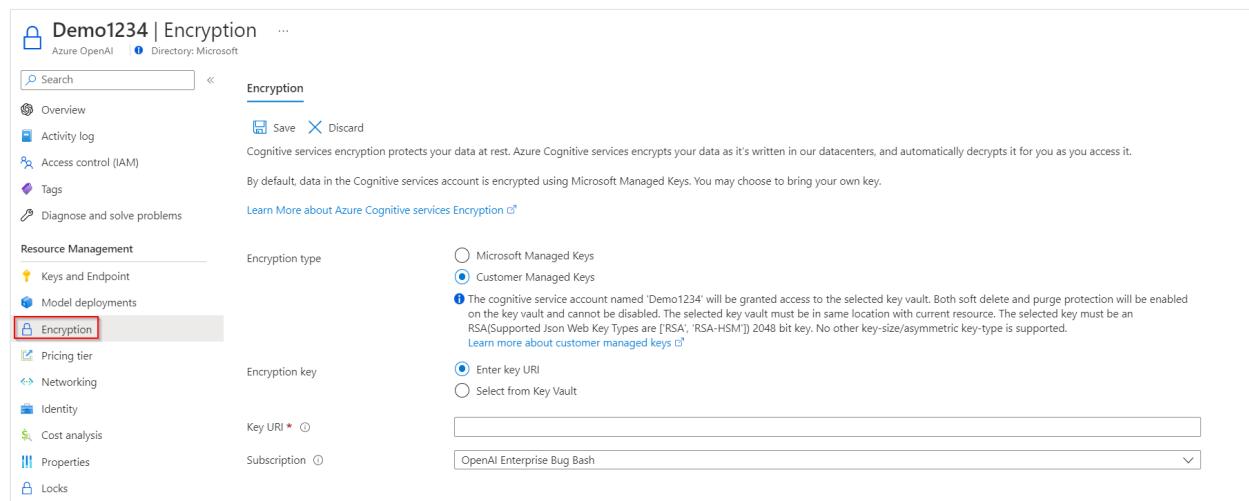
若要启用客户管理的密钥，还必须同时启用密钥保管库上的“软删除”和“不清除”属性。

认知服务加密只支持大小为 2048 的 RSA 密钥。有关密钥的详细信息，请参阅[关于 Azure Key Vault 密钥、机密和证书](#)中的“Key Vault 密钥”。

为你的资源启用客户管理的密钥

若要在 Azure 门户中启用客户管理的密钥，请执行以下步骤：

1. 转到认知服务资源。
2. 在左侧选择“加密”。
3. 在“加密类型”下选择“客户管理的密钥”，如以下屏幕截图所示。



指定密钥

启用客户管理的密钥后，可以指定要与认知服务资源关联的密钥。

将密钥指定为 URI

若要将某个密钥指定为 URI，请执行下列步骤：

1. 在 Azure 门户中转到你的密钥保管库。
2. 在“设置”下，选择“密钥”。
3. 选择所需的密钥，然后选择该密钥以查看其版本。选择一个密钥版本，查看该版本的设置。
4. 复制“密钥标识符”值，其中提供了 URI。

Properties

Key Type RSA

RSA Key Size 2048

Created 4/9/2019, 12:50:38 PM

Updated 4/9/2019, 12:50:38 PM

Key Identifier

<key-uri>

Settings

Set activation date?

Set expiration date?

Enabled? Yes No

Tags

0 tags

Permitted operations

Encrypt Sign Wrap Key

Decrypt Verify Unwrap Key

5. 返回你的认知服务资源，然后选择“加密”。

6. 在“加密密钥”下，选择“输入密钥 URI”。

7. 将复制的 URI 粘贴到“密钥 URI”框中。

CMK-Test - Encryption

Encryption

Save Discard

Cognitive services encryption protects your data at rest. Azure Cognitive services encrypts your data as it's written in our datacenters, and automatically decrypts it for you as you access it.

By default, data in the cognitive service account is encrypted using Microsoft Managed Keys. You may choose to bring your own key.

Please note that after enabling Cognitive Service Encryption, only new data will be encrypted, and any existing files in this cognitive service account will retroactively get encrypted by a background encryption process.

Learn More about Azure Cognitive services Encryption [↗](#)

Encryption type

Microsoft Managed Keys Customer Managed Keys

The cognitive service account named 'CMK-Test' will be granted access to the selected key vault. Both soft delete and purge protection will be enabled on the key vault and cannot be disabled.

Encryption key

Enter key URI Select from Key Vault

Key URI *

Subscription [ⓘ](#) AICP-DEV

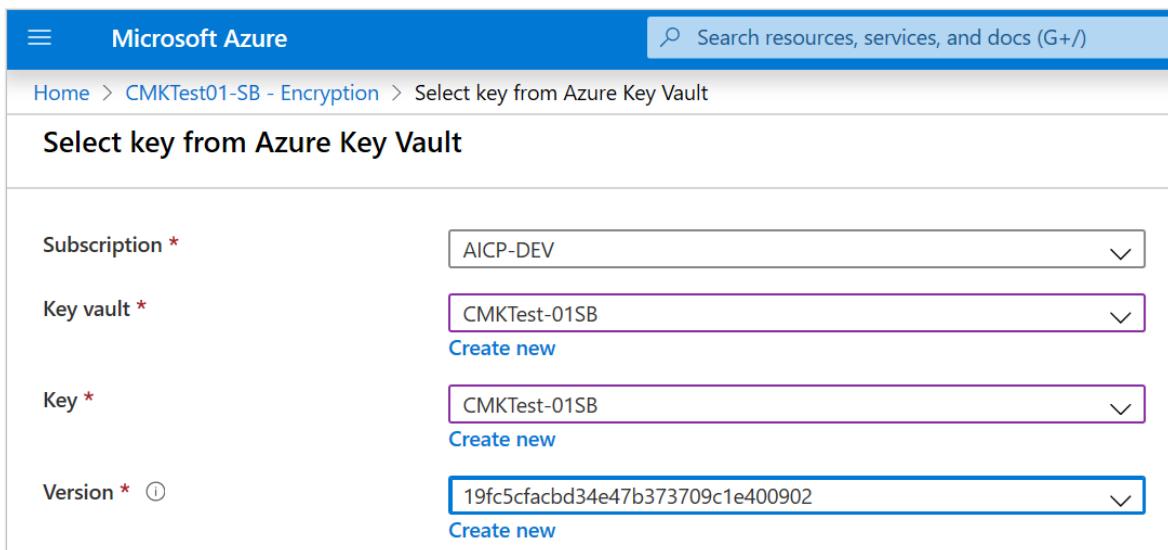
8. 在“订阅”下，选择包含密钥保管库的订阅。

9. 保存所做更改。

从 Key Vault 指定密钥

若要指定 Key Vault 中的密钥，请先请确保有一个包含密钥的 Key Vault。然后，执行以下步骤：

1. 返回你的认知服务资源，然后选择“加密”。
2. 在“加密密钥”下，选择“从 Key Vault 中选择”。
3. 选择包含要使用的密钥的密钥保管库。
4. 选择要使用的密钥。



Subscription * AICP-DEV

Key vault * CMKTest-01SB
Create new

Key * CMKTest-01SB
Create new

Version * 19fc5cfacbd34e47b373709c1e400902
Create new

5. 保存所做更改。

更新密钥版本

创建密钥的新版本时，请更新认知服务资源以使用新版本。执行以下步骤：

1. 返回你的认知服务资源，然后选择“加密”。
2. 输入新密钥版本的 URI。或者，可以选择密钥保管库，然后再次选择密钥以更新版本。
3. 保存所做更改。

使用其他密钥

若要更改用于加密的密钥，请执行以下步骤：

1. 返回你的认知服务资源，然后选择“加密”。
2. 输入新密钥的 URI。或者，可以选择密钥保管库，然后选择一个新密钥。
3. 保存所做更改。

轮换客户管理的密钥

可以根据合规性策略轮换 Key Vault 中客户管理的密钥。 轮换密钥后，必须更新认知服务资源才能使用新的密钥 URI。 若要了解如何更新资源以在 Azure 门户中使用新版本的密钥，请参阅[更新密钥版本](#)。

轮换密钥不会触发资源中数据的重新加密。 用户无需进一步执行操作。

撤销客户管理的密钥

可以撤销客户管理的加密密钥，方法是更改访问策略，或更改针对密钥保管库的权限，或者删除该密钥。

若要更改注册表使用的托管标识的访问策略，请运行 [az-keyvault-delete-policy](#) 命令：

Azure CLI

```
az keyvault delete-policy \
--resource-group <resource-group-name> \
--name <key-vault-name> \
--key_id <key-vault-key-id>
```

若要删除密钥的单个版本，请运行 [az-keyvault-key-delete](#) 命令。 此操作需要 keys/delete 权限。

Azure CLI

```
az keyvault key delete \
--name <key-vault-name> \
--object-id $identityPrincipalID \
```

① 重要

在 CMK 仍处于启用状态时撤销对活动的客户管理的密钥的访问权限，会阻止下载训练数据和结果文件、微调新模型以及部署微调后的模型。 但是，以前部署的微调模型将继续运行并提供流量，直到删除这些部署。

删除训练、验证和训练结果数据

文件 API 允许客户上传其训练数据，以便微调模型。 此数据存储在与资源位于同一区域的 Azure 存储中，并在逻辑上与其 Azure 订阅和 API 凭据隔离。 用户可以通过[DELETE API 操作](#)删除上传的文件。

删除微调的模型和部署

使用微调 API，客户可以基于你通过文件 API 上传到服务的训练数据创建其自己的微调版 OpenAI 模型。已训练的微调模型存储在同一区域的 Azure 存储中，经过静态加密，并在逻辑上与其 Azure 订阅和 API 凭据相隔离。用户可以通过调用 [DELETE API 操作](#) 来删除经过微调的模型和部署。

禁用客户托管密钥

当你禁用客户管理的密钥后，系统会使用 Microsoft 管理的密钥加密认知服务资源。若要禁用客户托管密钥，请执行以下步骤：

1. 返回你的认知服务资源，然后选择“加密”。
2. 选择“Microsoft 托管密钥”>“保存”。

以前启用客户管理的密钥时，还会启用系统分配的托管标识（Azure AD 的一项功能）。启用系统分配的托管标识后，此资源将注册到 Azure Active Directory。注册后，将向托管标识授予在设置客户管理的密钥期间选择的 Key Vault 的访问权限。你可以详细了解[托管标识](#)。

① 重要

如果禁用系统分配的托管标识，则会删除对密钥保管库的访问权限，而使用客户密钥加密的任何数据都将不再可供访问。任何依赖于此数据的功能都会失效。

① 重要

托管标识当前不支持跨目录方案。在 Azure 门户中配置客户管理的密钥时，系统会在幕后自动分配一个托管标识。如果随后将订阅、资源组或资源从一个 Azure AD 目录移动到另一个目录，则与资源关联的托管标识不会转移到新租户，因此，客户管理的密钥可能不再有效。有关详细信息，请参阅 [Azure 资源的常见问题解答和已知问题](#) 中的“在 Azure AD 目录之间转移订阅”。

后续步骤

- [语言服务客户管理的密钥请求表单](#)
- [详细了解 Azure 密钥保管库](#)

其他资源

📖 文档

[如何使用 Azure OpenAI 自定义模型 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 创建自己的自定义模型

[如何使用托管标识配置 Azure OpenAI - Azure OpenAI](#)

提供有关如何使用 Azure Active Directory 设置托管标识的指导

[Azure OpenAI 内容筛选 - Azure OpenAI](#)

了解 Azure 认知服务中 OpenAI 服务的内容筛选功能

[如何使用 Azure OpenAI 生成嵌入 - Azure OpenAI](#)

了解如何使用 Azure OpenAI 生成嵌入

[如何准备用于自定义模型训练的数据集 - Azure OpenAI](#)

了解如何准备用于微调的数据集

[Azure OpenAI 嵌入 - Azure OpenAI - embeddings and cosine similarity](#)

详细了解用于执行文档搜索以及获取余弦相似性的 Azure OpenAI 嵌入 API

[Azure OpenAI 中的新增功能有哪些？ - Azure Cognitive Services](#)

了解 Azure OpenAI 的最新资讯和功能更新

[显示另外 4 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务的业务连续性和灾难恢复 (BCDR) 注意事项

项目 • 2023/02/24

Azure OpenAI 在两个区域中提供。由于订阅密钥绑定到区域，因此当客户获取密钥时，他们会选择部署所在的区域，从那时起，所有操作都与该 Azure 服务器区域保持关联。

有时我们会遇到影响整个区域的网络问题，这种情况比较罕见，但也不是没有可能。如果服务需要始终保持可用，则应将其设计为可故障转移到另一区域，或者将工作负载分散到两个或更多个区域。这两种方法都至少需要两个不同区域中的 OpenAI 资源。本文提供有关如何为 Azure OpenAI 应用程序实现业务连续性和灾难恢复 (BCDR) 的一般建议。

最佳实践

当今的客户在部署期间会调用提供的终结点进行部署和推理。这些操作是无状态的，因此在区域不可用的情况下不会丢失任何数据。

如果某个区域处于不正常运行状态，客户必须采取措施来确保服务连续性。

业务连续性

以下说明集适用于使用默认终结点的客户和使用自定义终结点的客户。

默认终结点恢复

如果使用默认终结点，则应将客户端代码配置为监视错误，如果错误仍然存在，请准备好重定向到你选择的另一 Azure OpenAI 订阅区域。

请按照以下步骤配置客户端以监视错误：

1. 使用本页查看 OpenAI 服务的可用区域列表。
2. 从列表中选择一个主要区域以及一个次要/备份区域。
3. 为所选的每个区域创建 OpenAI 服务资源

4. 对于主要区域和任何备份区域，代码需要知道：
 - a. 资源的基 URI
 - b. 区域访问密钥或 Azure Active Directory 访问权限
5. 配置代码以监视连接错误（常见错误为连接超时和服务不可用）。
 - a. 由于网络遇到暂时性错误，因此对于出现的个别连接问题，建议重试。
 - b. 为实现持久性，请将流量重定向到区域中你创建的备份资源。

BCDR 需要自定义代码

对于此使用类型，从区域性故障中恢复可以即时执行，而且成本非常低。但是，这需要在应用程序的客户端自定义开发此功能。

其他资源

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

监视 Azure OpenAI 服务

项目 • 2023/02/18

当你的关键应用程序和业务流程依赖于 Azure 资源时，你需要监视这些资源的可用性、性能和操作。

本文介绍 Azure OpenAI 服务生成的监视数据。 Azure OpenAI 是认知服务的一部分，它使用 [Azure Monitor](#)。 如果你不熟悉所有 Azure 服务普遍使用的 Azure Monitor 功能，请参阅[使用 Azure Monitor 监视 Azure 资源](#)。

监视数据

Azure OpenAI 收集与[监视 Azure 资源中的数据](#)中所述的其他 Azure 资源类型相同的监视数据。

收集和路由

平台指标和活动日志会自动收集和存储，但你可以使用诊断设置将其路由到其他位置。

在创建诊断设置并将其路由到一个或多个位置之前，不会收集和存储资源日志。

有关使用 Azure 门户、CLI 或 PowerShell 创建诊断设置的详细过程，请参阅[创建诊断设置以收集 Azure 中的平台日志和指标](#)。 创建诊断设置时，请指定要收集的日志类别。

请记住，使用诊断设置并将数据发送到 Azure Monitor 日志会产生额外费用。 若要了解详细信息，请参阅[Azure Monitor 成本计算指南](#)。

以下部分将讨论可以收集的指标和日志。

分析指标

在 Azure 门户中查看 Azure OpenAI 资源时，可以通过打开“监视”部分下的“指标”来分析 Azure OpenAI 的指标。 有关使用此工具的详细信息，请参阅[Azure 指标资源管理器入门](#)。

Azure OpenAI 是认知服务的一部分。有关为认知服务和 Azure OpenAI 收集的所有平台指标的列表，请参阅[认知服务支持的指标](#)。

对于 Azure OpenAI 中可用指标的当前子集：

Azure OpenAI 指标

指标	是否可通过诊断设置导出？	指标显示名称	计价单位	聚合类型	说明	维度
BlockedCalls	是	阻止的调用数	计数	总计	超过速率或配额限制的调用数。	ApiName、OperationName、Region、RatelimtKey
ClientErrors	是	客户端错误数	计数	总计	引发客户端错误 (HTTP 响应代码 4xx) 的调用数。	ApiName、OperationName、Region、RatelimtKey
DataIn	是	数据输入	字节	总计	传入数据的大小 (字节)。	ApiName、OperationName、Region
DataOut	是	Data Out	字节	总计	传出数据的大小 (字节)。	ApiName、OperationName、Region
FineTunedTrainingHours	是	已处理的 FineTuned 训练小时数	计数	总计	在 OpenAI FineTuned 模型中处理的训练小时数	ApiName, ModelDeploymentName, FeatureName, UsageChannel, Region
延迟	是	延迟	毫秒	平均值	延迟 (毫秒)。	ApiName、OperationName、Region、RatelimtKey

指标	是否可通	指标显示名称	计价单位	聚合类型	说明	维度
	过诊 断设 置导 出？					
Ratelimt	是	Ratelimt	计数	总计	ratelimit键的当前速率限制。	Region、RatelimtKey
ServerErrors	是	服务器错误数	计数	总计	引发服务内部错误 (HTTP 响应代码 5xx) 的调用数。	ApiName、OperationName、Region、RatelimtKey
SuccessfulCalls	是	成功调用数	计数	总计	成功调用数。	ApiName、OperationName、Region、RatelimtKey
TokenTransaction	是	已处理的推理令牌	计数	总计	在 OpenAI 模型中处理的推理令牌数	ApiName, ModelDeploymentName, FeatureName, UsageChannel, Region
TotalCalls	是	总调用数	Count	总计	调用总数。	ApiName、OperationName、Region、RatelimtKey
TotalErrors	是	错误总数	Count	总计	引发错误响应 (HTTP 响应代码 4xx 或 5xx) 的调用总数。	ApiName、OperationName、Region、RatelimtKey

分析日志

Azure Monitor 日志中的数据以表形式存储，每个表具有自己独有的属性集。

Azure Monitor 中的所有资源日志都具有后跟服务特定字段的相同字段。 [Azure Monitor 资源日志架构](#)概述了常见架构。

活动日志是 Azure 中的一种平台日志，可用于深入了解订阅级别的事件。你可以单独查看它或将它路由到 Azure Monitor 日志，然后便可以在其中使用 Log Analytics 执行复杂的查询。

有关可用于 Azure OpenAI 和其他认知服务的资源日志类型的列表，请参阅[认知服务的资源提供程序操作](#)

Kusto 查询

① 重要

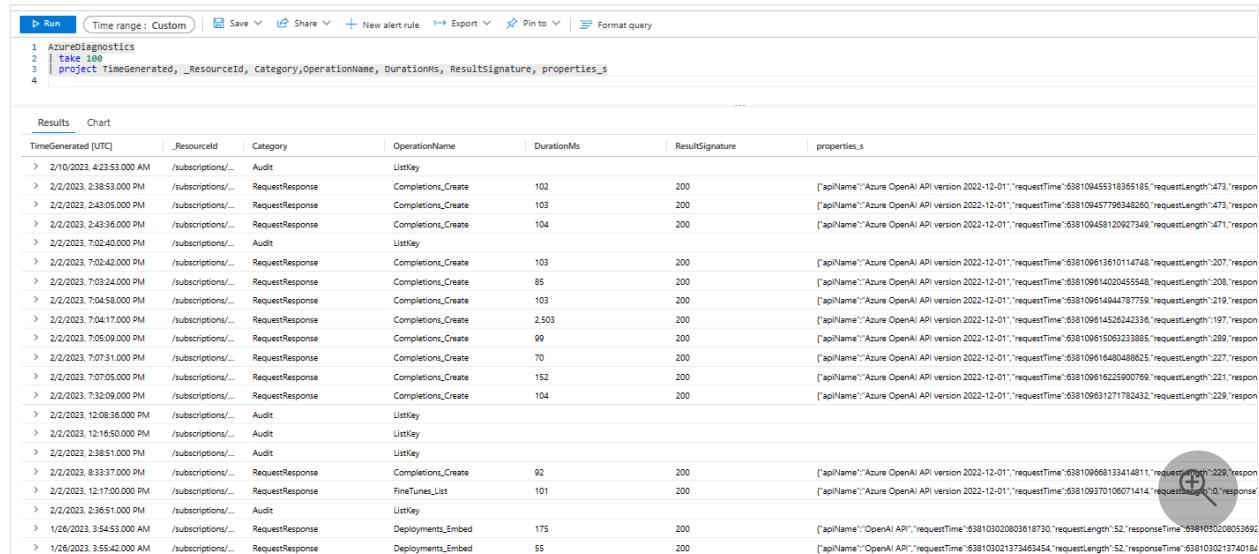
在“Azure OpenAI”菜单中选择“日志”时，Log Analytics 随即打开，其查询范围设置为当前 Azure OpenAI 资源。这意味着日志查询只包含来自该资源的数据。如果要运行的查询包含来自其他资源或其他 Azure 服务的数据，请从“Azure Monitor”菜单中选择“日志”。有关详细信息，请参阅 [Azure Monitor Log Analytics 中的日志查询范围和时间范围](#)。

若要浏览并了解可用于 Azure OpenAI 资源的信息类型，在部署模型并通过操场发送一些完成调用后，可以通过以下有用查询开始：

```
Kusto

AzureDiagnostics
| take 100
| project TimeGenerated, _ResourceId, Category, OperationName, DurationMs,
ResultSignature, properties_s
```

在这里，我们返回了包含 100 个条目的示例，并显示日志中可用数据列的子集。结果如下所示：



TimeGenerated (UTC)	_ResourceId	Category	OperationName	DurationMs	ResultSignature	properties_s																																																																																																																						
> 2/1/2023, 2:23:53.000 AM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 2/2/2023, 2:38:53.000 PM	/subscriptions/...	RequestResponse	Completions_Create	102	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096455318395185,"requestLength":473,"respon	> 2/2/2023, 2:43:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109645796348260,"requestLength":473,"respon	> 2/2/2023, 2:43:36.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096458120927349,"requestLength":471,"respon	> 2/2/2023, 7:02:40.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 7:02:42.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810965161011478,"requestLength":207,"respon	> 2/2/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614020455548,"requestLength":208,"respon	> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164
> 2/2/2023, 2:43:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109645796348260,"requestLength":473,"respon	> 2/2/2023, 2:43:36.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096458120927349,"requestLength":471,"respon	> 2/2/2023, 7:02:40.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 7:02:42.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810965161011478,"requestLength":207,"respon	> 2/2/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614020455548,"requestLength":208,"respon	> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164						
> 2/2/2023, 2:43:36.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096458120927349,"requestLength":471,"respon	> 2/2/2023, 7:02:40.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 7:02:42.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810965161011478,"requestLength":207,"respon	> 2/2/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614020455548,"requestLength":208,"respon	> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164												
> 2/2/2023, 7:02:40.000 PM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 2/2/2023, 7:02:42.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810965161011478,"requestLength":207,"respon	> 2/2/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614020455548,"requestLength":208,"respon	> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																									
> 2/2/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614020455548,"requestLength":208,"respon	> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																															
> 2/2/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":119,"respon	> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																					
> 2/2/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242338,"requestLength":197,"respon	> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																											
> 2/2/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109615065233885,"requestLength":289,"respon	> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																	
> 2/2/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096164808625,"requestLength":227,"respon	> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																							
> 2/2/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616223900769,"requestLength":221,"respon	> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																													
> 2/2/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109631271782432,"requestLength":229,"respon	> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200			> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																																			
> 2/2/2023, 12:09:36.000 PM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 2/2/2023, 12:16:50.000 PM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 2/2/2023, 2:38:51.000 PM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 2/2/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109668133414811,"requestLength":220,"respon	> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																																																														
> 2/2/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	["apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":0,"respon	> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200			> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																																																																				
> 2/2/2023, 2:36:51.000 PM	/subscriptions/...	Audit	listKey	200																																																																																																																								
> 1/26/2023, 3:55:43.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	["apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053369	> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																																																																																	
> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	["apiName":"OpenAI API","requestTime":638103021373463454,"requestLength":52,"responseTime":6381030213740164																																																																																																																							

如果想要查看所有可用的数据列，可以删除 `| project` 行提供的范围：

```
Kusto

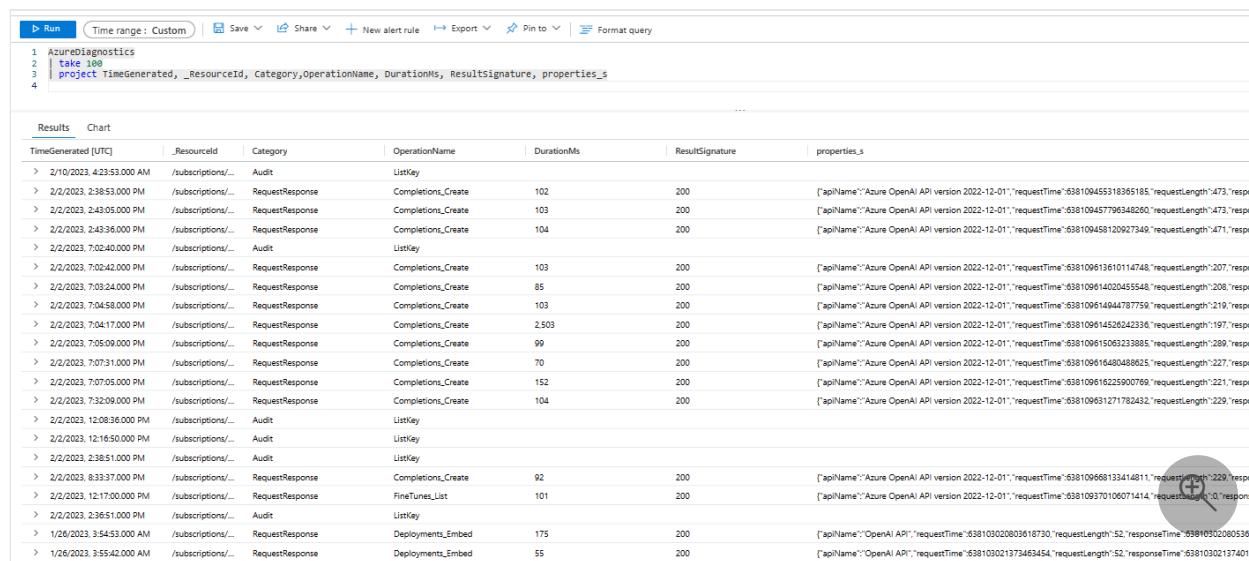
AzureDiagnostics
| take 100
```

还可以选择表名称旁边的箭头来查看所有可用列和关联的数据类型。

若要检查 AzureMetrics，请运行：

```
Kusto

AzureMetrics
| take 100
| project TimeGenerated, MetricName, Total, Count, TimeGrain, UnitName
```



TimeGenerated [UTC]	_ResourceId	Category	OperationName	DurationMs	ResultSignature	properties_s
> 2/1/2023, 4:23:33.000 AM	/subscriptions/...	Audit	LittleKey			
> 2/1/2023, 2:38:53.000 PM	/subscriptions/...	RequestResponse	Completions_Create	102	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810961402045548,"requestLength":473,"respon
> 2/1/2023, 2:43:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096140457796348260,"requestLength":473,"respon
> 2/1/2023, 2:43:36.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":6381096140927349,"requestLength":471,"respon
> 2/1/2023, 7:02:40.000 PM	/subscriptions/...	Audit	LittleKey			
> 2/1/2023, 7:02:42.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109613610114748,"requestLength":207,"respon
> 2/1/2023, 7:03:24.000 PM	/subscriptions/...	RequestResponse	Completions_Create	85	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810961402045548,"requestLength":208,"respon
> 2/1/2023, 7:04:58.000 PM	/subscriptions/...	RequestResponse	Completions_Create	103	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614044787759,"requestLength":219,"respon
> 2/1/2023, 7:04:17.000 PM	/subscriptions/...	RequestResponse	Completions_Create	2,503	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614526242336,"requestLength":197,"respon
> 2/1/2023, 7:05:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	99	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810961506233885,"requestLength":289,"respon
> 2/1/2023, 7:07:31.000 PM	/subscriptions/...	RequestResponse	Completions_Create	70	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109614480488052,"requestLength":227,"respon
> 2/1/2023, 7:07:05.000 PM	/subscriptions/...	RequestResponse	Completions_Create	152	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109616225900769,"requestLength":221,"respon
> 2/1/2023, 7:32:09.000 PM	/subscriptions/...	RequestResponse	Completions_Create	104	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810961271782432,"requestLength":229,"respon
> 2/1/2023, 12:08:36.000 PM	/subscriptions/...	Audit	LittleKey			
> 2/1/2023, 12:16:50.000 PM	/subscriptions/...	Audit	LittleKey			
> 2/1/2023, 2:38:51.000 PM	/subscriptions/...	Audit	LittleKey			
> 2/1/2023, 8:33:37.000 PM	/subscriptions/...	RequestResponse	Completions_Create	92	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":63810966133414811,"requestLength":229,"respon
> 2/1/2023, 12:17:00.000 PM	/subscriptions/...	RequestResponse	FineTunes_List	101	200	{"apiName":"Azure OpenAI API version 2022-12-01","requestTime":638109370106071414,"requestLength":10,"respon
> 2/1/2023, 2:36:51.000 PM	/subscriptions/...	Audit	LittleKey			
> 1/26/2023, 3:54:53.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	175	200	{"apiName":"OpenAI API","requestTime":638103020803618730,"requestLength":52,"responseTime":6381030208053692
> 1/26/2023, 3:55:42.000 AM	/subscriptions/...	RequestResponse	Deployments_Emb	55	200	{"apiName":"OpenAI API","requestTime":638103021373465454,"requestLength":52,"responseTime":6381030213740184}

警报

在监视数据中发现重要情况时，Azure Monitor 警报会主动通知你。有了警报，你就可以在客户注意到你的系统中的问题之前确定和解决它们。可以在[指标](#)、[日志](#)和[活动日志](#)上设置警报。不同类型的警报具有不同的优缺点。

每个组织的警报需求都会有所不同，而且还会随着时间的推移而变化。一般来说，所有警报都应具备可操作性，在发生警报时具有特定的预期响应。如果没有可执行的操作，则这可能是你想要在报表中而不是在警报中捕获的内容。某些用例可能需要在存在某些错误条件时发出警报。但在许多环境中，可能只有在错误超过某个阈值一段时间时，才会发出警报。

通常可以通过定期分析 Azure Monitor 日志中的数据来评估低于某些阈值的错误。在分析一段时间内的日志数据时，可能还会发现，在很长一段时间未发生的某个条件，对于使

用警报进行跟踪可能很有价值。有时，日志中事件的缺失与错误一样是重要的信号。

根据结合使用 Azure OpenAI 开发的应用程序类型，[Azure Monitor Application Insights](#) 可能会在应用程序层提供其他监视优势。

后续步骤

- 有关监视 Azure 资源的详细信息，请参阅[使用 Azure Monitor 监视 Azure 资源](#)。
 - 阅读[了解 Azure Monitor 日志中的日志搜索](#)。
-

其他资源

■ 文档

[Deployments - REST API \(Azure Cognitive Services\)](#)

详细了解 [认知服务部署操作]。如何[创建、删除、获取、列出、更新]。

[Fine Tunes - REST API \(Azure Cognitive Services\)](#)

详细了解 [认知服务微调操作]。如何 [取消，创建，删除，获取，获取事件，列表]。

[Deployments - REST API \(Azure Cognitive Services\)](#)

详细了解 [认知服务部署操作]。如何[创建、删除、获取、列出、更新]。

[Files - REST API \(Azure Cognitive Services\)](#)

详细了解 [认知服务文件操作]。如何[删除、获取、获取内容、导入、列表、上传]。

[自定义文本分类常见问题解答 - Azure Cognitive Services](#)

了解使用自定义文本分类 API 时的常见问题解答。

[如何标记数据以用于自定义分类 - Azure 认知服务 - Azure Cognitive Services](#)

了解如何标记数据以用于自定义文本分类。

[Fine Tunes - Create - REST API \(Azure Cognitive Services\)](#)

创建从给定训练文件微调指定模型的作业。响应包括排队作业的详细信息，包括作业状态和超参数

[快速入门：使用 SDK 创建和使用学习循环 - 个性化体验创建服务 - Azure Cognitive Services](#)

本快速入门介绍如何使用个性化体验创建服务客户端库创建和管理知识库。

[显示另外 5 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

计划管理 Azure OpenAI 服务的成本

项目 • 2023/03/02

本文介绍如何计划和管理 Azure OpenAI 服务的成本。在部署服务之前，可以使用 Azure 定价计算器估算 Azure OpenAI 的成本。之后在部署 Azure 资源时，查看预估成本。在开始使用 Azure OpenAI 资源之后，请使用“成本管理”功能来设置预算和监视成本。你还可以查看预测成本并确定支出趋势，以确定你可能想要对其采取措施的区域。Azure OpenAI 服务的成本仅是 Azure 帐单中每月成本的一部分。尽管本文介绍了如何为 Azure OpenAI 计划和管理成本，但你需要为 Azure 订阅中使用的所有 Azure 服务和资源（包括第三方服务）付费。

先决条件

成本管理中的成本分析支持大多数 Azure 帐户类型，但不支持所有帐户类型。若要查看支持的帐户类型的完整列表，请参阅[了解成本管理数据](#)。若要查看成本数据，你至少需要对 Azure 帐户具有读取访问权限。若要了解如何分配对 Azure 成本管理数据的访问权限，请参阅[分配对数据的访问权限](#)。

使用 Azure OpenAI 之前请估算成本

使用 [Azure 定价计算器](#) 估算使用 Azure OpenAI 的成本。

了解 Azure OpenAI 服务的完整计费模式

Azure OpenAI 服务在 Azure 基础结构上运行，部署新资源时，该基础结构会随之产生成本。用户务必了解，是否可能产生其他基础结构成本。

Azure OpenAI 服务的收费方式

基础系列和 Codex 系列模型

Azure OpenAI 基础系列和 Codex 系列模型每 1,000 个标记进行收费。成本因所选模型系列而异：Ada、Babbage、Curie、DaVinci 或 Code-Cushman。

我们的模型通过将文本分解为标记来理解和处理文本。作为参考，对于典型的英文文本，每个标记大约是四个字符。

令牌成本是针对输入和输出计算的。例如，如果你有 1,000 个令牌的 JavaScript 代码示例，且要求 Azure OpenAI 模型将代码转换为 Python。首先你要为发送的初始输入请求的 1,000 个令牌支付费用，然后还要针对接收的响应支付输出的 1,000 个令牌的费用，总共涉及的是 2,000 个令牌。

在完成这种类型的调用时，其实令牌的输入/输出不会是完全 1 : 1 的关系。在不同的编程语言间进行转换时，输出可能会变长或变短，具体取决于许多不同的因素，包括分配给 `max_tokens` 参数的值。

基本系列和 Codex 系列微调模型

Azure OpenAI 微调模型根据三个因素收费：

- 训练小时数
- 托管小时数
- 每 1,000 个标记的推理

了解托管小时数成本非常重要，因为一旦部署了微调模型，无论你是否主动使用它，每小时都会持续产生成本。应密切监视微调模型成本。

Azure OpenAI 服务可能产生的其他成本

请记住，启用将数据发送到 Azure Monitor 日志、发出警报等功能时，这些服务会产生额外的成本。这些成本在其他服务下和订阅级别可见，但如果范围仅限于 Azure OpenAI 资源，则不可见。

对 Azure OpenAI 服务使用 Azure 预付款

可以使用 Azure 预付款额度支付 Azure OpenAI 服务费用。但是，不能使用 Azure 预付款额度来支付第三方产品和服务（包括 Azure 市场中的）的费用。

监视成本

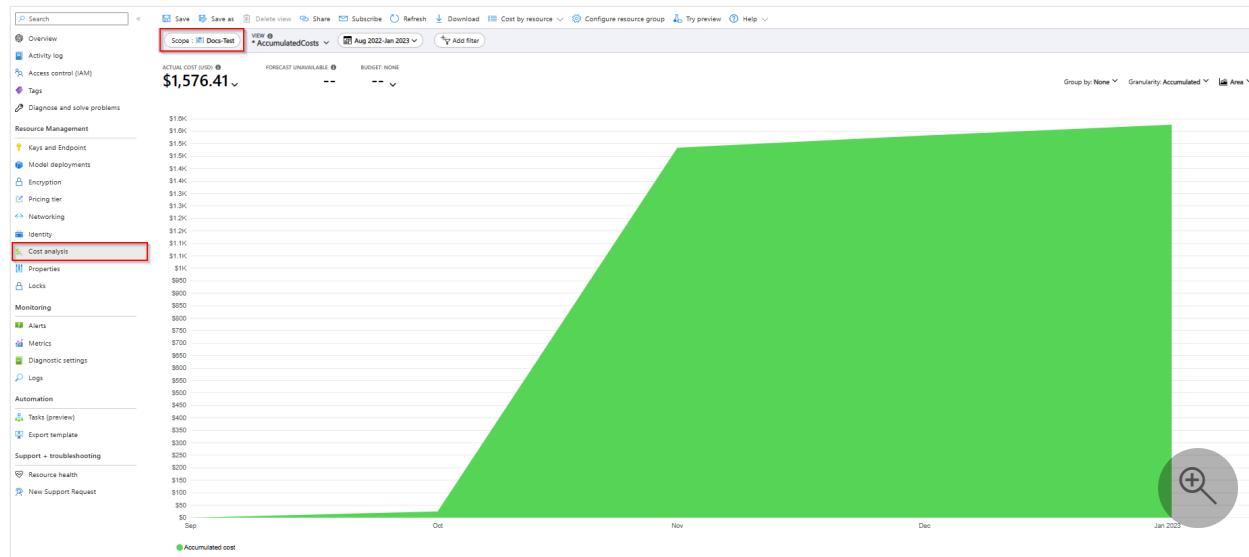
将 Azure 资源用于 Azure OpenAI 时，会产生成本。Azure 资源使用情况的单位成本因时间间隔（秒、分钟、小时和天）或单位使用情况（字节、MB 等）而异。一旦开始使用 Azure OpenAI，就会产生费用，你可以在[成本分析](#)中看到这些费用。

使用成本分析时，可以在关系图和表中查看不同时间间隔的 Azure OpenAI 成本。有些示例是按天、当前、上个月和年划分的。还可以对比预算和预测的成本来查看成本。随着

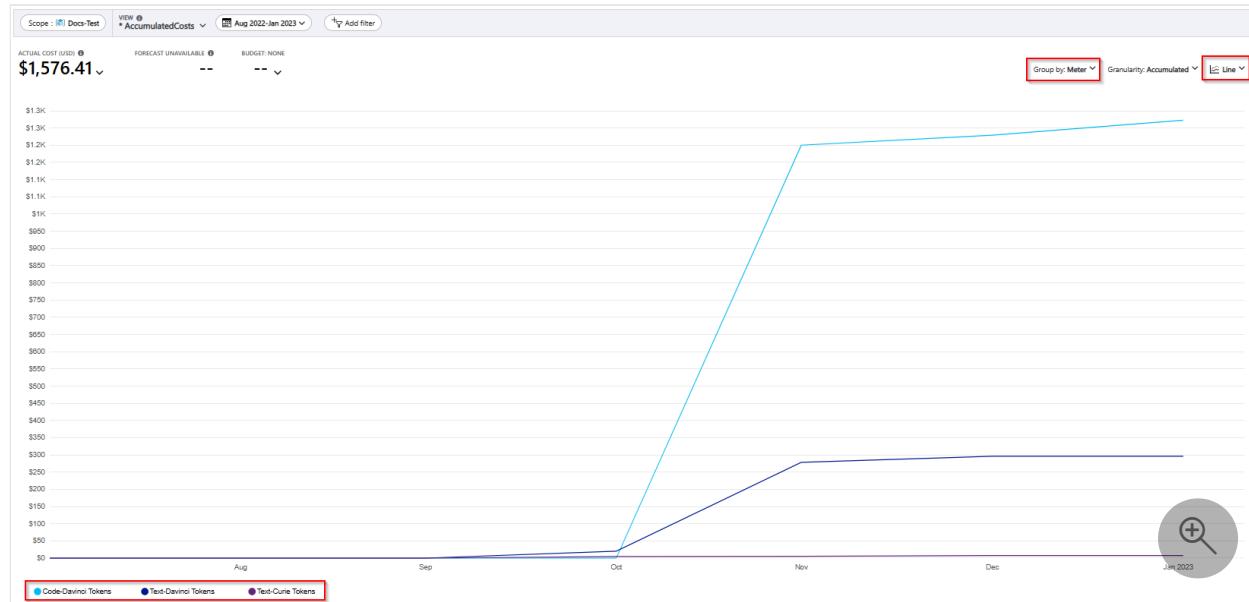
时间的推移切换到较长的视图，可帮助你确定支出趋势。你会看到可能出现超支的地方。如果已创建预算，还可以轻松查看超支的地方。

在成本分析中查看 Azure OpenAI 成本：

1. 登录到 Azure 门户。
2. 选择一个 Azure OpenAI 资源。
3. 在“资源管理”下，选择“成本分析”
4. 默认情况下，成本分析的范围仅限于单个 Azure OpenAI 资源。

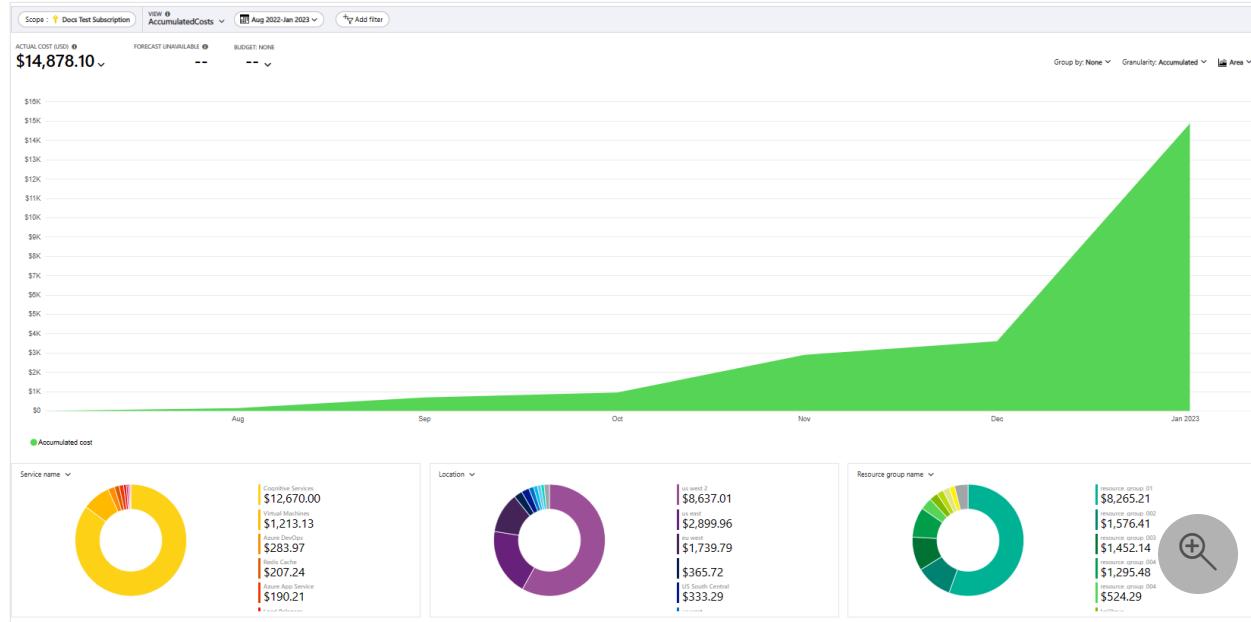


若要了解构成该成本的明细，可以将“分组依据”修改为“计量”，在本例中将图表类型切换为“折线图”。现在可以看到，对于此特定资源，成本源自三个不同的模型系列，其中“Text-Davinci 标记”表示大部分成本。

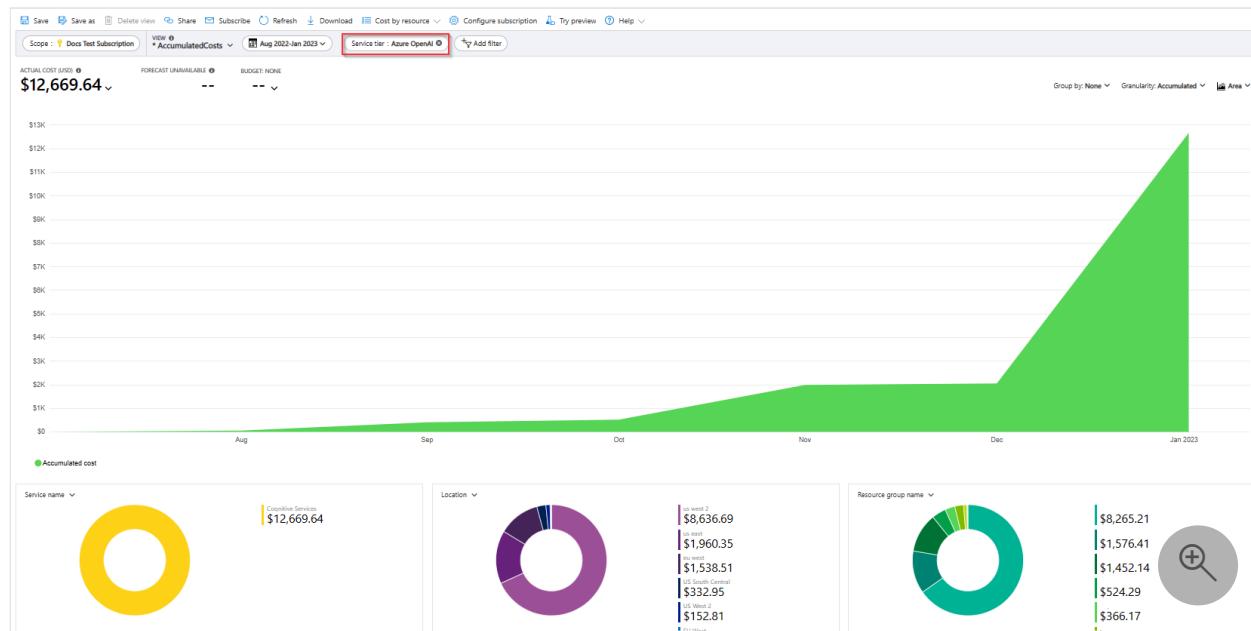


评估与 Azure OpenAI 相关的成本时，必须了解范围。如果资源是同一资源组的一部分，则可以在该级别确定成本分析的范围，以了解对成本的影响。如果资源分布在多个资源组中，则可以将范围限定为订阅级别。

但是，当范围限定在较高级别时，通常需要添加其他筛选器，以便能够准确跟踪 Azure OpenAI 的使用情况。当范围限定在订阅级别时，我们会在 Azure OpenAI 成本管理上下文中看到许多我们不在意的其他资源。在订阅级别确定范围时，建议导航到“成本管理”服务下的完整“成本分析工具”。在顶部的 Azure 搜索栏中搜索“成本管理”，导航到完整的服务体验，其中包括创建预算等更多选项。



如果尝试按服务添加筛选器，你会发现在列表中找不到 Azure OpenAI。这是因为从技术上讲，Azure OpenAI 是认知服务的一部分，因此服务级别筛选器是“认知服务”，但如果查看订阅中的所有 Azure OpenAI 资源，，而不查看任何其他类型的认知服务资源，则需要转而将范围限定为“服务层级: Azure OpenAI”：



创建预算

可以创建[预算](#)来管理成本，并创建[警报](#)以自动通知利益干系人支出异常和超支风险。警报基于与预算和成本阈值相比的支出。预算和警报是针对 Azure 订阅和资源组创建的，作为总体成本监视策略的一部分，它们非常有用。

如果希望增加监视中的精度，可以在 Azure 中使用筛选器为特定资源或服务创建预算。筛选器可帮助确保不会意外创建会产生额外成本的新资源。有关创建预算时可用的筛选选项的详细信息，请参阅[分组和筛选选项](#)。

① 重要

虽然 OpenAI 有一个硬限制选项，可以防止你超出预算，但 Azure OpenAI 目前不提供此功能。作为预算通知的一部分，你可以从操作组启动自动化，以采取更高级的操作，但这需要你进行额外的自定义开发。

导出成本数据

还可以将[成本数据导出](#)到存储帐户。当你或其他人需要进行有关成本的额外数据分析时，这非常有用。例如，财务团队可以使用 Excel 或 Power BI 来分析数据。可以按每天、每周或每月计划导出成本，并设置自定义的日期范围。建议导出成本数据来检索成本数据集。

后续步骤

- 了解[如何通过 Azure 成本管理优化云投资](#)。
- 详细了解[如何通过成本分析](#)来管理成本。
- 了解[如何防止意外成本](#)。
- 参与[成本管理引导式学习课程](#)。

其他资源

文 档

[监视 Azure OpenAI 服务 - Azure Cognitive Services](#)

从此处开始了解如何监视 Azure OpenAI 服务

[Azure OpenAI 服务静态数据加密 - Azure Cognitive Services](#)

了解 Azure OpenAI 在数据持久化到云中时如何加密数据。

[Fine Tunes - Create - REST API \(Azure Cognitive Services\)](#)

创建从给定训练文件微调指定模型的作业。响应包括排队作业的详细信息，包括作业状态和超参数

如何使用托管标识配置 Azure OpenAI 服务 - Azure OpenAI

提供有关如何使用 Azure Active Directory 设置托管标识的指导

Azure OpenAI 服务嵌入 - Azure OpenAI - embeddings and cosine similarity

详细了解用于执行文档搜索以及获取余弦相似性的 Azure OpenAI 嵌入 API

如何使用 Azure OpenAI 服务生成嵌入 - Azure OpenAI

了解如何使用 Azure OpenAI 生成嵌入

Models - REST API (Azure Cognitive Services)

详细了解 [认知服务模型操作]。 如何 [获取, 列出]。

Azure OpenAI 内容筛选 - Azure OpenAI

了解 Azure 认知服务中 OpenAI 服务的内容筛选功能

[显示另外 5 个](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

教程：探索 Azure OpenAI 服务嵌入和文档搜索

项目 • 2023/03/09

本教程将引导你使用 Azure OpenAI 嵌入 API 执行文档搜索，你将通过此操作查询知识库以查找最相关的文档。

在本教程中，你将了解如何执行以下操作：

- ✓ 安装 Azure OpenAI 和其他依赖 Python 库。
- ✓ 下载 BillSum 数据集并将其准备就绪进行分析。
- ✓ 为资源终结点和 API 密钥创建环境变量。
- ✓ 使用 text-search-curie-doc-001 和 text-search-curie-query-001 模型。
- ✓ 使用余弦相似性对搜索结果进行排名。

先决条件

- Azure 订阅 - [免费创建订阅](#)
- 已在所需的 Azure 订阅中授予对 Azure OpenAI 的访问权限。目前，仅应用程序授予对此服务的访问权限。可以通过在 <https://aka.ms/oai/access> 上填写表单来申请对 Azure OpenAI 的访问权限。如果有任何问题，请在此存储库上提出问题以联系我们。
- [Python 3.7.1 或更高版本](#)
- 以下 Python 库：openai、num2words、matplotlib、plotly、scipy、scikit-learn、pandas、transformers。
- 部署了带有 text-search-curie-doc-001 和 text-search-curie-query-001 模型的 Azure OpenAI 资源。这些模型当前仅在[特定区域](#)中可用。如果你没有资源，我们的[资源部署指南](#)中阐述了部署过程。

① 备注

如果从未使用过 Hugging Face Transformers 库，它有自己特定的[先决条件](#)，必须满足这些条件才能成功运行 `pip install transformers`。

设置

Python 库

如果尚未安装，则需要安装以下库：

Windows 命令提示符

```
pip install openai num2words matplotlib plotly scipy scikit-learn pandas
transformers
```

或者，可以使用我们的 [requirements.txt 文件](#)。

下载 BillSum 数据集

BillSum 是美国国会法案和加州法案的数据集。出于说明目的，我们只探讨美国法案。语料库由国会第 103 - 115 届 (1993 - 2018) 会议的法案组成。数据被拆分为 18,949 个训练法案和 3,269 个测试法案。BillSum 语料库侧重于长度从 5,000 到 20,000 个字符的中等长度立法。有关该项目的详细信息以及此数据集派生自的原始学术论文，请参阅 [BillSum 项目的 GitHub 存储库](#)

本教程使用 `bill_sum_data.csv` 文件，可从我们的 [GitHub 示例数据](#) 下载此文件

还可以在本地计算机上运行以下命令来下载示例数据：

Windows 命令提示符

```
curl "https://raw.githubusercontent.com/Azure-Samples/Azure-OpenAI-Docs-
Samples/main/Samples/Tutorials/Embeddings/data/bill_sum_data.csv" --output
bill_sum_data.csv
```

检索密钥和终结点

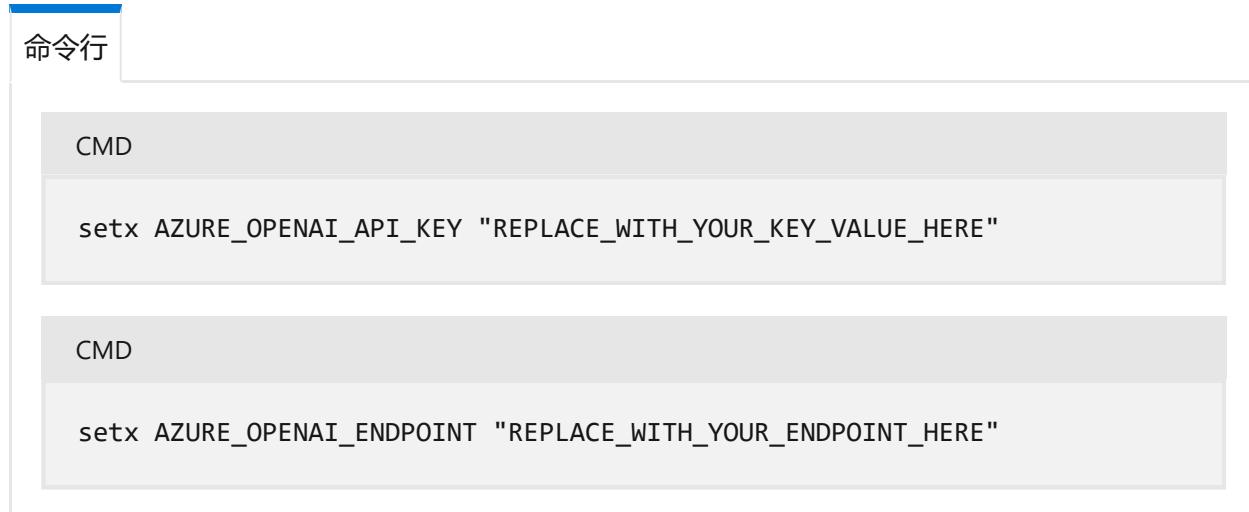
若要成功对 Azure OpenAI 发出调用，需要一个**终结点**和一个**密钥**。

变量名称	值
ENDPOINT	从 Azure 门户检查资源时，可在“密钥和终结点”部分中找到此值。也可在“Azure OpenAI Studio”>“操场”>“代码视图”中找到该值。示例终结点为： https://docs-test-001.openai.azure.com/ 。
API-KEY	从 Azure 门户检查资源时，可在“密钥和终结点”部分中找到此值。可以使用 <code>KEY1</code> 或 <code>KEY2</code> 。

在 Azure 门户中转到你的资源。可以在“资源管理”部分找到“终结点和密钥”。复制终结点和访问密钥，因为在对 API 调用进行身份验证时需要这两项。可以使用 `KEY1` 或 `KEY2`。始终准备好两个密钥可以安全地轮换和重新生成密钥，而不会导致服务中断。

为密钥和终结点创建和分配持久环境变量。

环境变量



The screenshot shows the Azure portal's "Environment variables" section. It includes two environment variables: `AZURE_OPENAI_API_KEY` and `AZURE_OPENAI_ENDPOINT`, both set to placeholder values. The portal also provides CMD and PowerShell snippets for setting these variables.

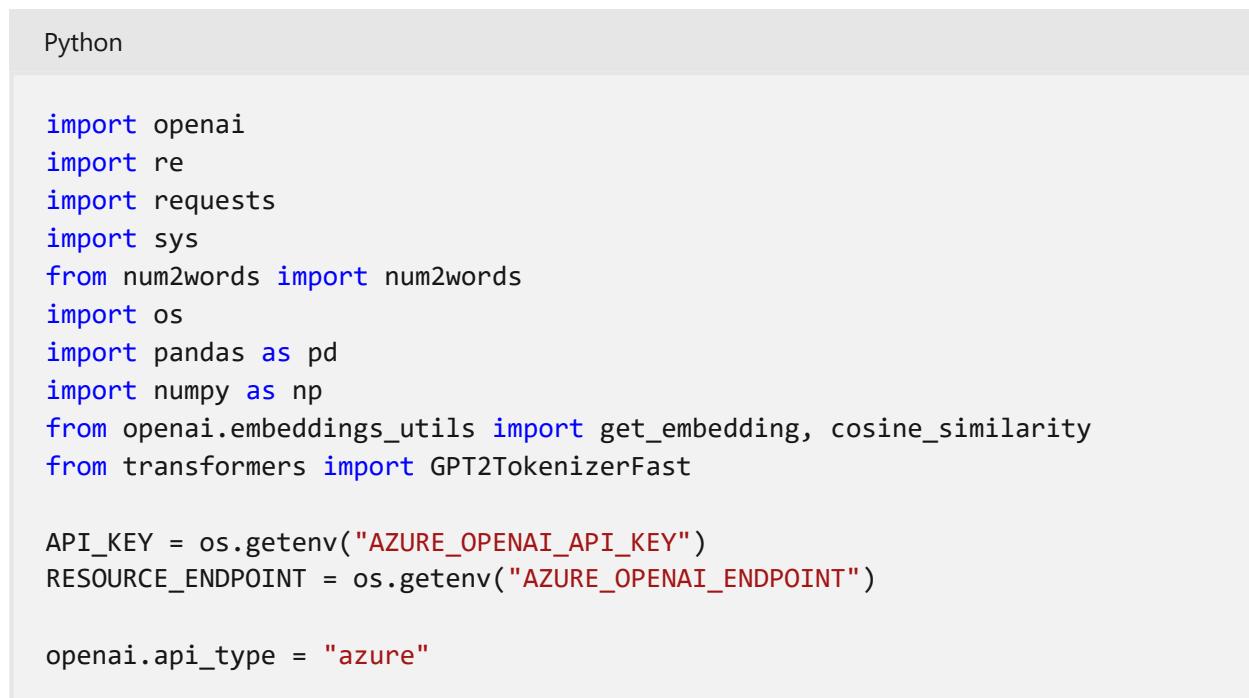
环境变量	值
<code>AZURE_OPENAI_API_KEY</code>	"REPLACE_WITH_YOUR_KEY_VALUE_HERE"
<code>AZURE_OPENAI_ENDPOINT</code>	"REPLACE_WITH_YOUR_ENDPOINT_HERE"

设置环境变量后，可能需要关闭并重新打开 Jupyter 笔记本或正在使用的任何 IDE，以便可以访问环境变量。

在首选 Python IDE 中运行以下代码：

如果要查看与本教程对应的 Jupyter 笔记本，可以从[示例存储库](#)下载该教程。

导入库并列出模型



The screenshot shows a Python code editor with the following code:

```
Python

import openai
import re
import requests
import sys
from num2words import num2words
import os
import pandas as pd
import numpy as np
from openai.embeddings_utils import get_embedding, cosine_similarity
from transformers import GPT2TokenizerFast

API_KEY = os.getenv("AZURE_OPENAI_API_KEY")
RESOURCE_ENDPOINT = os.getenv("AZURE_OPENAI_ENDPOINT")

openai.api_type = "azure"
```

```
openai.api_key = API_KEY
openai.api_base = RESOURCE_ENDPOINT
openai.api_version = "2022-12-01"

url = openai.api_base + "/openai/deployments?api-version=2022-12-01"

r = requests.get(url, headers={"api-key": API_KEY})

print(r.text)
```

输出：

Windows 命令提示符

```
{
  "data": [
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "text-davinci-002",
      "owner": "organization-owner",
      "id": "text-davinci-002",
      "status": "succeeded",
      "created_at": 1657572678,
      "updated_at": 1657572678,
      "object": "deployment"
    },
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "code-cushman-001",
      "owner": "organization-owner",
      "id": "code-cushman-001",
      "status": "succeeded",
      "created_at": 1657572712,
      "updated_at": 1657572712,
      "object": "deployment"
    },
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "text-search-curie-doc-001",
      "owner": "organization-owner",
      "id": "text-search-curie-doc-001",
      "status": "succeeded",
      "created_at": 1668620345,
      "updated_at": 1668620345,
      "object": "deployment"
    },
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "text-search-curie-doc-002",
      "owner": "organization-owner",
      "id": "text-search-curie-doc-002",
      "status": "succeeded",
      "created_at": 1668620345,
      "updated_at": 1668620345,
      "object": "deployment"
    }
  ]
}
```

```

    "scale_settings": {
        "scale_type": "standard"
    },
    "model": "text-search-curiel-query-001",
    "owner": "organization-owner",
    "id": "text-search-curiel-query-001",
    "status": "succeeded",
    "created_at": 1669048765,
    "updated_at": 1669048765,
    "object": "deployment"
}
],
"object": "list"
}

```

此命令的输出将因部署的模型的数量和类型而异。在这种情况下，需要确认我们有 text-search-curiel-doc-001 和 text-search-curiel-query-001 的条目。如果发现缺少其中一个模型，则需要先向资源[模型部署](#)，然后再继续操作。

现在，需要读取 csv 文件并创建 pandas 数据帧。创建初始数据帧后，可以通过运行 `df` 来查看表的内容。

Python

```

df = pd.read_csv("INSERT LOCAL PATH TO BILL_SUM_DATA.CSV") # example: df =
pd.read_csv("c:\\test\\bill_sum_data.csv")df
df

```

输出：

Unnamed: 0	bill_id	text	summary	title	text_len	sum_len
0	0	110_hr37	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	National Science Education Tax Incentive for B...	To amend the Internal Revenue Code of 1986 to ...	8494 321
1	1	112_hr2873	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Small Business Expansion and Hiring Act of 201...	To amend the Internal Revenue Code of 1986 to ...	6522 1424
2	2	109_s2408	SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR...	Requires the Director of National Intelligence...	A bill to require the Director of National Int...	6154 463
3	3	108_s1899	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	National Cancer Act of 2003 - Amends the Publi...	A bill to improve data collection and dissemin...	19853 1400
4	4	107_s1531	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Military Call-up Relief Act - Amends the Inter...	A bill to amend the Internal Revenue Code of 1...	6273 278
5	5	107_hr4541	SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR...	Requires the Customs Service to reliquidate ce...	To provide for reliquidation of entries premat...	11691 114
6	6	111_s1495	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Service Dogs for Veterans Act of 2009 - Direct...	A bill to require the Secretary of Veterans Af...	5328 379
7	7	111_s3885	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Race to the Top Act of 2010 - Directs the Secr...	A bill to provide incentives for States and lo...	16668 1525
8	8	113_hr1796	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Troop Talent Act of 2013 - Directs the Secreta...	Troop Talent Act of 2013	15352 2151
9	9	103_hr1987	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Taxpayer's Right to View Act of 1993 - Amends ...	Taxpayer's Right to View Act of 1993	5633 894
10	10	103_hr1677	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Full-Service Schools Act - Establishes the Fed...	Full-Service Schools Act	12472 1107
11	11	111_s3149	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Wall Street Compensation Reform Act of 2010 - ...	A bill to amend the Internal Revenue Code of 1...	18226 1297
12	12	110_hr1007	SECTION 1. FINDINGS.\\n\\n\\n The Congress f...	Amends the Marine Mammal Protection Act of 197...	To amend the Marine Mammal Protection Act of 1...	5261 276
13	13	113_hr3137	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Freedom and Mobility in Consumer Banking Act -...	Freedom and Mobility in Consumer Banking Act	17690 2044
14	14	115_hr1634	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Education and Training for Health Act of 2017 ...	Education and Training for Health Act of 2017	9037 772
15	15	103_hr1815	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Recreational Hunting Safety and Preservation A...	Recreational Hunting Safety and Preservation A...	13024 475
16	16	113_s1773	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Andrew Prior Act or Andrew's Law - Amends the ...	Andrew's Law	5149 613
17	17	106_hr5585	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Directs the President, in coordination with de...	Energy Independence Act of 2000	800 810
18	18	114_hr2499	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	This measure has not been amended since it was...	Veterans Entrepreneurship Act of 2015	7539 1421
19	19	111_hr3141	SECTION 1. SHORT TITLE\\n\\n\\n This Act ma...	Strengthening the Health Care Safety Net Act o...	To amend title XIX of the Social Security Act ...	18429 514

初始表的列数超过所需列数，我们将创建一个名为 `df_bills` 的较小的新数据帧，该数据帧仅包含 `text` `summary` 和 `title` 的列。

Python

```
df_bills = df[['text', 'summary', 'title']]  
df_bills
```

输出：

	text	summary	title
0	SECTION 1. SHORT TITLE. This Act may be cited ...	National Science Education Tax Incentive for B...	To amend the Internal Revenue Code of 1986 to ...
1	SECTION 1. SHORT TITLE. This Act may be cited ...	Small Business Expansion and Hiring Act of 201...	To amend the Internal Revenue Code of 1986 to ...
2	SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR...	Requires the Director of National Intelligence...	A bill to require the Director of National Int...
3	SECTION 1. SHORT TITLE. This Act may be cited ...	National Cancer Act of 2003 - Amends the Publi...	A bill to improve data collection and dissemin...
4	SECTION 1. SHORT TITLE. This Act may be cited ...	Military Call-up Relief Act - Amends the Inter...	A bill to amend the Internal Revenue Code of 1...
5	SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR...	Requires the Customs Service to reliquidate ce...	To provide for reliquidation of entries premat...
6	SECTION 1. SHORT TITLE. This Act may be cited ...	Service Dogs for Veterans Act of 2009 - Direct...	A bill to require the Secretary of Veterans Af...
7	SECTION 1. SHORT TITLE. This Act may be cited ...	Race to the Top Act of 2010 - Directs the Secr...	A bill to provide incentives for States and lo...
8	SECTION 1. SHORT TITLE. This Act may be cited ...	Troop Talent Act of 2013 - Directs the Secreta...	Troop Talent Act of 2013
9	SECTION 1. SHORT TITLE. This Act may be cited ...	Taxpayer's Right to View Act of 1993 - Amends ...	Taxpayer's Right to View Act of 1993
10	SECTION 1. SHORT TITLE. This Act may be cited ...	Full-Service Schools Act - Establishes the Fed...	Full-Service Schools Act
11	SECTION 1. SHORT TITLE. This Act may be cited ...	Wall Street Compensation Reform Act of 2010 - ...	A bill to amend the Internal Revenue Code of 1...
12	SECTION 1. FINDINGS. The Congress finds the fo...	Amends the Marine Mammal Protection Act of 197...	To amend the Marine Mammal Protection Act of 1...
13	SECTION 1. SHORT TITLE. This Act may be cited ...	Freedom and Mobility in Consumer Banking Act -...	Freedom and Mobility in Consumer Banking Act
14	SECTION 1. SHORT TITLE. This Act may be cited ...	Education and Training for Health Act of 2017 ...	Education and Training for Health Act of 2017
15	SECTION 1. SHORT TITLE. This Act may be cited ...	Recreational Hunting Safety and Preservation A...	Recreational Hunting Safety and Preservation A...
16	SECTION 1. SHORT TITLE. This Act may be cited ...	Andrew Prior Act or Andrew's Law - Amends the ...	Andrew's Law
17	SECTION 1. SHORT TITLE. This Act may be cited ...	Directs the President, in coordination with de...	Energy Independence Act of 2000
18	SECTION 1. SHORT TITLE. This Act may be cited ...	This measure has not been amended since it was...	Veterans Entrepreneurship Act of 2015
19	SECTION 1. SHORT TITLE. This Act may be cited ...	Strengthening the Health Care Safety Net Act o...	To amend title XIX of the Social Security Act ...

接下来，通过删除冗余的空格和清理标点来执行一些轻数据清理，以便为标记化准备数据。

Python

```
# s is input text  
def normalize_text(s, sep_token = " \n "):  
    s = re.sub(r'\s+', ' ', s).strip()  
    s = re.sub(r". ,",".",s)  
    # remove all instances of multiple spaces  
    s = s.replace(.., ".")  
    s = s.replace(.. ., ".")  
    s = s.replace("\n", "")  
    s = s.strip()  
  
    return s  
  
df_bills['text'] = df_bills["text"].apply(lambda x : normalize_text(x))
```

① 备注

如果收到指示 *A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row_indexer,col_indexer] = value instead 的警告，可以放心地忽略此消息。

现在，需要删除对于令牌限制（大约 2000 个令牌）来说太长的任何法案。

Python

```
tokenizer = GPT2TokenizerFast.from_pretrained("gpt2")
df_bills['n_tokens'] = df_bills["text"].apply(lambda x:
len(tokenizer.encode(x)))
df_bills = df_bills[df_bills.n_tokens<2000]
len(df_bills)
```

输出：

Windows 命令提示符

```
12
```

① 备注

可以忽略该消息：Token indices sequence length is longer than the specified maximum sequence length for this model (1480 > 1024). Running this sequence through the model will result in indexing errors. A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row_indexer,col_indexer] = value instead.。

我们将再次探讨 df_bills。请注意，正如预期的那样，尽管它们在第一列中保留了原始索引，现在仍仅返回 12 个结果，并且现在我们添加了一个名为 n_tokens 的列。

Python

```
df_bills
```

输出：

	text	summary	title	n_tokens
0	SECTION 1. SHORT TITLE. This Act may be cited ...	National Science Education Tax Incentive for B...	To amend the Internal Revenue Code of 1986 to ...	1480
1	SECTION 1. SHORT TITLE. This Act may be cited ...	Small Business Expansion and Hiring Act of 201...	To amend the Internal Revenue Code of 1986 to ...	1152
2	SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR...	Requires the Director of National Intelligence...	A bill to require the Director of National Int...	930
4	SECTION 1. SHORT TITLE. This Act may be cited ...	Military Call-up Relief Act - Amends the Inter...	A bill to amend the Internal Revenue Code of 1...	1048
5	SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR...	Requires the Customs Service to reliquidate ce...	To provide for reliquidation of entries premat...	1846
6	SECTION 1. SHORT TITLE. This Act may be cited ...	Service Dogs for Veterans Act of 2009 - Direct...	A bill to require the Secretary of Veterans Af...	872
9	SECTION 1. SHORT TITLE. This Act may be cited ...	Taxpayer's Right to View Act of 1993 - Amends ...	Taxpayer's Right to View Act of 1993	946
12	SECTION 1. FINDINGS. The Congress finds the fo...	Amends the Marine Mammal Protection Act of 197...	To amend the Marine Mammal Protection Act of 1...	1223
14	SECTION 1. SHORT TITLE. This Act may be cited ...	Education and Training for Health Act of 2017 ...	Education and Training for Health Act of 2017	1596
16	SECTION 1. SHORT TITLE. This Act may be cited ...	Andrew Prior Act or Andrew's Law - Amends the ...	Andrew's Law	608
17	SECTION 1. SHORT TITLE. This Act may be cited ...	Directs the President, in coordination with de...	Energy Independence Act of 2000	1341
18	SECTION 1. SHORT TITLE. This Act may be cited ...	This measure has not been amended since it was...	Veterans Entrepreneurship Act of 2015	1404

若要更深入地了解 n_tokens 列以及文本的标记化方式，运行以下代码可能会有所帮助：

Python

```
understand_tokenization = tokenizer.tokenize(df_bills.text[0])
understand_tokenization
```

对于我们的文档，我们有意截断输出，但在你的环境中运行此命令将返回第一个标记化为区块的索引的全文。

输出：

Windows 命令提示符

```
['S',
'ECITION',
'G1',
'.',
'GSH',
'ORT',
'GTIT',
'LE',
'.',
'GThis',
'GAct',
'Gmay',
'Gbe',
'Gcited',
'Gas',
'Gthe',
'G`',
'National',
'GScience',
'GEducation',
'GTax',
'GIn',
'cent',
'ive',
```

```
'for',  
'Business',  
...
```

如果随后检查 `understand_tokenization` 变量的长度，你会发现它与 `n_tokens` 列中的第一个数字匹配。

Python

```
len(understand_tokenization)
```

输出：

Windows 命令提示符

```
1480
```

现在，我们详细了解了标记化的工作原理，可以继续嵌入。在搜索之前，我们将嵌入文本文档并保存相应的嵌入。我们使用文档模型（在本例中为 `text-search-curie-doc-001`）嵌入每个区块。这些嵌入可以存储在本地，也可以存储在 Azure DB 中。因此，每个技术文档在数据帧右侧的新 `curie` 搜索列中都有相应的嵌入向量。

Python

```
df_bills['curie_search'] = df_bills["text"].apply(lambda x :  
get_embedding(x, engine = 'text-search-curie-doc-001'))
```

Python

```
df_bills
```

输出：

	text	summary	title	n_tokens	curie_search
0	SECTION 1. SHORT TITLE. This Act may be cited ...	National Science Education Tax Incentive for B...	To amend the Internal Revenue Code of 1986 to ...	1480	[-0.019770914688706398, 0.01169900186359882, ...
1	SECTION 1. SHORT TITLE. This Act may be cited ...	Small Business Expansion and Hiring Act of 201...	To amend the Internal Revenue Code of 1986 to ...	1152	[-0.007850012741982937, 0.01001765951514244, 0...
2	SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR...	Requires the Director of National Intelligence...	A bill to require the Director of National Int...	930	[0.00012103027984267101, 0.01845593340694904, ...
4	SECTION 1. SHORT TITLE. This Act may be cited ...	Military Call-up Relief Act - Amends the Inter...	A bill to amend the Internal Revenue Code of 1...	1048	[-0.005481021944433451, 0.00856819562613964, ...
5	SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR...	Requires the Customs Service to reliquidate ce...	To provide for reliquidation of entries premat...	1846	[-0.008310390636324883, -0.004660653416067362, ...
6	SECTION 1. SHORT TITLE. This Act may be cited ...	Service Dogs for Veterans Act of 2009 - Direct...	A bill to require the Secretary of Veterans Af...	872	[-0.017687108367681503, 0.01164870113134384, ...
9	SECTION 1. SHORT TITLE. This Act may be cited ...	Taxpayer's Right to View Act of 1993 - Amends ...	Taxpayer's Right to View Act of 1993	946	[0.0021867561154067516, -0.004219848196953535, ...
12	SECTION 1. FINDINGS. The Congress finds the fo...	Amends the Marine Mammal Protection Act of 197...	To amend the Marine Mammal Protection Act of 1...	1223	[-0.015813011676073074, 0.009919906966388226, ...
14	SECTION 1. SHORT TITLE. This Act may be cited ...	Education and Training for Health Act of 2017 ...	Education and Training for Health Act of 2017	1596	[-0.0150684155523777, 0.005073960404843092, 0...
16	SECTION 1. SHORT TITLE. This Act may be cited ...	Andrew Prior Act or Andrew's Law - Amends the ...	Andrew's Law	608	[-0.011593054980039597, 0.0275289676561356, ...
17	SECTION 1. SHORT TITLE. This Act may be cited ...	Directs the President, in coordination with de...	Energy Independence Act of 2000	1341	[-0.008348068222403526, 0.00272438395768404, 0...
18	SECTION 1. SHORT TITLE. This Act may be cited ...	This measure has not been amended since it was...	Veterans Entrepreneurship Act of 2015	1404	[-0.020315825939178467, 0.0011716989101842046, ...

在搜索（实时计算）时，我们将使用相应的查询模型（text-search-query-001）嵌入搜索查询。接下来，查找数据库中最接近的嵌入，按余弦相似性排名。

在示例中，用户提供了查询“我是否可以获取有线电视公司税收的信息”。查询通过一个函数传递，该函数使用相应的查询模型嵌入查询，并从上一步中以前嵌入的文档中查找最接近它的嵌入。

Python

```
# search through the reviews for a specific product
def search_docs(df, user_query, top_n=3, to_print=True):
    embedding = get_embedding(
        user_query,
        engine="text-search-curie-query-001"
    )
    df["similarities"] = df.curie_search.apply(lambda x:
cosine_similarity(x, embedding))

    res = (
        df.sort_values("similarities", ascending=False)
        .head(top_n)
    )
    if to_print:
        display(res)
    return res

res = search_docs(df_bills, "can i get information on cable company tax
revenue", top_n=4)
```

输出：

	text	summary	title	n_tokens	curie_search	similarities
9	SECTION 1. SHORT TITLE. This Act may be cited ...	Taxpayer's Right to View Act of 1993 - Amends ...	Taxpayer's Right to View Act of 1993	946	[0.0021867561154067516, -0.004219848196953535,...	0.363270
0	SECTION 1. SHORT TITLE. This Act may be cited ...	National Science Education Tax Incentive for B...	To amend the Internal Revenue Code of 1986 to ...	1480	[-0.019770914688706398, 0.01169900186359882, ...	0.314105
1	SECTION 1. SHORT TITLE. This Act may be cited ...	Small Business Expansion and Hiring Act of 201...	To amend the Internal Revenue Code of 1986 to ...	1152	[-0.007850012741982937, 0.01001765951514244, 0...	0.297908
18	SECTION 1. SHORT TITLE. This Act may be cited ...	This measure has not been amended since it was...	Veterans Entrepreneurship Act of 2015	1404	[-0.020315825939178467, 0.0011716989101842046, ...	0.295586

最后，我们将展示基于对整个知识库的用户查询的文档搜索的排名靠前结果。这返回了排名靠前的结果“1993 年纳税人观看权法案”。此文档的查询与文档之间的余弦相似性分为 0.36：

Python

```
res["summary"][9]
```

输出：

Windows 命令提示符

```
"Taxpayer's Right to View Act of 1993 - Amends the Communications Act of 1934 to prohibit a cable operator from assessing separate charges for any video programming of a sporting, theatrical, or other entertainment event if that event is performed at a facility constructed, renovated, or maintained with tax revenues or by an organization that receives public financial support. Authorizes the Federal Communications Commission and local franchising authorities to make determinations concerning the applicability of such prohibition. Sets forth conditions under which a facility is considered to have been constructed, maintained, or renovated with tax revenues. Considers events performed by nonprofit or public organizations that receive tax subsidies to be subject to this Act if the event is sponsored by, or includes the participation of a team that is part of, a tax exempt organization."
```

使用此方法，可以将嵌入用作知识库中跨文档的搜索机制。然后，用户可以获取排名靠前的搜索结果，并将其用于其下游任务，这会提示其初始查询。

视频

本教程的视频演练包括可在此[社区 YouTube 帖子](#)上查看的先决条件步骤。

清理资源

如果只是为了完成本教程而创建了 OpenAI 资源，并且想要清理和删除 OpenAI 资源，则需要删除已部署的模型，然后删除专用于测试资源的资源或关联的资源组。删除资源组

同时也会删除与之相关联的任何其他资源。

- [Portal](#)
- [Azure CLI](#)

后续步骤

详细了解 Azure OpenAI 的模型：

[Azure OpenAI 服务模型](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Use cases for Azure OpenAI Service

项目 • 2023/02/16 • 21 分钟可看完

What is a Transparency Note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the [Microsoft's AI principles](#) ↗.

The basics of Azure OpenAI

Introduction

Azure OpenAI provides customers with a fully managed AI service that lets developers and data scientists apply OpenAI's powerful language models including their GPT-3 and Codex series. GPT-3 models analyze and generate natural language, while Codex models analyze and generate code and plain text code commentary. These models use an autoregressive architecture meaning they use data from prior observations to predict the most probable next word. This process is then repeated by appending the newly generated content to the original text to produce the complete generated response. Because the response is conditioned on the input text, these models can be applied to a variety of tasks simply by changing the input text.

The GPT-3 series of models are pretrained on a wide body of publicly available free text data. This data is sourced from a combination of web crawling (specifically, a filtered version of [Common Crawl](#)), which includes a broad range of text from the internet and comprises sixty percent of the weighted pre-training dataset) and higher-quality datasets, including an expanded version of the WebText dataset, two internet-based books corpora and English-language Wikipedia. The training data for Codex contains both natural language and billions of lines of public code from GitHub. Given the wide breadth of knowledge and data that the models are trained on and their ability to generate dynamic content, special care must be taken to ensure responsible use in applications.

Learn more about the training and modeling techniques in OpenAI's [GPT-3](#) and [Codex](#) research papers. The guidance below is also drawn from [OpenAI's safety best practices](#).

Key terms

Term	Definition
Prompt	<p>The text you send to the service in the API call. This text is then input into the model. For example, one might input the following prompt:</p> <pre>Convert the questions to a command: Q: Ask Constance if we need some bread A: send-msg `find constance` Do we need some bread? Q: Send a message to Greg to figure out if things are ready for Wednesday. A:</pre>
Completion or Generation	<p>The text Azure OpenAI outputs in response. For example, the service may respond with the following answer to the above prompt:</p> <pre>send-msg `find greg` figure out if things are ready for Wednesday.</pre>
Token	<p>Azure OpenAI processes text by breaking it down into tokens. Tokens can be words or just chunks of characters. For example, the word "hamburger" gets broken up into the tokens "ham", "bur" and "ger", while a short and common word like "pear" is a single token. Many tokens start with a whitespace, for example " hello" and "bye".</p>

Capabilities

System behavior

The Azure OpenAI Service models use natural language instructions and examples in the prompt to identify the task. The model then completes the task by predicting the most probable next text. This technique is known as "in-context" learning. These models are not re-trained during this step but instead give predictions based on the context you include in the prompt.

There are three main approaches for in-context learning. These approaches vary based on the amount of task-specific data that is given to the model:

Few-shot: In this case, a user includes several examples in the prompt that demonstrate the expected answer format and content. The following example shows a few-shot prompt providing multiple examples:

```
Convert the questions to a command:  
Q: Ask Constance if we need some bread  
A: send-msg `find constance` Do we need some bread?  
Q: Send a message to Greg to figure out if things are ready for  
Wednesday.  
A: send-msg `find greg` Is everything ready for Wednesday?  
Q: Ask Ilya if we're still having our meeting this evening  
A: send-msg `find ilya` Are we still having a meeting this evening?  
Q: Contact the ski store and figure out if I can get my skis fixed  
before I leave on Thursday  
A: send-msg `find ski store` Would it be possible to get my skis fixed  
before I leave on Thursday?  
Q: Thank Nicolas for lunch  
A: send-msg `find nicolas` Thank you for lunch!  
Q: Tell Constance that I won't be home before 19:30 tonight – unmovable  
meeting.  
A: send-msg `find constance` I won't be home before 19:30 tonight. I  
have a meeting I can't move.  
Q: Tell John that I need to book an appointment at 10:30  
A:
```

The number of examples typically ranges from 0 to 100 depending on how many can fit in the maximum input length for a single prompt. Few-shot learning enables a major reduction in the amount of task-specific data required for accurate predictions.

One-shot: This case is the same as the few-shot approach except only one example is provided. The following example shows a one-shot prompt:

```
Convert the questions to a command:  
Q: Ask Constance if we need some bread  
A: send-msg `find constance` Do we need some bread?
```

Q: Send a message to Greg to figure out if things are ready for Wednesday.

A:

Zero-shot: In this case, no examples are provided to the model and only the task request is provided. The following example shows a zero-shot prompt:

Convert the question to a command:

Q: Ask Constance if we need some bread

A:

Use cases

Intended uses

Azure OpenAI can be used in multiple scenarios. The system's intended uses include:

- **Chat and conversation interaction:** Users can interact with a conversational agent that responds with responses drawn from trusted documents such as internal company documentation or tech support documentation; conversations must be limited to answering scoped questions.
- **Chat and conversation creation:** Users can create a conversational agent that responds with responses drawn from trusted documents such as internal company documentation or tech support documentation; conversations must be limited to answering scoped questions.
- **Code generation or transformation scenarios:** For example, converting one programming language to another, generating docstrings for functions, converting natural language to SQL.
- **Journalistic content:** For use to create new journalistic content or to rewrite journalistic content submitted by the user as a writing aid for pre-defined topics. Users cannot use the application as a general content creation tool for all topics. May not be used to generate content for political campaigns.
- **Question-answering:** Users can ask questions and receive answers from trusted source documents such as internal company documentation. The application does not generate answers ungrounded in trusted source documentation.
- **Reason over structured and unstructured data:** Users can analyze inputs using classification, sentiment analysis of text, or entity extraction. Examples include analyzing product feedback sentiment, analyzing support calls and transcripts, and refining text-based search with embeddings.

- **Search:** Users can search trusted source documents such as internal company documentation. The application does not generate results ungrounded in trusted source documentation.
- **Summarization:** Users can submit content to be summarized for pre-defined topics built into the application and cannot use the application as an open-ended summarizer. Examples include summarization of internal company documentation, call center transcripts, technical reports, and product reviews.
- **Writing assistance on specific topics:** Users can create new content or rewrite content submitted by the user as a writing aid for business content or pre-defined topics. Users can only rewrite or create content for specific business purposes or pre-defined topics and cannot use the application as a general content creation tool for all topics. Examples of business content include proposals and reports. For journalistic use, see above **Journalistic content** use case.

Considerations when choosing a use case

We encourage customers to leverage Azure OpenAI in their innovative solutions or applications. However, here are some considerations when choosing a use case:

- **Not suitable for open-ended, unconstrained content generation.** Scenarios where users can generate content on any topic are more likely to produce offensive or harmful text. The same is true of longer generations.
- **Not suitable for scenarios where up-to-date, factually accurate information is crucial** unless you have human reviewers or are using the models to search your own documents and have verified suitability for your scenario. The service does not have information about events that occur after its training date, likely has missing knowledge about some topics, and may not always produce factually accurate information.
- **Avoid scenarios where use or misuse of the system could result in significant physical or psychological injury to an individual.** For example, scenarios that diagnose patients or prescribe medications have the potential to cause significant harm.
- **Avoid scenarios where use or misuse of the system could have a consequential impact on life opportunities or legal status.** Examples include scenarios where the AI system could affect an individual's legal status, legal rights, or their access to credit, education, employment, healthcare, housing, insurance, social welfare benefits, services, opportunities, or the terms on which they are provided.
- **Avoid high stakes scenarios that could lead to harm.** The models hosted by Azure OpenAI service reflect certain societal views, biases and other undesirable content present in the training data or the examples provided in the prompt. As a result,

we caution against using the models in high-stakes scenarios where unfair, unreliable, or offensive behavior might be extremely costly or lead to harm.

- **Carefully consider use cases in high stakes domains or industry:** Examples include but are not limited to healthcare, medicine, finance or legal.
- **Carefully consider well-scoped chatbot scenarios.** Limiting the use of the service in chatbots to a narrow domain reduces the risk of generating unintended or undesirable responses.
- **Carefully consider all generative use cases.** Content generation scenarios may be more likely to produce unintended outputs and these scenarios require careful consideration and mitigations.

Limitations

When it comes to large-scale natural language models, there are particular fairness and responsible AI issues to consider. People use language to describe the world and to express their beliefs, assumptions, attitudes, and values. As a result, publicly available text data typically used to train large-scale natural language processing models contains societal biases relating to race, gender, religion, age, and other groups of people, as well as other undesirable content. These societal biases are reflected in the distributions of words, phrases, and syntactic structures.

Technical limitations, operational factors and ranges

Large-scale natural language models trained with such data can potentially behave in ways that are unfair, unreliable, or offensive, in turn causing harms. There are several different ways in which a large-scale natural language processing model can cause harms. Some of the ways are listed here. We emphasize that these types of harms aren't mutually exclusive. A single model can exhibit more than one type of harm, potentially relating to multiple different groups of people. For example:

- **Allocation:** Language models can be used in ways that lead to unfair allocation of resources or opportunities. For example, automated resume screening systems can withhold employment opportunities from one gender if they're trained on resume data that reflects the existing gender imbalance in a particular industry.
- **Quality of service:** Language models can fail to provide the same quality of service to some people as they do to others. For example, sentence completion systems may not work as well for some dialects or language varieties because of their lack of representation in the training data. The Azure OpenAI models are trained primarily on English text. Languages besides English will experience worse

performance. English language varieties with less representation in the training data may experience worse performance.

- **Stereotyping:** Language models can reinforce stereotypes. For example, when translating "He is a nurse" and "She is a doctor" into a genderless language such as Turkish and then back into English, many machine translation systems yield the stereotypical (and incorrect) results of "She is a nurse" and "He is a doctor."
- **Demeaning:** Language models can demean people. For example, an open-ended content generation system with inappropriate mitigations might produce offensive text targeted at a particular group of people.
- **Over- and underrepresentation:** Language models can over- or underrepresent groups of people, or even erase them entirely. For example, toxicity detection systems that rate text containing the word "gay" as toxic might lead to the underrepresentation or even erasure of legitimate text written by or about the LGBTQIA+ community.
- **Inappropriate or offensive content:** Language models can produce other types of inappropriate or offensive content. Examples include hate speech; text that contains profane words or phrases; text that relates to illicit activities; text that relates to contested, controversial, or ideologically polarizing topics; misinformation; text that's manipulative; and text that relates to sensitive or emotionally charged topics. For example, suggested-reply systems that are restricted to positive replies can suggest inappropriate or insensitive replies for messages about negative events.
- **False information:** Azure OpenAI service doesn't fact check or verify content provided by customers or users. Depending on how you've developed your application, it might promote false information unless you've built in mitigations.

System performance

In many AI systems, performance is often defined in relation to accuracy—that is, how often the AI system offers a correct prediction or output. With large-scale natural language models, two different users might look at the same output and have different opinions of how useful or relevant it is, which means that performance for these systems must be defined more flexibly. Here, we broadly consider performance to mean that the application performs as you and your users expect, including not generating harmful outputs.

Azure OpenAI service can support a wide range of applications like search, classification, and code generation, each with different performance metrics and mitigation strategies. There are several steps you can take to mitigate some of the concerns listed under "Limitations" and to improve performance. Additional important mitigation techniques are outlined in the section Evaluating and integrating Azure OpenAI for your use below.

- **Show and tell when designing prompts.** Make it clear to the model what kind of outputs you expect through instructions, examples, or a combination of the two. If you want the model to rank a list of items in alphabetical order or to classify a paragraph by sentiment, show it that's what you want.
- **Keep your application on-topic.** Carefully structure prompts to reduce the chance of producing undesired content, even if a user tries to use it for this purpose. For instance, you might indicate in your prompt that a chatbot only engages in conversations about mathematics and otherwise responds "I'm sorry. I'm afraid I can't answer that." Adding adjectives like "polite" and examples in your desired tone to your prompt can also help steer outputs. Consider nudging users toward acceptable queries, either by listing such examples upfront or by offering them as suggestions upon receiving an off-topic request. Consider training a classifier to determine whether an input is on- or off-topic.
- **Provide quality data.** If you're trying to build a classifier or get the model to follow a pattern, make sure that there are enough examples. Be sure to proofread your examples — the model is usually smart enough to see through basic spelling mistakes and give you a response, but it also might assume this is intentional and it could affect the response.
- **Measure model quality.** As part of general model quality, consider measuring and improving fairness-related metrics and other metrics related to responsible AI in addition to traditional accuracy measures for your scenario. Consider resources like this checklist when you measure the fairness of the system. These measurements come with limitations, which you should acknowledge and communicate to stakeholders along with evaluation results.
- **Limit the length, structure, rate, and source of inputs and outputs.** Restricting the length or structure of inputs and outputs can increase the likelihood that the application will stay on task and mitigate, at least in part, any potential unfair, unreliable, or offensive behavior. Restricting the source of inputs (for example, limiting inputs to a fixed list of items, a particular domain, or to authenticated users rather than anyone on the internet) or restricting the source of outputs (for example, only surfacing answers from approved, vetted documents rather than the open web) can further mitigate the risk of the harms. Putting usage rate limits in place can also reduce misuse.
- **Implement additional scenario-specific mitigations.** Refer to the mitigations outlined in [Evaluating and integrating Azure OpenAI for your use](#) below, including content moderation strategies. They do not represent every mitigation that may be required for your application but point to the general minimum baseline we check for when approving use cases for the Azure OpenAI service.

Evaluation of Azure OpenAI

GPT-3 models were trained on a total of 300 billion tokens and evaluated in zero-shot, one-shot and few-shot settings on a range of different NLP benchmarks and tasks, such as language modeling and completion tasks, and on datasets that involve commonsense reasoning, reading comprehension and question answering. GPT-3 shows strong performance on many NLP benchmarks and tasks, it nearly matches the performance of state-of-the-art fine-tuned systems in some cases, and demonstrates strong qualitative performance at tasks defined on-the-fly. More information and benchmarking statistics can be found in the OpenAI [GPT-3 research paper](#).

Evaluating and integrating Azure OpenAI for your use

Practices for responsible use

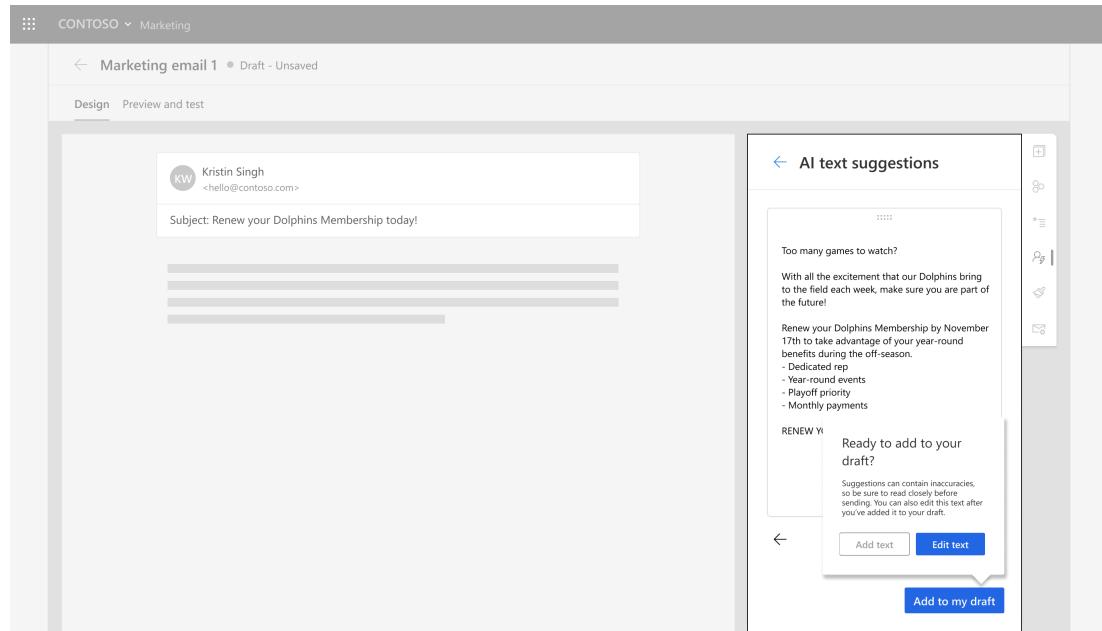
The practices below are baseline requirements for production applications.

1. Ensure human oversight. Ensure appropriate human oversight for your system, including ensuring the people responsible for oversight understand the system's intended uses; how to effectively interact with the system; how to effectively interpret system behavior; when and how to override, intervene, or interrupt the system; and ways to remain aware of the possible tendency of over-relying on outputs produced by the system ("automation bias"). Especially in high-stakes scenarios, make sure people have a role in decision-making and can intervene in real time to prevent harm. For people to make good decisions about system outputs, they need to (a) understand how the system works, (b) have awareness of the system status and how well it's working in their scenario, and (c) have the opportunity and the tools to bring the system into alignment with their expectations and goals. When using the Azure OpenAI service, human oversight might include some or all of the following:

- **1a Let people edit generated outputs.** (*For detailed guidance and examples, see [HAX G9-B: Rich and detailed edits](#).*) In this example, users are given the option to edit each generated output before using the text. Our research showed that users expected to be able to edit content and even combine parts of different generations.
- **1b. Highlight potential inaccuracies in generated outputs.** (*For detailed guidance and examples, see [HAX G2-C: Report system performance information](#).*) Here, numbers and terms having to do with time periods are underlined. When a user hovers over an underlined item, a tooltip appears that reminds them to check its accuracy. Depending on your scenario, you may want to highlight numbers, days, dates, names, URLs, titles, quotations,

addresses, phone numbers, and hashtags. Our research suggests that users find this kind of flagging useful as a reminder to fact-check the content.

- **1c. Remind users that they are accountable for final decisions and/or final content.** In the example below, before a user is allowed to insert generated content into their document, they must go through a pop-up dialog and edit the text or acknowledge that they've reviewed it. Our research suggests that users appreciate having these reminders and believe them to be especially beneficial for beginners.

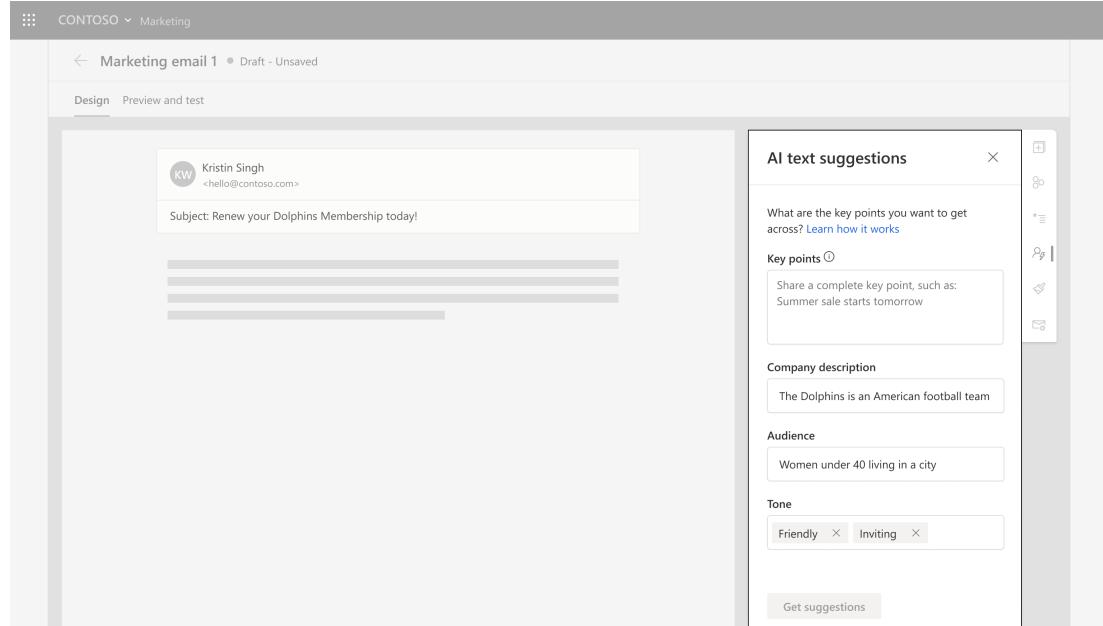


- **1d. Limit how people can automate your product or service.** For example, don't allow automated posting of generated content to external sites (including social media), or automated execution of generated code.
- **1e. Disclose AI's role in generated content.** In some cases, letting content consumers know when published content is partly or fully generated by Azure OpenAI can help them use their own judgment about how to read it. If generated content does not include meaningful human oversight before being shared or published--including opportunities for an expert to understand, review, and edit the content--disclosure may be critical to preventing misinformation.

2. Implement technical limits on inputs and outputs

- **2a. Limit the length of inputs and outputs.** Restricting input and output length can reduce the likelihood of producing undesirable content, misuse of the system beyond its intended use cases, or other harmful or unintended scenarios.

- **2b. Structure inputs to limit open-ended responses and to give users more-refined control.** The better the instructions that users give each time they interact with the system, the better the results they'll get. Restrict users from creating custom prompts that let them operate as if interacting directly with the API. You can also limit outputs to be structured in certain formats or patterns.



In the example above, the prompt has been structured to require that users provide limiting details such as audience and tone. When setting prompt fields, consider what information will be easy for users to provide, and run experiments to learn what information changes output quality. Smart defaults can help people get started quickly and can also be used to demonstrate best practices for prompt format, length, and style.

- **2c. Return outputs from validated, reliable source materials**, such as existing support articles, rather than returning non-vetted content from the internet. This can help your application stay on task and mitigate unfair, unreliable, or offensive behavior.
- **2d. Implement blocklists and content moderation.** Keep your application on topic by checking both inputs and outputs for undesired content. The definition of undesired content depends on your scenario and changes over time. It might include hate speech, text that contains profane words or phrases, misinformation, and text that relates to sensitive or emotionally charged topics. Checking inputs can help keep your application on topic, even if a malicious user tries to produce undesired content. Checking API outputs can allow you to detect undesired content produced by the system

and take action. You can replace it, report it, ask the user to enter different input, or provide input examples.

- **2e. Put rate limits in place** (in other words, frequency and quantity of API calls) to further reduce misuse.

3. Authenticate users. To make misuse more difficult, consider requiring that customers sign in and, if appropriate, link a valid payment method. Consider working only with known, trusted customers in the early stages of development. Applications that do not authenticate users may require other, stricter mitigations to ensure the application cannot be used beyond its intended purpose.

4. Test your application thoroughly to ensure it responds in a way that is fit for the application's purpose. This includes conducting adversarial testing where trusted testers attempt to find system failures, poor performance, or undesirable behaviors. This information helps you to understand risks and consider appropriate mitigations. Communicate the capabilities and limitations to stakeholders.

5. Establish feedback channels for users and impacted groups. AI-powered products and features require ongoing monitoring and improvement. Establish channels to collect questions and concerns from users as well as people affected by the system. For example:

- **5a. Build feedback features into the user experience.** Invite feedback on the usefulness and accuracy of outputs, and give users a separate and clear path to report outputs that are problematic, offensive, biased, or otherwise inappropriate. For detailed guidance and examples, see [HAX Guideline 15: Encourage granular feedback.](#) ↗
- **5b. Publish an easy-to-remember email address for public feedback.**

Scenario-specific practices

1. **If your application powers chatbots or other conversational AI systems**, follow the [Microsoft guidelines for responsible development of conversational AI systems](#) ↗.
2. **If you are developing an application in a high stakes domain or industry**, such as healthcare, human resources, education, or the legal field, thoroughly assess how well the application works in your scenario, implement strong human oversight, thoroughly evaluate how well users understand the limitations of the application, and comply with all relevant laws. Consider additional mitigations based on your scenario.

Additional practices

1. Use Microsoft's [Inclusive Design Guidelines](#) to build inclusive solutions.
2. **Conduct research to test the product and solicit feedback.** Include a diverse group of stakeholders (for example, direct users, consumers of generated results, admins, and so on) to your research structure to seek their feedback at different stages of deployment. Depending on the research goal, you can use various methodologies, such as [Community Jury](#), online experiments, beta testing, and testing with real users after deployment. Consider including stakeholders from different demographic groups to gather a wider range of feedback.
3. **Conduct a legal review.** Obtain appropriate legal advice to review your solution, particularly if you plan to use it in sensitive or high-risk applications.
4. Learn more about responsible AI [here](#).

Learn more about responsible AI

- Microsoft AI principles
- Microsoft responsible AI resources
- Microsoft Azure Learning courses on responsible AI

Learn more about Azure OpenAI

- Limited access to Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn
- Code of Conduct for the Azure OpenAI Service | Microsoft Learn
- Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn

See also

- Limited access to Azure OpenAI Service
- Code of conduct for Azure OpenAI Service integrations
- Data, privacy, and security for Azure OpenAI Service

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Limited access to Azure OpenAI Service

项目 • 2023/02/09 • 2 分钟可看完

As part of Microsoft's commitment to responsible AI, we are designing and releasing Azure OpenAI Service with the intention of protecting the rights of individuals and society and fostering transparent human-computer interaction. For this reason, we currently limit the access and use of Azure OpenAI, including limiting access to the ability to modify content filters and modify abuse monitoring.

Registration process

Azure OpenAI requires registration and is currently only available to managed customers and partners working with Microsoft account teams. Customers who wish to use Azure OpenAI are required to submit [a registration form](#) .

Customers must attest to any and all use cases for which they will use the service. Customers who wish to add additional use cases after initial onboarding must submit the additional use cases using [this form](#) . The use of Azure OpenAI is limited to use cases that have been selected in a registration form. Microsoft may require customers to re-verify this information. Read more about example use cases and use cases to avoid [here](#).

Customers who wish to modify content filters and modify abuse monitoring after they have onboarded to the service are subject to additional scenario restrictions and are required to register [here](#) .

Access to the Azure OpenAI Service is subject to Microsoft's sole discretion based on eligibility criteria and a vetting process, and customers must acknowledge that they have read and understand the Azure terms of service for Azure OpenAI Service.

Azure OpenAI Service is made available to customers under the terms governing their subscription to Microsoft Azure Services, including the Azure OpenAI section of the [Microsoft Product Terms](#) . Please review these terms carefully as they contain important conditions and obligations governing your use of Azure OpenAI Service.

Important links

- Register to use Azure OpenAI [↗](#)
- Add additional use cases [↗](#) (if needed)
- Register to modify content filters and abuse monitoring [↗](#) (if needed)

Help and support

FAQ about Limited Access can be found [here](#). If you need help with Azure OpenAI, find support [here](#). Report abuse of Azure OpenAI [here](#) [↗](#).

Report problematic content to cscraireport@microsoft.com.

See also

- [Code of conduct for Azure OpenAI Service integrations](#)
- [Transparency note for Azure OpenAI Service](#)
- [Characteristics and limitations for Azure OpenAI Service](#)
- [Data, privacy, and security for Azure OpenAI Service](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Code of conduct for Azure OpenAI Service

项目 • 2023/01/30 • 6 分钟可看完

The following Code of Conduct defines the requirements that all Azure OpenAI Service implementations must adhere to in good faith. This code of conduct is in addition to the Acceptable Use Policy in the [Microsoft Online Services Terms](#) .

Access requirements

Azure OpenAI Service is a Limited Access service that requires registration and is only available to managed customers and partners with Microsoft account teams. Customers who wish to use this feature are required to [register through this form](#) both for initial access for experimentation and for approval to move from experimentation to production. To learn more, see [Limited Access to Azure OpenAI Service](#).

Responsible AI mitigation requirements

Integrations with AzureOpen AI Service must:

- Implement meaningful human oversight
- Implement strong technical limits on inputs and outputs to reduce the likelihood of misuse beyond the application's intended purpose
- Test applications thoroughly to find and mitigate undesirable behaviors
- Establish feedback channels
- Implement additional scenario-specific mitigations

To learn more, see the [Azure OpenAI transparency note](#).

Integrations with Azure OpenAI Service must not:

- be used in any way that violates Microsoft's [Acceptable Use Policy](#), including but not limited to any use prohibited by law, regulation, government order, or decree, or any use that violates the rights of others;
- be used in any way that is inconsistent with this code of conduct, including the Limited Access requirements, the Responsible AI mitigation requirements, and the Content requirements;
- exceed the documented use case you provided to Microsoft in connection with your request to use the service;
- interact with individuals under the age of consent in any way that could result in exploitation or manipulation or otherwise prohibited by law or regulation;
- generate or interact with content prohibited in this Code of Conduct;
- be presented alongside or monetize content prohibited in this Code of Conduct;
- make decisions without appropriate human oversight if your application may have a consequential impact on any individual's legal position, financial position, life opportunities, employment opportunities, human rights, or result in physical or psychological injury to an individual;
- infer sensitive information about people without their explicit consent unless if used in a lawful manner by a law enforcement entity, court, or government official subject to judicial oversight in a jurisdiction that maintains a fair and independent judiciary; or
- be used for chatbots that (i) are erotic, romantic, or used for companionship purposes, or which are otherwise prohibited by this Code of Conduct; (ii) are personas of specific people without their explicit consent; (iii) claim to have special wisdom/insight/knowledge, unless very clearly labeled as being for entertainment purposes only; or (iv) enable end users to create their own chatbots without oversight.

Content requirements

We prohibit the use of our service for generating content that can inflict harm on individuals or society. Our content policies are intended to improve the safety of our platform.

These content requirements apply to the output of all models developed by OpenAI and hosted in Azure OpenAI, such as GPT-3 and Codex models, and includes content provided as input to the service and content generated as output from the service.

Exploitation and Abuse

Child sexual exploitation and abuse

Azure OpenAI Service prohibits content that describes, features, or promotes child sexual exploitation or abuse, whether or not prohibited by law. This includes sexual content involving a child or that sexualizes a child.

Grooming

Azure OpenAI Service prohibits content that describes or is used for purposes of grooming of children. Grooming is the act of an adult building a relationship with a child for the purposes of exploitation, especially sexual exploitation. This includes communicating with a child for the purpose of sexual exploitation, trafficking, or other forms of exploitation.

Non-consensual intimate content

Azure OpenAI Service prohibits content that describes, features or promotes non-consensual intimate activity.

Sexual solicitation

Azure OpenAI Service prohibits content that describes, features, or promotes, or used for, purposes of solicitation of commercial sexual activity and sexual services. This includes encouragement and coordination of real sexual activity.

Trafficking

Azure OpenAI Service prohibits content describing or used for purposes of human trafficking. This includes the recruitment of individuals, facilitation of transport and payment for, and the promotion of, exploitation of people such as forced labor, domestic servitude, sexual slavery, forced marriages, and forced medical procedures.

Suicide and Self-Injury

Azure OpenAI Service prohibits content that describes, praises, supports, promotes, glorifies, encourages and/or instructs individual(s) on self-injury or to take their life.

Violent Content and Conduct

Graphic violence and gore

Azure OpenAI Service prohibits content that describes, features, or promotes graphic violence or gore.

Terrorism and Violent Extremism

Azure OpenAI Service prohibits content that depicts an act of terrorism, praises, or supports a terrorist organization, terrorist actor, or violent terrorist ideology, encourages terrorist activities, offers aid to terrorist organizations or terrorist causes, or aids in recruitment to a terrorist organization.

Violent Threats, Incitement, and Glorification of Violence

Azure OpenAI Service prohibits content advocating or promoting violence toward others through violent threats or incitement.

Harmful Content

Hate speech and discrimination

Azure OpenAI Service prohibits content that attacks, denigrates, intimidates, degrades, targets, or excludes individuals or groups on the basis of traits such as actual or perceived race, ethnicity, national origin, gender, gender identity, sexual orientation, religious affiliation, age, disability status, caste, or any other characteristic that is associated with systemic prejudice or marginalization.

Bullying and harassment

Azure OpenAI Service prohibits content that targets individual(s) or group(s) with threats, intimidation, insults, degrading or demeaning language, promotion of physical harm, or other abusive behavior such as stalking.

Deception, disinformation, and inauthentic activity

Azure OpenAI Service prohibits content that is intentionally deceptive and likely to adversely affect the public interest, including deceptive or untrue content relating to health, safety, election integrity, or civic participation). Azure OpenAI Service also prohibits inauthentic interactions, such as fake accounts, automated inauthentic activity, impersonation to gain unauthorized information or privileges, and claims to be from any person, company, government body, or entity without explicit permission to make that representation.

Active malware or exploits

Content that directly supports unlawful active attacks or malware campaigns that cause technical harms, such as delivering malicious executables, organizing denial of service attacks, or managing command and control servers.

Additional content policies

We prohibit the use of our Azure OpenAI Service for scenarios in which the system is likely to generate undesired content due to limitations in the models or scenarios in which the system cannot be applied in a way that properly manages potential negative consequences to people and society. Without limiting the foregoing restriction, Microsoft reserves the right to revise and expand the above Content requirements to address specific harms to people and society.

This includes content that is sexually graphic, including consensual pornographic content and intimate descriptions of sexual acts, as well as content that may influence the political process, such as an election, passage of legislation, and content for campaigning purposes;

We may at times limit our service's ability to respond to particular topics, such as probing for personal information or seeking opinions on sensitive topics or current events.

We prohibit the use of Azure OpenAI Service for activities that significantly harm other individuals, organizations, or society, including but not limited to use of the service for purposes in conflict with the applicable [Azure Legal Terms](#) and the [Microsoft Product Terms](#).

Report abuse

If you suspect that Azure OpenAI Service is being used in a manner that is abusive or illegal, infringes on your rights or the rights of other people, or violates these policies you can report it at the [Report Abuse Portal](#).

Report problematic content

If Azure OpenAI Service outputs problematic content that you believe should have been filtered, report it at cscraireport@microsoft.com.

See also

- [Limited access to Azure OpenAI Service](#)
- [Transparency note for Azure OpenAI Service](#)
- [Data, privacy, and security for Azure OpenAI Service](#)

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Data, privacy, and security for Azure OpenAI Service

项目 • 2023/01/30 • 5 分钟可看完

This article provides details regarding how data provided by you to the Azure OpenAI service is processed, used, and stored. Azure OpenAI stores and processes data to provide the service, monitor for abusive use, and to develop and improve the quality of Azure's Responsible AI systems. Please also see the [Microsoft Products and Services Data Protection Addendum](#), which governs data processing by the Azure OpenAI Service except as otherwise provided in the applicable [Product Terms](#).

Azure OpenAI was designed with compliance, privacy, and security in mind; however, the customer is responsible for its use and the implementation of this technology.

What data does the Azure OpenAI Service process?

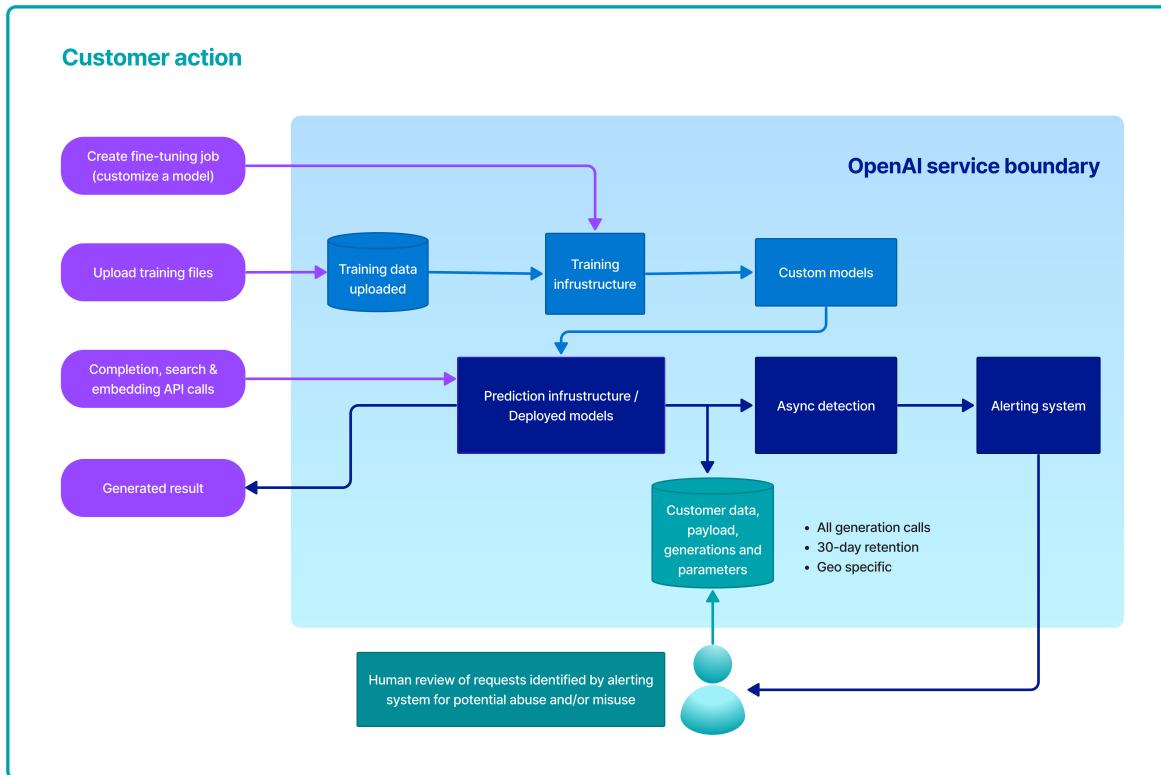
Azure OpenAI processes the following types of data:

- **Text prompts, queries and responses** submitted by the user via the completions, search, and embeddings operations.
- **Training & validation data.** You can provide your own training data consisting of prompt-completion pairs for the purposes of fine-tuning an OpenAI model.
- **Results data from training process.** After training a fine-tuned model, the service will output meta-data on the job which includes tokens processed and validation scores at each step.

How does the Azure OpenAI Service process data?

The diagram below illustrates how your data is processed. This diagram covers three different types of processing:

1. How the Azure OpenAI Service creates a fine-tuned (custom) model with your training data
2. How the Azure OpenAI Service processes your text prompts to generate completions, embeddings, and search results; and
3. How the Azure OpenAI Service and Microsoft personnel analyze prompts & completions for abuse, misuse or harmful content generation.



Training data for purposes of fine-tuning an OpenAI model

The training data (prompt-completion pairs) submitted to the Fine-tunes API through the Azure OpenAI Studio is pre-processed using automated tools for quality checking including data format check. The training data is then imported to the model training component on the Azure OpenAI platform. During the training process, the training data are decomposed into batches and used to modify the weights of the OpenAI models.

Training data provided by the customer is only used to fine-tune the customer's model and is not used by Microsoft to train or improve any Microsoft models.

Text prompts to generate completions, embeddings and search results

Once a model is deployed, you can generate text using this model using our Completions operation through the REST API, client libraries or Azure OpenAI Studio.

Abuse and harmful content generation

The Azure OpenAI Service stores prompts & completions from the service to monitor for abusive use and to develop and improve the quality of Azure OpenAI's content management systems. [Learn more about our content management and filtering ↗](#).

Authorized Microsoft employees can access your prompt & completion data that has triggered our automated systems for the purposes of investigating and verifying potential abuse; for customers who have deployed Azure OpenAI Service in the European Union, the authorized Microsoft employees will be located in the European Union. This data may be used to improve our content management systems.

In the event of a confirmed policy violation, we may ask you to take immediate action to remediate the issue to and to prevent further abuse. Failure to address the issue may result in suspension or termination of Azure OpenAI resource access.

How is data retained and what Customer controls are available?

- **Training, validation, and training results data.** The Files API allows customers to upload their training data for the purpose of fine-tuning a model. This data is stored in Azure Storage, encrypted at rest by Microsoft Managed keys, within the same region as the resource and logically isolated with their Azure subscription and API Credentials. Uploaded files can be deleted by the user via the DELETE API operation.
- **Fine-tuned OpenAI models.** The Fine-tunes API allows customers to create their own fine-tuned version of the OpenAI models based on the training data that you have uploaded to the service via the Files APIs. The trained fine-tuned models are stored in Azure Storage in the same region, encrypted at rest and logically isolated with their Azure subscription and API credentials. Fine-tuned models can be deleted by the user by calling the DELETE API operation.
- **Text prompts, queries and responses.** The requests & response data may be temporarily stored by the Azure OpenAI Service for up to 30 days. This data is encrypted and is only accessible to authorized engineers for (1) debugging purposes in the event of a failure, (2) investigating patterns of abuse and misuse or (3) improving the content filtering system through using the prompts and completions flagged for abuse or misuse.

To learn more about Microsoft's privacy and security commitments visit the [Microsoft Trust Center](#).

Frequently asked questions

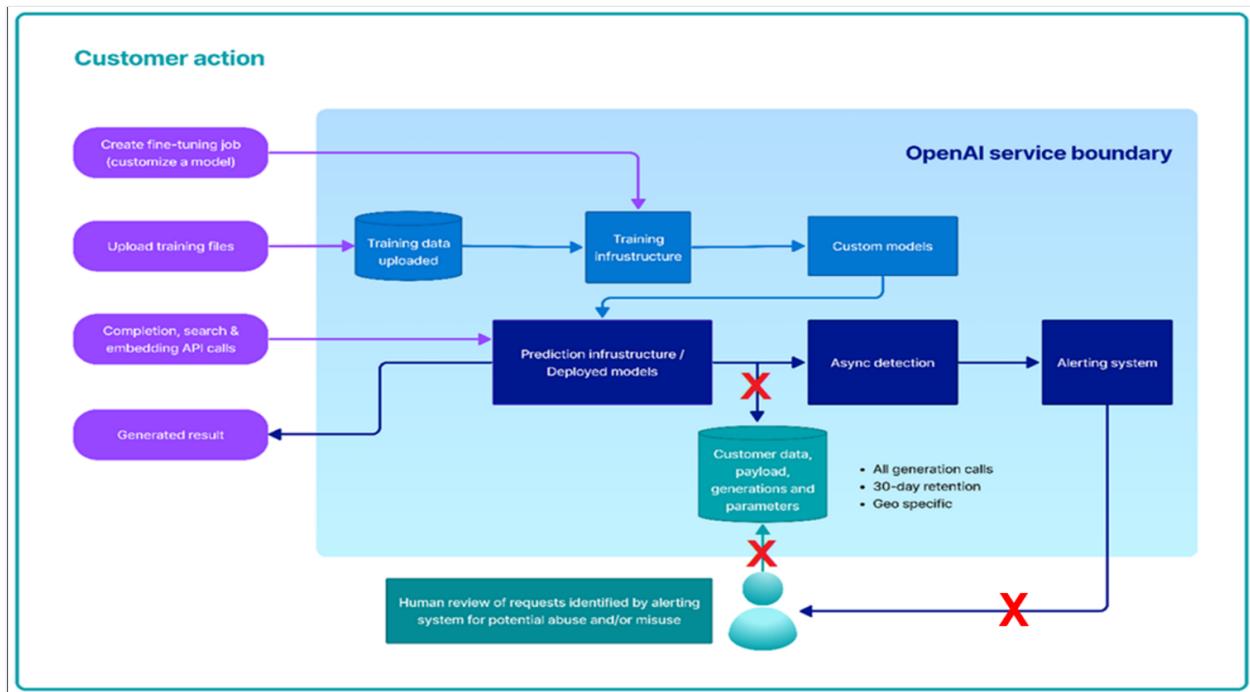
Can a customer opt out of the logging and human review process?

Some customers in highly regulated industries with low risk use cases process sensitive data with less likelihood of misuse. Because of the nature of the data or use case, these customers do not want or do not have the right to permit Microsoft to process such data for abuse detection due to their internal policies or applicable legal regulations.

To empower its enterprise customers and to strike a balance between regulatory / privacy needs and abuse prevention, the Azure Open AI Service will include a set of Limited Access features to provide potential customers with the option to modify following:

1. abuse monitoring
2. content filtering

These Limited Access features will enable potential customers to opt out of the human review and data logging processes subject to eligibility criteria governed by Microsoft's Limited Access framework. Customers who meet Microsoft's Limited Access eligibility criteria and have a low-risk use case can apply for the ability to opt-out of both data logging and human review process. This allows trusted customers with low-risk scenarios the data and privacy controls they require while also allowing us to offer AOAI models to all other customers in a way that minimizes the risk of harm and abuse.



If Microsoft approves a customer's request to access Limited Access features with the capability to (i) modify abuse monitoring and (ii) modify content filtering, then Microsoft will not store the associated request or response. Since no request or response data will be stored at rest in the Service Results Store in this case, the human review process will no longer be feasible. Therefore, both CMK and Lockbox will be deemed out-of-scope for harm and abuse detection.

See also

- [Limited access to Azure OpenAI Service](#)
- [Code of conduct for Azure OpenAI Service integrations](#)
- [Transparency note and use cases for Azure OpenAI Service](#)
- [Characteristics and limitations for Azure OpenAI Service](#)
- Report abuse of Azure OpenAI Service through the [Report Abuse Portal](#) ↗
- Report problematic content to cscraireport@microsoft.com

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure OpenAI 服务 REST API 参考

项目 · 2023/03/12

本文详细介绍 Azure OpenAI (Azure 认知服务套件中的一项服务) 的 REST API 终结点。 REST API 分为两类：

- **管理 API**：Azure 资源管理器 (ARM) 在 Azure 中提供了管理层，可用于在 Azure 中创建、更新和删除资源。所有服务都对这些操作使用通用结构。 [了解详细信息](#)
- 服务 API：Azure OpenAI 提供一组 REST API，用于与通过管理 API 部署的资源和模型进行交互。

管理 API

Azure OpenAI 作为 Azure 认知服务的一部分进行部署。所有认知服务都依赖于同一组管理 API 来执行创建、更新和删除操作。管理 API 还用于在 OpenAI 资源中部署模型。

[管理 API 参考文档](#)

身份验证

Azure OpenAI 提供两种身份验证方法。可以使用 API 密钥或 Azure Active Directory。

- **API 密钥身份验证**：对于这种类型的身份验证，所有 API 请求都必须在 `api-key` HTTP 标头中包含 API 密钥。本[快速入门](#)提供有关如何通过此类身份验证进行调用的教程
- **Azure Active Directory 身份验证**：可以使用 Azure Active Directory 令牌对 API 调用进行身份验证。身份验证令牌作为 `Authorization` 标头包含在请求中。提供的令牌必须以 `Bearer` 开头，例如 `Bearer YOUR_AUTH_TOKEN`。可以阅读有关如何[使用 Azure Active Directory 进行身份验证](#)的操作指南。

REST API 版本控制

服务 API 使用 `api-version` 查询参数进行版本控制。所有版本都遵循 YYYY-MM-DD 日期结构。例如：

HTTP

POST

https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2022-12-01

我们目前提供以下版本：`2022-12-01`

完成

通过完成操作，模型将根据提供的提示生成一个或多个预测完成。该服务还可以返回每个位置的替代令牌的概率。

创建完成

HTTP

POST `https://{{your-resource-name}}.openai.azure.com/openai/deployments/{{deployment-id}}/completions?api-version={{api-version}}`

路径参数

参数	类型	必需？	说明
<code>your-resource-name</code>	字符串	必须	Azure OpenAI 资源的名称。
<code>deployment-id</code>	字符串	必须	模型部署的名称。必须先部署模型，然后才能进行调用。
<code>api-version</code>	字符串	必须	要用于此操作的 API 版本。它遵循 YYYY-MM-DD 格式。

支持的版本

- `2022-12-01`

请求正文

参数	类型	必 需？	默认 值	说明
<code>prompt</code>	字符串 或数组	可选	<code><\ endoftext\ \></code>	生成完成的提示，其编码为字符串、字符串列表或令牌列表。请注意， <code><\ endoftext\ \></code> 是模型在训练期间看到的文档分隔符，因此如果未指定提示，则模型将像从新文档的开头一样生成。

参数	类型	必需 需?	默认	说明
max_tokens	integer	可选	16	在完成中生成的最大令牌数。 提示加上 max_tokens 的令牌计数不能超过模型的上下文长度。 大多数模型的上下文长度为 2048 个令牌 (davinci-codex 除外，它的长度为 4096 个)。
temperature	数字	可选	1	要使用的采样温度。 较高的值意味着模型将承担更多风险。 尝试将值设为 0.9 以获得更有创意的应用程序，将值设为 0 (argmax sampling) 以获得具有明确答案的应用程序。 我们通常建议更改此设置或 top_p，但不要同时更改这两者。
top_p	数字	可选	1	温度采样的替代方法，称为核采样，其中模型考虑具有 top_p 概率质量的令牌的结果。 所以 0.1 意味着只考虑包含前 10% 概率质量的令牌。 我们通常建议更改此设置或温度，但不要同时更改这两者。
n	integer	可选	1	要为每个提示生成的完成数。 注意：由于此参数会生成许多完成，因此可能会快速消耗你的令牌配额。 谨慎使用并确保对 max_tokens 和 stop 进行了合理的设置。
stream	boolean	可选	False	是否流式传输回部分进度。 如果设置，令牌将在可用时作为仅数据服务器发送的事件发送，流由数据终止：[DONE] 消息。
logprobs	integer	可选	Null	包含有关 logprobs 最有可能的令牌和已选择的令牌的日志概率。 例如，如果 logprobs 为 10，则 API 将返回包含 10 个最有可能的令牌的列表。 API 将始终返回采样令牌的 logprob，因此响应中可能最多有 logprobs+1 个元素。 此参数不能与 gpt-35-turbo 一起使用。
echo	boolean	可选	False	除了完成之外，还要回显提示。 此参数不能与 gpt-35-turbo 一起使用。
stop	字符串 或数组	可选	Null	最多四个序列，其中 API 将停止生成进一步的令牌。 返回的文本不包含停止序列。

参数	类型	必需 需?	默认	说明
presence_penalty	数字	可选	0	介于 -2.0 到 2.0 之间的数字。正值会根据它们到目前为止是否出现在文本中来惩罚新令牌，从而增加模型谈论新主题的可能性。
frequency_penalty	数字	可选	0	介于 -2.0 到 2.0 之间的数字。正值会根据它们到目前为止是否出现在文本中来惩罚新令牌，从而增加模型谈论新主题的可能性。
best_of	integer	可选	1	在服务器端生成 best_of 完成并返回“最佳”（每个令牌的日志概率最低的参数）。无法流式传输结果。与 n 一起使用时，best_of 控制候选完成数，n 指定返回的完成数，best_of 的值必须大于 n 的值。注意：由于此参数会生成许多完成，因此可能会快速消耗你的令牌配额。谨慎使用并确保对 max_tokens 和 stop 进行了合理的设置。此参数不能与 gpt-35-turbo 一起使用。
logit_bias	map	可选	Null	修改指定令牌在完成中出现的可能性。接受 json 对象，该对象将令牌（由其在 GPT tokenizer 中的令牌 ID 指定）映射到从 -100 到 100 的关联偏差。可以使用此 tokenizer 工具（适用于 GPT-2 和 GPT-3）将文本转换为令牌 ID。在数学上，采样之前会将偏差添加到由模型生成的 logit 中。具体效果因模型而异，但 -1 和 1 之间的值会减少或增加选择的可能性；-100 或 100 等值会导致相关令牌的禁止或独占选择。例如，可以传递 {"50256": -100} 以防止生成 < endoftext > 令牌。

示例请求

控制台

```
curl
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2022-12-01\
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d "{
  \"prompt\": \"Once upon a time\",
  \"max_tokens\": 100
}"
```

```
\\"max_tokens\\": 5
}"
```

示例响应

JSON

```
{
  "id": "cmpl-4kGh7iXtjW4lc9eGhff6Hp8C7btdQ",
  "object": "text_completion",
  "created": 1646932609,
  "model": "ada",
  "choices": [
    {
      "text": ", a dark line crossed",
      "index": 0,
      "logprobs": null,
      "finish_reason": "length"
    }
  ]
}
```

嵌入

获取给定输入的向量表示形式，该输入可由机器学习模型和其他算法轻松使用。

创建嵌入

HTTP

```
POST https://{{your-resource-name}}.openai.azure.com/openai/deployments/{{deployment-id}}/embeddings?api-version={{api-version}}
```

路径参数

参数	类型	必需？	说明
your-resource-name	字符串	必须	Azure OpenAI 资源的名称。
deployment-id	字符串	必须	模型部署的名称。必须先部署模型，然后才能进行调用
api-version	字符串	必须	要用于此操作的 API 版本。它遵循 YYYY-MM-DD 格式。

支持的版本

请求正文

参数	类 型	必 需?	默 认	说明
input	字 符 串 或 数 组	是	空 值	获取嵌入的输入文本，编码为字符串或令牌数组。要在单个请求中获取多个输入的嵌入，请传递一个由字符串构成的数组或一个由令牌数组构成的数组。每个输入的长度不得超过 2048 个令牌。目前，我们接受的最大数组数为 1。除非嵌入代码，否则我们建议将输入中的换行符 (\n) 替换为一个空格，因为我们观察到出现换行符时会产生较差的结果。
user	字 符 串	否	Null	表示最终用户的唯一标识符。这将帮助 Azure OpenAI 监视和检测滥用行为。不要传递 PII 标识符，而是使用伪名化值，例如 GUID

示例请求

控制台

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/embeddings?api-version=2022-12-01\  
-H "Content-Type: application/json" \  
-H "api-key: YOUR_API_KEY" \  
-d "{\"input\": \"The food was delicious and the waiter...\"}"
```

示例响应

JSON

```
{  
  "object": "list",  
  "data": [  
    {  
      "object": "embedding",  
      "embedding": [  
        0.018990106880664825,  
        -0.0073809814639389515,  
        .... (1024 floats total for ada)  
        0.021276434883475304,  
      ],  
      "index": 0  
    },  
  ],
```

```
  "model": "text-similarity-babbage:001"  
}
```

后续步骤

了解如何[使用 REST API 管理部署、模型和进行微调](#)。详细了解[为 Azure OpenAI 提供支持的基础模型](#)。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Deployments - Create

参考

Service: Cognitive Services

API Version: 2022-12-01

根据给定的规范为 Azure OpenAI 资源创建新部署。

本文内容

[URI 参数](#)

[请求头](#)

[请求正文](#)

[响应](#)

[安全性](#)

[示例](#)

[定义](#)

HTTP

```
POST {endpoint}/openai/deployments?api-version=2022-12-01
```

URI 参数

Name	In	Required	Type	Description
endpoint	path	True	string url	支持的认知服务终结点 (协议和主机名，例如： https://aoairesource.openai.azure.com 。将“aoairesource”替换为 Azure OpenAI 帐户名称)。
api-version	query	True	string	请求的 API 版本。

请求头

Name	Required	Type	Description
api-key	True	string	在此处提供认知服务 Azure OpenAI 帐户密钥。

请求正文

Name	Required	Type	Description
model	True	string	OpenAI 模型标识符 (要部署的 model-id)。可以是基础模型或微调。
scale_settings	True	ScaleSettings: ManualScale Settings Standard ScaleSettings	ScaleSettings 部署的缩放设置。 它定义缩放模式和预留容量。
error		Error	错误 Microsoft REST 准则中定义的错误内容 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。

响应

Name	Type	Description
201 Created	Deployment	已成功创建部署。 Headers Location: string
Other Status Codes	Error Response	出现了错误。

安全性

api-key

在此处提供认知服务 Azure OpenAI 帐户密钥。

Type: apiKey

In: header

示例

Creating a deployment.

Sample Request

HTTP

HTTP

```
POST https://aoairesource.openai.azure.com/openai/deployments?api-version=2022-12-01
```

```
{
  "scale_settings": {
    "capacity": 2,
    "scale_type": "manual"
  },
  "model": "curie"
}
```

Sample Response

Status code: 201

HTTP

```
location:
https://aoairesource.openai.azure.com/openai/deployments/deployment-afa0669ca01e4693ae3a93baf40f26d6
```

Response Body

JSON

```
{
  "scale_settings": {
    "capacity": 2,
    "scale_type": "manual"
  },
  "model": "curie",
  "owner": "organization-owner",
  "id": "deployment-afa0669ca01e4693ae3a93baf40f26d6",
  "status": "notRunning",
  "created_at": 1646126127,
  "updated_at": 1646127311,
  "object": "deployment"
}
```

定义

Deployment	部署
Error	错误
ErrorCode	ErrorCode
ErrorResponse	ErrorResponse
InnerError	InnerError
InnerErrorCode	InnerErrorCode
ManualScaleSettings	ManualScaleSettings
ScaleType	ScaleType
StandardScaleSettings	StandardScaleSettings
State	状态
TypeDiscriminator	TypeDiscriminator

Deployment

部署

Name	Type	Description
created_at	integer	创建此作业或项时的时间戳 (unix) ep。
error	Error	错误 Microsoft REST 准则中定义的错误内容 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。
id	string	此项的标识。
model	string	OpenAI 模型标识符 (要部署的 model-id)。 可以是基础模型或微调。
object	Type Discriminator	TypeDiscriminator 定义 对象的类型。
owner	string	此部署的所有者。 对于 Azure OpenAI，仅支持“组织所有者”。
scale_settings	ScaleSettings: ManualScale Settings Standard ScaleSettings	ScaleSettings 部署的缩放设置。 它定义缩放模式和预留容量。
status	State	状态 作业或项的状态。
updated_at	integer	在 unix epochs 中，上次 (修改此作业或项时的时间戳。

Error

错误

Name	Type	Description

code	ErrorCode Code	ErrorCode Microsoft REST 准则中定义的错误代码 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。
details	Error[]	错误详细信息 (如果可用)。
innererror	Inner Error	InnerError Microsoft REST 准则中定义的内部错误 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。
message	string	此错误的消息。
target	string	发生错误的位置 (如果可用)。

ErrorCode

ErrorCode

Name	Type	Description
conflict	string	请求的操作与当前资源状态冲突。
fileImportFailed	string	导入文件失败。
forbidden	string	当前用户/API 密钥禁止该操作。
internalFailure	string	内部错误。请重试。
invalidPayload	string	此操作的请求数据无效。
itemDoesAlreadyExist	string	该项已存在。
jsonValidationFailed	string	json 数据验证失败。
notFound	string	找不到资源。

quotaExceeded	string	已超出配额。
serviceUnavailable	string	该服务当前不可用。
unexpectedEntityState	string	操作不能在当前资源的状态下执行。

ErrorResponse

ErrorResponse

Name	Type	Description
error	Error	<p>错误</p> <p>Microsoft REST 准则中定义的错误内容 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。</p>

InnerError

InnerError

Name	Type	Description
code	Inner Error Code	<p>InnerErrorCode</p> <p>Microsoft REST 准则中定义的内部错误代码 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。</p>
innererror	Inner Error	<p>InnerError</p> <p>Microsoft REST 准则中定义的内部错误 (https://github.com/microsoft/api-guidelines/blob/vNext/Guidelines.md#7102-error-condition-responses)。</p>

InnerErrorCode

InnerErrorCode

Name	Type	Description
invalidPayload	string	此操作的请求数据无效。

ManualScaleSettings

ManualScaleSettings

Name	Type	Description
capacity	integer	此部署的推理终结点的常量保留容量。
scale_type	string: manual	ScaleType 定义缩放操作的执行方式。

ScaleType

ScaleType

Name	Type	Description
manual	string	可以通过手动指定模型的容量来缩放部署。
standard	string	根据使用情况自动缩放部署。

StandardScaleSettings

StandardScaleSettings

Name	Type	Description
scale_type	string: standard	ScaleType 定义缩放操作的执行方式。

State

状态

Name	Type	Description
canceled	string	操作已取消且不完整。
deleted	string	该实体已被删除，但仍可能被删除前的其他实体引用。
failed	string	操作已完成处理但失败，无法进一步使用。
notRunning	string	操作已创建，并且不会排队等待将来处理。
running	string	已开始处理操作。
succeeded	string	操作已成功处理并已准备好使用。

TypeDiscriminator

TypeDiscriminator

Name	Type	Description
deployment	string	此对象表示部署。
file	string	此对象表示文件。
fine-tune	string	此对象表示微调作业。
fine-tune-event	string	此对象表示微调作业的事件。
list	string	此对象表示其他对象的列表。
model	string	此对象表示模型 (可以是基础模型或微调作业结果)。

你目前正在访问 Microsoft Azure Global Edition 技术文档网站。如果需要访问由世纪互联运营的 Microsoft Azure 中国技术文档网站，请访问 <https://docs.azure.cn>。

Azure 认知服务的支持和帮助选项

项目 · 2022/12/05

你是否刚刚开始探索 Azure 认知服务的功能？也许你正在应用程序中实现一项新功能。或者，在使用该服务后，你是否对如何改进它有任何建议？对于认知服务，你可以通过以下方式获取支持、随时了解最新信息、提供反馈和报告 bug。

创建 Azure 支持请求



浏览 [Azure 支持选项的范围并选择最适合的计划](#)，无论你是刚刚开始使用云的开发人员，还是部署业务关键型战略应用程序的大型组织。Azure 客户可在 Azure 门户中创建和管理支持请求。

- [Azure 门户](#)
- [适用于美国政府的 Azure 门户](#)

在 Microsoft Q&A 上发布问题

若要快速地从 Microsoft 工程师、Azure 最有价值专家 (MVP) 或我们的专家社区那里获得技术产品问题的可靠答案，请在 [Microsoft Q&A](#) 上与我们联系 - 这是 Azure 的首选社区支持位置。

如果通过搜索无法找到问题的解答，请向 Microsoft Q&A 提交新问题。提问时，请使用以下标记之一：

- [认知服务](#)

影像

- [计算机视觉](#)
- [自定义视觉](#)
- [人脸](#)
- [表单识别器](#)
- [视频索引器](#)

语言

- [沉浸式阅读器](#)
- [语言理解 \(LUIS\)](#)
- [QnA Maker](#)
- [语言服务](#)
- [翻译](#)

语音

- [语音服务](#)

决策

- [异常检测器](#)
- [内容审查器](#)
- [指标顾问](#)
- [个性化体验创建服务](#)

Azure OpenAI

- [Azure OpenAI](#)

在 Stack Overflow 上发布问题



要通过最大的社区开发者生态系统获取开发人员问题的解答，请在 Stack Overflow 上提问。

如果确实要向 Stack Overflow 提交新问题，请在创建问题时使用以下一个或多个标记：

- [认知服务](#)

影像

- [计算机视觉](#)
- [自定义视觉](#)
- [人脸](#)
- [表单识别器](#)
- [视频索引器](#)

语言

- [沉浸式阅读器](#)
- [语言理解 \(LUIS\)](#)

- [QnA Maker](#)
- [语言服务](#)
- [翻译](#)

语音

- [语音服务](#)

决策

- [异常检测器](#)
- [内容审查器](#)
- [指标顾问](#)
- [个性化体验创建服务](#)

Azure OpenAI

- [Azure OpenAI](#)

提交反馈

若要请求新功能，请在 <https://feedback.azure.com> 中发布请求。针对如何让认知服务及其 API 更好地服务于你开发的应用程序，分享你的想法。

- [认知服务](#)

影像

- [计算机视觉](#)
- [自定义视觉](#)
- [人脸](#)
- [表单识别器](#)
- [视频索引器](#)

语言

- [沉浸式阅读器](#)
- [语言理解 \(LUIS\)](#)
- [QnA Maker](#)
- [语言服务](#)
- [翻译](#)

语音

- [语音服务](#)

决策

- [异常检测器](#)
- [内容审查器](#)
- [指标顾问](#)
- [个性化体验创建服务](#)

随时获取最新信息

在 Azure 博客中随时了解新版本的功能或新闻，这可以帮助你了解编程错误、服务 bug 与认知服务中尚未提供的功能之间的差异。

- 在 [Azure 更新](#) 中详细了解产品更新、路线图和公告。
- 有关认知服务的新闻在 [Azure 博客](#) 中共享。
- 加入 [Reddit 上有关认知服务的对话](#)。

后续步骤

[什么是 Azure 认知服务？](#)