Microsoft Azure

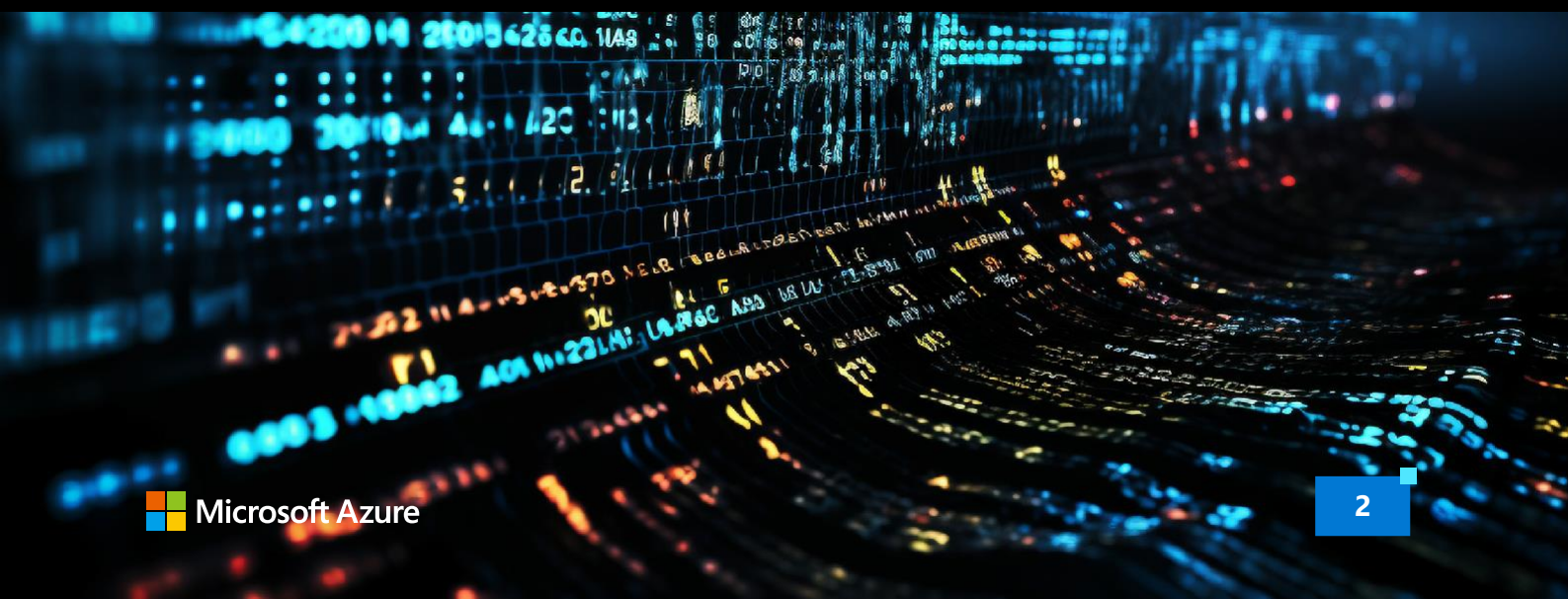
Azure OpenAI
生成式人工智能白皮书

前言

生成式 AI 成为人工智能领域新的关键词。吸纳从机器智能到机器学习、深度学习的关键技术，生成式 AI 更进一步，能够根据提示或现有数据创建新的书面、视觉和听觉内容。在此基础上，大模型和大模型应用一时涌现，并迅速确立 AI 落地新范式。据 data.ai intelligence 数据显示，2023 年生成式 AI 应用激增。2024 年 AI 将会带动 10% 的应用下载量，包含 GenAI 功能的应用下载量将同比增长 40%。

面对生成式 AI 及其应用落地的迅猛发展，微软期待用 AI 重新定义软件开发与工作的未来。从 Azure OpenAI、Copilot Stack、开发工具到协作应用等领域，微软将 AI 融入现有的软件和服务生态，从提供 AI 工具到构筑 AI 平台和生态，全方位帮助开发者应对技术革命，予力人们运用 AI，让每个人都可以在工作和生活中，受益于这些突破性技术，探索全新机遇与无限可能。微软的 AI 战略包括三个部分：将 AI 副驾融入每个微软云解决方案、帮助客户通过 AI 创新和转型以及负责任的 AI。

本白皮书将全面介绍 Microsoft Azure 生成式人工智能领域的解决方案、工具指南、最佳实践以及支持 AI 的云端算力及架构优势，旨在帮助处于 AI 不同阶段的客户选择适合您战略的落地路径，本白皮书介绍的落地方案涵盖从直接使用微软 Azure AI 和 Azure OpenAI 服务，到自建大模型等四层路径。



目录

第一章
Azure AI 及 Azure OpenAI 服务概括

Azure AI 服务

Azure AI 服务包括了 Azure 机器学习平台、认知服务和应用 AI 服务。其中，Azure 认知服务有五大支柱，分别是视觉、语音、语言、决策和 Azure OpenAI 服务 。

Azure AI 服务通过现成的预生成可定制的 API 和模型，帮助开发人员和组织快速创建智能、前沿、面向市场且负责任的应用程序，包括对话、搜索、监视、翻译、语音、视觉和决策的自然语言处理。这意味着，您可以使用 Azure AI 组合构建企业规模的智能应用，并使用生成式 AI 重新构想你的业务。

Azure AI 组合

Azure OpenAI 服务

生成自己的助手和生成式 AI 应用程序。

Azure AI 搜索

基于你自己的数据构建、微调和训练自定义 AI 模型，获得独特的优势。

Azure AI Content Safety

在应用程序和服务中检测用户生成的和 AI 生成的有害内容。

负责任的 AI 仪表板

使用统一的仪表板实践负责任、高质量的 AI，能够轻松评估和调试机器学习模型。

Azure AI 提示流

创建可执行的 Prompt Flow，通过可视化图形链接大型语言模型、提示和 Python 工具。

机器学习运营 (MLOps)

加速机器学习工作流的自动化、协作和可重现性。

在 Azure AI 的帮助下，您可以为您的组织构建可以立即推向市场的先进应用程序。以微软为例，Microsoft 产品及应用平台如 Microsoft 365 Copilot、Dynamics 365 Copilot、Power Platform 都由 Azure AI 驱动，为每个人和组织提效生产力。

微软 AI 服务一览

Azure OpenAI 服务

微软 Azure 作为 OpenAI 的独家云服务提供商，自 2019 年开始便以出色的、面向 AI 时代的领先架构，为 OpenAI 的快速发展提供助力。基于双方的战略合作，现在 Azure 全球版客户可以通过 Azure OpenAI 服务直接调用 OpenAI 全部模型，包括 ChatGPT、GPT-4、GPT-4 vision、Codex 和 DALL.E 等模型，并享有 Azure 企业级 99.9% 的可用性 SLA、企业级安全保障和为人工智能优化的基础设施。目前有 18,000 多家组织使用 Azure OpenAI 服务。

Azure OpenAI 目前为各类企业提供两类服务模式：

- **Pay As You Go 即用即付型模式：**

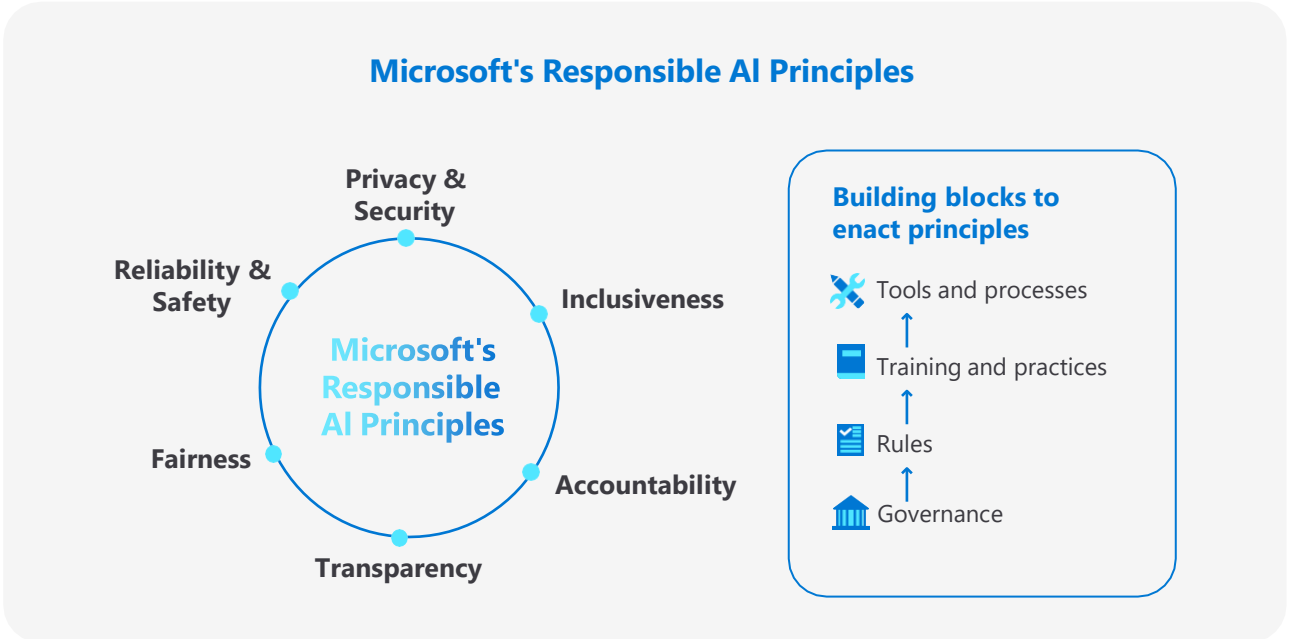
  适合各类 AI 场景的测试及上线应用，可在 Azure Portal 后台直接启用。

- **预配置吞吐容量的专属模式：适合各类追求稳定性能及稳定延时的 AI 生产应用。**
  - 通过提供稳定的最大延迟和吞吐量保障稳定可预测的性能
  - 预留专属的容量
  - 成本节省：与按消耗 Tokens 计费的 Pay As You Go 即用即付模式相比，可节省成本

负责任的人工智能原则

在科技行业里，微软率先提出了打造负责任的人工智能的决心。2016 年，微软 CEO 萨提亚纳德拉发表了一篇关于人工共同责任的专栏文章，几个月后，第一次公开提出了微软的人工智能准则：公平、可靠和安全、隐私和保障、包容、透明、责任。

当 AI 进入主流产品和服务时，这些原则对于创建负责任且值得信赖的 AI 至关重要。它们以两个角度为指导：道德和可解释。阅读标题链接可阅读详细细节。



第二章
生成式人工智能落地实践的四种路径

微软为各行业客户 根据 AI 战略阶段不同提供四个层次的 AI 创新支持：

结合您的现有数据直接使用 Azure OpenAI 模型：适合各类 AI 场景的实现 1

Prompt engineer 提示工程优化：提高大语言模型生成响应的准确性 2

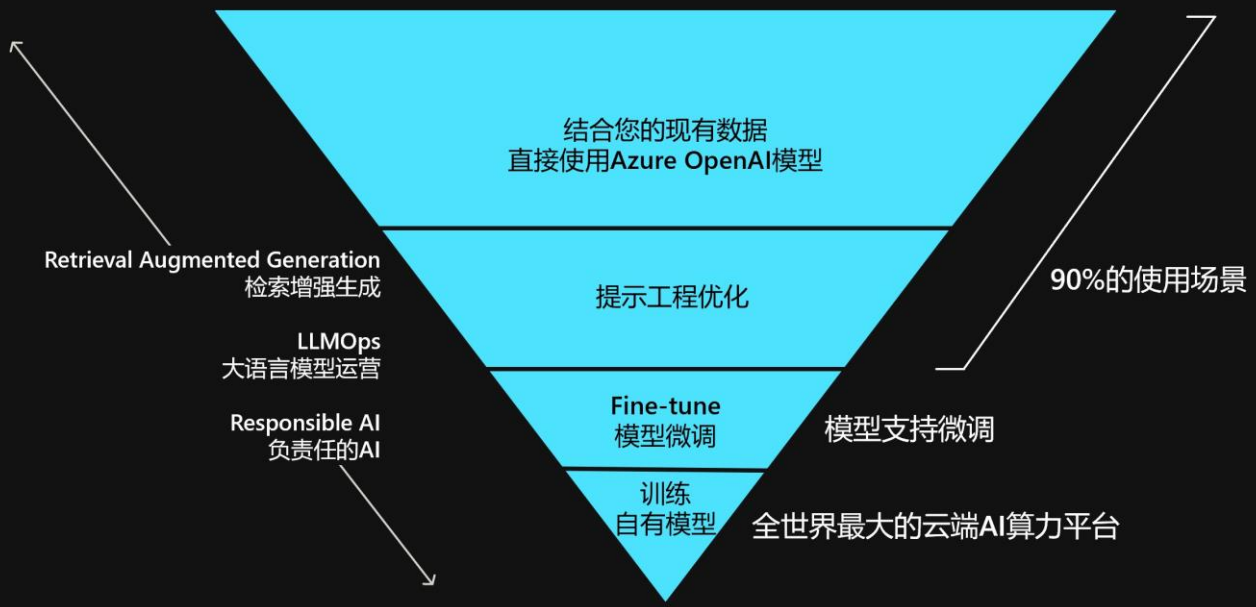
基于现有模型进行 Fine-tuning 微调：使用示例数据重新训练现有的大型语言模型，从而生成使用提供的示例经过优化的新的“自定义”大型语言模型。 3

训练您的自有模型：Azure AI Infra 云端大规模 AI 算力平台方案 4

其中，90% 的使用场景可以通过前两层的落地方案实现。在任何路径下，您都可以通过使用检索增强生成 (RAG) 模式，与预训练的大型语言模型 (LLM) 和你自己的数据配合使用以生成响应。

同时，也可以通过使用 Azure 大语言模型运营功能 (LLMOps) 实现高效提示工程和开发部署，以简化的过程来管理大模型应用的端到端生命周期。

微软为各行业提供四个层次的 AI 创新支持



路径一 直接使用 Azure OpenAI 模型：添加您的数据至 Azure OpenAI 模型

1.1 Azure OpenAI 服务上的可用模型

语言模型

GPT-4-0125-preview：GPT-4-0125-preview 已发布，新模型旨在减少模型未完成任务的“懒惰”情况，及其他升级。

GPT-4：国际版 Azure OpenAI 服务的客户和合作伙伴可以申请访问 GPT-4，并开始使用 OpenAI 目前最先进的模型构建应用。通过基于 Azure AI 优化的基础设施、企业级可用性、合规性、数据安全和隐私控制提供的支持，以及与其它 Azure 服务的多种集成，实现您的大模型应用架构。

GPT-4-Turbo：GPT-4 Turbo 功能更强大，模型信息更新到 2023 年 4 月。它具有 128K 上下文窗口，因 此您可以使用 RAG（检索增强生成）等技术，基于企业用例所需的自定义数据定制应用程序。

GPT-3.5-Turbo：GPT-3.5-Turbo 现已更新至 1025 版本。

Assistants API

客户可以开启 Code Interpreter, Function calling 等工具在自己的应用程序内构建拥有指令的 AI 助手，并利用模型、工具和知识来响应用户查询。Azure Assistants API 目前已支持代码解释器和函数调用，检索功能将很快发布。

TTS 模型

OpenAI 的 TTS 模型在 AOAI 和 Azure AI Speech 同时上线。新的 TTS 模型能生成 6 种预设不同个性和风格的人类品质语音。

多模态模型

GPT-4-Turbo with Vision：GPT-4 Turbo with Vision 是由 OpenAI 开发的大型多模态模型（LMM），支持图像分析并能对图像有关问题生成文本响应。它结合了自然语言处理和视觉理解能力。GPT-4 Turbo with Vision达到了图像理解的先进水平。它不仅仅能够识别图片中的对象，更注重理解上下文和细节，比如创建详细的图像标题、提供丰富的上下文描述、回答关于视觉内容的问题或分配智能标签。

微调

Azure OpenAI 服务推出了三款模型的 Fine-tuning 功能（Babbage-002、Davinci-002 和 GPT-3.5-Turbo）。用户可以使用 Azure OpenAI 服务或 Azure 机器学习对 Babbage/Davinci-002 和 GPT-3.5-Turbo 进行 Fine-tuning。GPT-3.5-Turbo 1106 模型已支持 fine-tuning, 同时 Training 和 Hosting 的成本比 GPT-3.5-Turbo 都降低 50%。Babbage-002 和 Davinci-002 支持 completion，Turbo 支持对话式交互。通过几个简单的命令，您就可以指定基本模型、提供数据、进行训练和部署。点击阅读关于微调的介绍文章

图像

DALL·E 3：DALL · E 3 是一种支持文本提示的图像生成模型，助力用户探索创意表达的新领域，通过语言和视觉艺术的交融提供独特的体验。提示越详细，图像效果就越好。您甚至可以在已创建的图像中添加文本。

转录和翻译

Whisper：Whisper 模型提供转录和翻译音频内容的功能，同时支持大规模高质量的批量转录。

1.2 添加您的数据

Azure OpenAI Service on your data，基于这项全新功能，用户可以使用自己的数据驱动 OpenAI 模型，无需训练或微调，即可释放全部数据潜力。

Azure OpenAI Service on your data 能够获取并打通所有来源的数据。无论数据是存储在本地还是云端，该功能将提供无缝连接以解锁数据的全部潜力。借助这一先进工具，您可以高效处理、组织、优化数据，获得有价值、高质量的洞察。同时，用户友好的 API 和 SDK，将与您的现有系统轻松集成；定制化示例应用程序助力快速部署。此外，数据共享和利用也将变得更轻松，您可以在企业内部或面向客户快速实现信息分发。

Azure OpenAI Service on your data连接多个数据源，包括：

- **Azure 认知搜索索引：**您可以将数据连接到 Azure 认知搜索索引，实现与 OpenAI 模型的无缝集成。
- **Azure Blob 存储容器：**将数据连接到 Azure Blob 存储容器，使用 Azure OpenAI 服务轻松获取数据，用于后续的分析和对话。
- **本地文件：**连接您的 Azure AI 门户文件，为数据连接提供灵活性和便利性。数据在摄取、切分之后，将导入 Azure 认知搜索索引。txt、md、html、Word 文件、PowerPoint、PDF 等格式的文件都可以用于分析和对话。

使用 Azure OpenAI 服务数据支持的步骤如下：

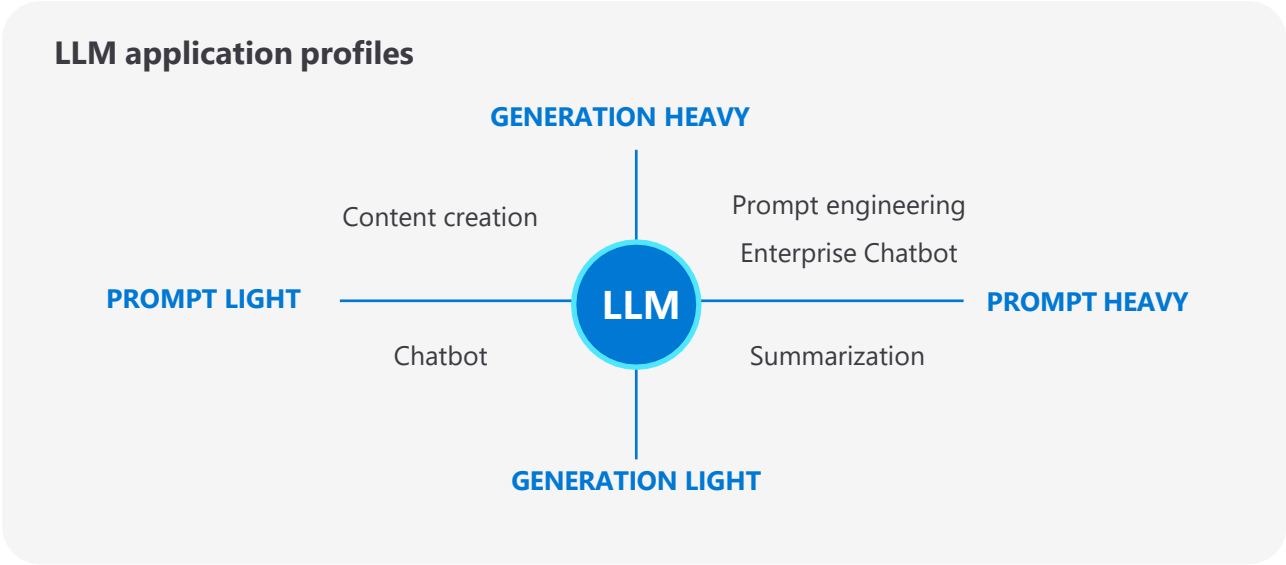
- **连接数据源：**使用 Azure AI Studio 连接您所需的数据源，可通过 Azure 认知搜索索引、Blob 存储容器、上传本地文件等途径完成连接。
- **基于数据进行提问和聊天：**基于数据进行提问和聊天：连接数据源之后，您就可以通过 Azure AI Studio，向 OpenAI 模型提问和对话。这将使您获得有价值的洞察，并在大量信息的支撑下做出商业决策。

1.3 结合 Azure OpenAI 的 Embedding 嵌入向量生成模型

将企业现有的结构化知识库与提示词引擎结合起来，让 GPT 模型生成更正确、更稳定、更可靠的结果。

Embedding 是一种特殊的数据表示格式，机器学习模型和算法可以轻松使用。 Embedding 是一段文本的语义含义的信息密集表示。Embedding 支持在 Azure 数据库中进行矢量相似性搜索，例如 Azure Cosmos DB for MongoDB vCore 或 Azure Database for PostgreSQL - 灵活服务器。

路径二 Prompt engineer 提示工程优化



在 AI 领域，特别是在大型语言模型中，提示指的是用户为引起特定类型的响应而给出的输入或指令。要充分利用 GPT-4 等大型语言模型，就必须精心设计能产生有效结果的提示。挑战在于如何选择词语、表达方式、符号和结构的最佳组合，以引导模型生成准确而贴切的内容，包括回答问题、以最喜欢的作家的口吻创作故事、创作诗歌、执行代码相关的任务等。

提示有助于 AI 明确用户意图和对其生成内容的期望，更精确的提示会带来更准确、相关性更强的结果。

需要注意的是，相似的提示可能会引发不同的响应，这取决于底层模型、训练数据，甚至是用户请求措辞的细微变化。

您可以使用本文介绍的 15 个提示工程优化技巧，跃升为更好的 AI 提示词工程师。

路径三 基于现有模型进行 Fine-tuning 微调

1. 什么是 Fine-tune 微调

Fine-tuning 是开发人员和数据科学家用来定制大型语言模型，以满足特定任务需求的方法之一。与“检索增强生成”（Retrieval Augmented Generation，RAG）和“提示工程”等方法通过在提示中注入正确的信息和指令不同，Fine-tuning 则是通过对大型语言模型本身进行个性化定制来实现的。

Azure OpenAI 服务和 Azure 机器学习提供了监督式 Fine-tuning，允许您提供自定义数据（提示 / 补全或对话式聊天，具体取决于选择的模型），以教授基本模型新的技能。

2. 何时应该使用微调

在开始使用 Fine-tuning 之前，我们建议您首先尝试提示工程或 RAG（检索增强生成）——这是最快的入手方式。微软提供了如 Prompt Flow 或 On Your Data 工具来使此过程变得更简单。您可以在需要 Fine-tuning 模型的场景下进行比较，以选择从提示工程还是 RAG 开始入手。大多数模型都会结合提示工程和 Fine-tuning，从而避免浪费精力。

想要了解何时 / 是否应该进行 Fine-tuning？

一些基本规则可以为您提供指导：

1. 如果您希望简单而快速地获取结果，请不要立即开始使用 Fine-tuning：因为这需要大量的数据和时间来训练和评估新模型。如果时间有限，通常只需通过提示工程即可取得相当大的进展。

2. 如果您需要最新或域外数据，这是使用 RAG 和提示工程的完美用例。

3. 如果您希望确保您的模型具有良好的基础，避免产生幻觉（hallucinations），那么 RAG 在这方面表现出色。

可以考虑使用 Fine-tuning 的场景包括：

1. 教会模型一项新技能，以便它在特定任务上表现出色，如分类、摘要或始终以特定格式或语调进行回应。有时，通过 Fine-tuning 较小的模型，也能使其在特定任务上与较大的模型一样出色。

2. 通过示例向模型演示如何执行某些操作，但在提示中很难解释清楚，或者示例太多，上下文窗口中无法容纳。这些场景有很多边缘情况，比如自然语言查询，或者教模型用特定的声音或语调说话。

3. 减少延迟。较长的提示可能需要更长的处理时间，而 Fine-tuning 允许您将这些较长的提示集成到模型本身中。

微调与提示工程、RAG 的使用情景一览

要求	开始	为什么
用几个示例来引导模型	提示工程	易于制作和快速实验，门槛非常低
简单、快速部署	提示工程、RAG	在 Azure OpenAI 上轻松使用您的数据，使用提示流程，LangChain
提升模型相关性	RAG	从您自己的数据集中检索相关信息以插入提示
最新数据信息	RAG	从您自己的数据库、搜索引擎等查询最新信息，以插入提示
事实基础	RAG	引用和检查检索到的数据的能力
优化具体任务	微调	微调经常用于引导模型执行特定任务，例如以特定的格式、语调或声音总结数据
指令无法适应提示	微调	微调将少样本示例移入训练步骤，但增加了训练所需的样本数量
更低的延迟	微调	在提供足够且高质量的数据的情况下，可以微调较小、更快的模型以在特定任务上提供良好的性能，而不是使用类似 GPT4 的通用模型
更低的成本	适情况而定	提示工程和 RAG 的初始成本较低，但长提示更昂贵；微调训练是昂贵的，但可能会减少提示长度。选择始终取决于用例和数据。
复杂、新颖的数据或领域	提示工 +RAG+ 微调	这是一个高风险领域。微调可以重新训练模型以识别新领域，但需要使用 RAG 来避免合理的混淆。确保客户不尝试为未经批准的用途重新训练

微调是一种高级技术，需要专业知识才能正确使用。下面的问题将帮助您评估是否已经准备好使用微调。

1. 为什么要微调?

2. 您目前尝试了哪些其他办法?

3. 这些方法哪些情况下行不通?

4. 你打算使用哪些数据进行微调?

5. 如何衡量微调模型的质量?

如果您可以明确回答以上问题，那么，您可以开始下一步的微调计划了。

3.3 在 Azure 上进行微调

使用 Azure Open AI 服务进行 Fine-tuning 将为您带来双重优势：您可以自定义先进的 OpenAI 大语言模型 (LLMs)，同时部署在 Azure 的安全、企业级云服务上。Azure 的内容审查功能让您使用所需的数据进行 Fine-tuning 时，可以过滤掉任何有合规风险的数据及响应。

如果您是 Azure OpenAI 服务和 LLMs 的新用户，欢迎您使用微软提供的简单易用的 API 来训练和部署模型。如果您更愿意使用 GUI，可以尝试 Azure OpenAI Studio。如果您正在从 OpenAI 迁移到 Azure，API 可两者兼容。

Fine-tuning 分为两个部分：训练 Fine-tuning 模型以及使用新定制的模型进行推理。

训练

指定您的基本模型、训练和验证数据，并设置超参数，就可以开始了！您可以使用 Azure OpenAI Studio 进行简单的 GUI 操作，或者对于更高级的用户，我们还提供了 REST API 或 OpenAI Python SDK。

推理

训练完成后，新模型将在您的资源中可用。当使用您的模型用于推理时，该定制模型可像任何其他 OpenAI LLM 一样进行部署！

经过 Fine-tuning 的模型部署后将按小时收取托管费用，基于输入和输出 token 计价。

如果您熟悉使用 Azure Machine Learning Studio 来开发、监控和部署模型，您可以将 Fine-tuning 集成至 AML 工作流程中的现有模型。除了 OpenAI 模型，Azure 机器学习还支持对开源模型进行 Fine-tuning，例如 LLaMa。

YouTube 视频：Alicia & Seth “如何在 Turbo 上进行微调”

路径四 训练您的自有模型

如果您决定训练您的自有模型，Azure 云端大规模 AI 算力平台是您最佳的合作伙伴！

先进 AI，离不开算力基础设施、服务与专业知识。我们将微软过去十年的超级计算经验和支持超大型 AI 训练工作负载的经验应用于搭建具备规模化高性能的 AI 基础架构。

您可以在 Azure AI Studio 中使用预构建和可自定义的 AI 模型，开发生成式 AI 解决方案和自定义 copilot。

现在 Azure AI 模型目录中添加了新的基础和生成式 AI 模型。在 Hugging Face 中，我们引入了一系列不同的稳定扩散模型、falcon 模型、CLIP、Whisper V3、BLIP 和 SAM 模型。此外，我们还添加了 分别来自 Meta 和 NVIDIA 的 Code Llama 和 Nemotron 模型，以及微软研究的尖端 Phi 模型。模型目录中新增了 40 个新模型和 4 种新模式，包括文本到图像和图像嵌入模型。

借助我们的模型即服务，专业开发人员很快就能轻松集成最新的 AI 模型，例如 Meta 的 Llama 2、Cohere 的 Command、G42 的 Jais 以及 Mistral 的高级模型作为 API 端点到应用程序中。他们还可以使用自己的数据微调这些模型，Azure 庞大的 GPU 基础设施能力，将帮助您降低配置 GPU 资源和管理托管的复杂性。

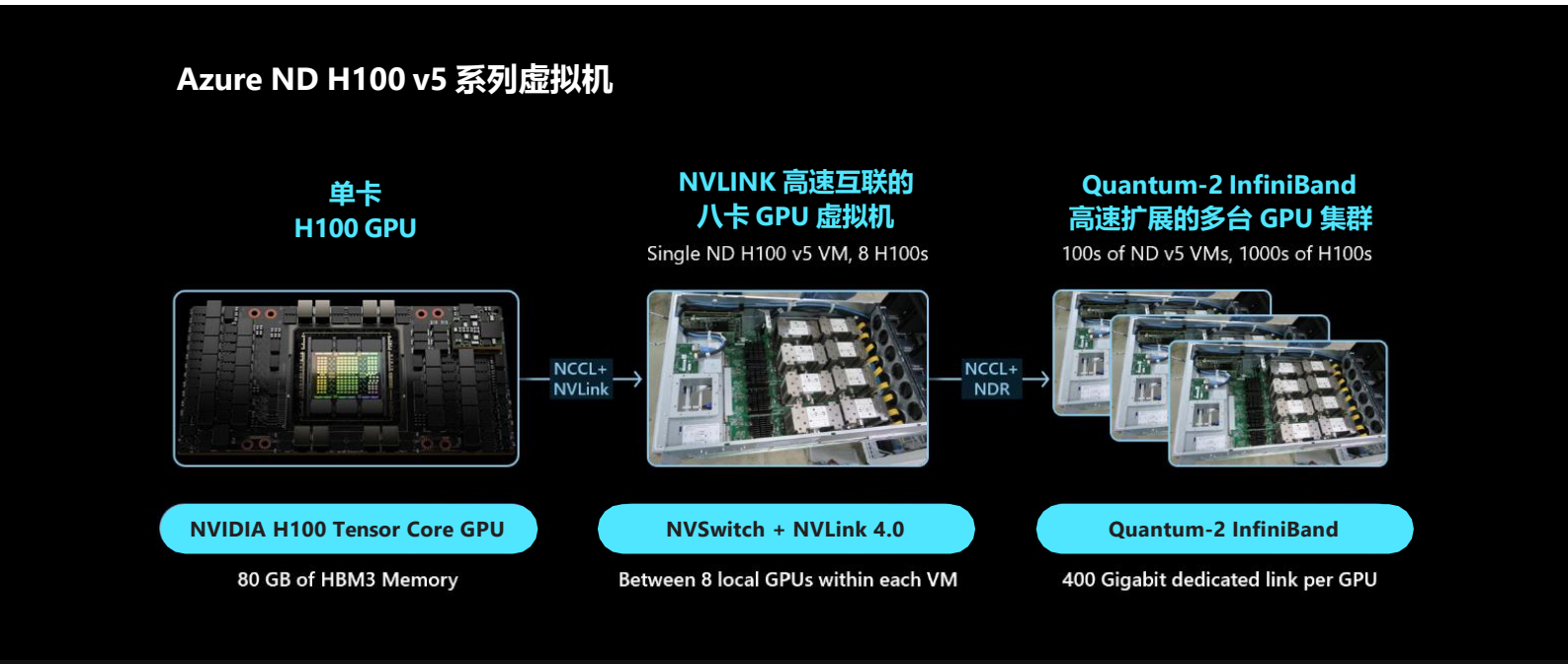
4.1 Azure 为任何规模的 AI 提供世界一流的算力

Azure 领先的 GPU 和网络解决方案的独特架构设计，为计算最密集的 AI 训练和推理工作负载提供了一流的性能和规模。这就是世界领先的 AI 公司（包括 OpenAI、Meta 等）选择 Azure 来推进 AI 创新的原因。Azure 目前位列全球 TOP500 超级计算平台第三名，是唯一的云服务提供商。

4.2 Azure AI Infra GPU 虚拟机类型

Microsoft Azure 推出基于 NVIDIA H100 Tensor Core GPU 的 ND H100 v5 VM，该虚拟机支持的按需配置可达 8 至上千个通过 Quantum-2 InfiniBand 网络互连的 NVIDIA H100 GPU，使得 AI 模型的性能明显提高。

在 2023 年 11 月份发布的 MLperf 性能基准报告里，我们将 175B 参数 GPT-3 模型的训练时间从之前的 10.94 分钟降低到 4.0085 分钟，使用的算力是 1344 台通过 InfiniBand 互联的 Azure H100 虚机，这只比 Nvidia 裸金属机器慢了 2%。



此外 Microsoft Azure 的 AI Infra 机型还包括：

基于 Nvidia H100 用于中等规模模型训练以及推理密集型负载的 H100 虚拟机类型 NC_H100。

基于 Nvidia H100 用于机密型 AI 应用的 H100 机型 NCC H100 v5。

基于 Nvidia H200 用于大模型训练和推理的 ND H200 v5 机型。

基于 AMD MI300X 用于支持大模型的高带宽内存 GPU 机型 ND MI300X v5 虚拟机。

> **与 Azure 共同设计超级计算机对于扩展我们苛刻的 AI 训练需求至关重要，使我们在 ChatGPT 等系统上的研究和调试工作成为可能。**
>
> **—— Greg Brockman**
>
> OpenAI 总裁兼联合创始人

> **我们对于对话式 AI 的关注促使我们开发和训练一些最复杂的大型语言模型。Azure 的 AI 基础结构为我们提供了必要的性能，以便大规模、可靠地、高效处理这些模型。我们对 Azure 的新 VM 及其将为我们的 AI 开发工作带来的性能提升感到非常兴奋。**
>
> **—— Mustafa Suleyman**
>
> Inflection 首席执行官

第三章
生成式人工智能落地
成功案例参考

01 汽车行业 梅赛德斯 - 奔驰

梅赛德斯 - 奔驰（奔驰）是世界知名的汽车品牌，致力于为用户提供更互联、更智能、更个性化的驾驶体验。为了实现这一目标，奔驰开发了 MBUX 语音助手“嘿，梅赛德斯”，让用户可以通过自然语音命令控制车内的各种功能，如导航、娱乐、空调等。这种语音助手不仅可以提高用户的便捷性和生产力，还可以增强用户的安全性和舒适度。

企业需求

1. 奔驰希望尽快将 GPT 模型应用到 MBUX 语音助手中，以抢占市场先机，但 GPT 模型的部署和集成需要大量的时间和资源。
2. 奔驰需要确保 GPT 模型与 MBUX 语音助手的兼容性和稳定性，以及与其他车内系统的协调性。
3. 奔驰需要保证 GPT 模型的安全性和可靠性，以及用户的隐私和数据的保护。

使用 Azure OpenAI 服务

微软 Azure OpenAI 服务为奔驰提供了一个集成了 GPT 模型的车载语音助手，让用户可以通过自然语言与车辆进行交互。这个语音助手可以理解用户的需求和意图，提供全面和相关的回答，还可以集成第三方服务，让用户可以在驾驶过程中完成更多的任务。这个方案通过 Azure 的企业功能和基础设施，保证了语音助手的性能和安全性，同时遵循了微软的负责任的人工智能原则。这个方案正在美国进行 Beta 测试，预计将为奔驰用户带来更互联、更智能、更个性化的驾驶体验。

成功亮点

奔驰可以在三个月内完成 GPT 模型的 Beta 测试，并根据用户的反馈进行改进和优化，从而快速推出增强版的 MBUX 语音助手；

奔驰可以使用 Azure OpenAI 服务的预留预配置吞吐量功能，以便大规模控制 GPT 模型的配置和性能，从而支持更多的用户和车型；

奔驰可以通过 GPT 模型提供更直观、更具对话感的语音助手，让用户可以通过自然语音命令完成更多的任务和功能，从而提升用户的满意度和忠诚度。

02 零售行业 沃尔玛

沃尔玛是全球最大的零售商，以低价和多样化的商品吸引了数亿消费者。沃尔玛在 2023 财年实现了超过 820 亿美元的电子商务销售额，并持续扩大其活跃数字客户群。该公司的目标是为顾客提供最佳的数字购物体验，让他们能够轻松地找到和购买自己想要的产品。

企业需求

1 传统搜索功能无法满足高度个性化的需求。沃尔玛致力于提供更个性化的商品推荐和更直观的购物体验。

2 沃尔玛希望通过引入生成式 AI 技术，帮助员工完成各种任务，包括文件总结和内容创建。

使用 Azure OpenAI 服务

沃尔玛选择了微软 Azure OpenAI 服务，以访问世界领先的 AI 模型，同时获得微软智能云 Azure 企业级功能（包括安全性、合规性和区域可用性）的支持。结合沃尔玛专有数据和技术、大型语言模型以及沃尔玛建立的零售专用模型，沃尔玛在 iOS、Android 和自有网站上建立了一个全新的生成式 AI 支持的搜索功能。这项新功能专门用于理解顾客查询的上下文并生成个性化回复。很快，顾客将利用该功能获得更具互动性和对话性的体验，获取特定问题的答案，并收到个性化的产品建议。借助这种前沿的生成式 AI 技术，用户可以从“滚动搜索”转变为“目标搜索”，从而使数字购物体验更加流畅和直观。

成功亮点

- 生成式 AI 技术改进搜索功能，提供更个性化和直观的购物体验；
- 5 万名非门店员工通过“我的助手”应用显著提升工作效率；
- 沃尔玛与微软战略合作加速了数字化转型，带来业务创新。

04 游戏行业 完美世界

完美世界游戏是中国最早自主研发 3D 游戏引擎的游戏企业。作为全球化的游戏开发商、发行商、运营商，并在端游、手游、主机游戏、VR 游戏以及云游戏等多个领域进行布局，旗下产品出口 100 多个国家和地区，其出品的《诛仙》《天龙八部》《幻塔》等游戏，搭载中国文化，为全球玩家带来了优质的游戏体验。为了在 MMO 游戏研发中实现效能突破，完美世界游戏于 2022 年开始基于 Azure OpenAI 及其他 Azure AI 服务探索 GenAI 时代的游戏创新。

企业需求

1 提升产品品质，打造精品游戏。大模型技术在图像与文本生产领域的广泛应用为游戏行业带来了新的机遇，然而生成式 AI 的调试精度难以满足游戏文本与美术资产的创作。

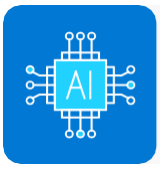
2 升级游戏玩家的 AI 交互体验。AI 对话缺少与生成文本对应的声画表现，难以直接用在 MMO 项目中，玩家的游戏体验需要进一步升级。

3 革新技术中台对工作室的赋能。完美世界游戏内部早已成立 AI 中心，但如何将 GenAI 的技术能力落地为各个工作室开箱即用的实用功能也是一项巨大挑战。

使用 Azure OpenAI 服务

完美世界游戏在生产管线、核心业务和开发工具上实现了全面优化。策划团队引入 GPT 3.5 和 GPT-4 等 Azure OpenAI 模型，为 AI 文本创意、剧情拓展和定制世界观环节提供支持；结合 GPT 4-8K 模型和 GPT 3.5 微调模型以及基于 Azure TTS 文本转语音服务，自研自动生成游戏过场动画的工具 D+，生成角色情绪、动作、分镜、文本、表情、口型、语音等；在美术环节，基于 GPT4 + DALL-E3 模型和微软 3D 优化工具 Simplygon 实现智能化建模、生成分镜预览；此外，完美世界将 Azure OpenAI 模型强大的归纳、理解及总结能力应用在量化运营场景与安全维护场景；并基于 Azure OpenAI 服务的 GPT-4、Codex 模型和 GitHub Copilot 工具，实现了智能代码生成和智能测试。

成功亮点



目前生产管线优化已经完成了大于 38% 的 AI 化，核心业务优化的 AI 渗透率占比超过 60%，技术中心提供给各个工作室的技术开发工具也完成了超过 47% 的 AI 能力接入；

AI 技术在完美世界游戏在研 MMO 项目组中达到了全流程应用，产品开发效率提升 35%，美术产出效率提升 23%；

完全基于微软 Azure OpenAI 服务及 GPT-4 模型能力，上线 AI 原生剧本杀游戏。

05 专业服务行业 KPMG

毕马威会计师事务所在 143 个国家和地区开展业务，共有超过 270,000 名合伙人和员工（至 2023 财年），为全球企业提供审计、税务和咨询等专业服务。毕马威与微软扩展了全球合作伙伴关系，重塑多个关键业务领域的专业服务，包括劳动力现代化、安全可靠的开发以及为广泛的客户、行业和社会提供 AI 解决方案。

企业需求

跟上快速发展的 AI 技术的步伐，为全球客户提供基于 AI 的智能解决方案，提供更优质的专业服务和决策洞察。 1

增强员工体验，使毕马威全球 20 多万员工提升分析效率，从而释放创造力，将更多时间用于为客户提供战略建议上。 2

毕马威多年积累的数据需不断扩充和微调，同时确保满足治理、风险和监管要求。 3

使用 Azure OpenAI 服务

毕马威将微软 Azure 的 AI 创新与自身的税务、审计和咨询专业知识结合起来，利用多学科模型的强大能力，为员工提供支持，为客户输出洞察。通过将数据分析、AI 和 Azure 认知服务融入审计流程，员工能够实时审计，并更密切地关注高风险领域的审计风险和挑战；将 Azure OpenAI 服务和 Microsoft Fabric 集成到 KPMG Digital Gateway，客户能够轻松访问全套毕马威税务和法律知识；基于 Azure OpenAI 服务开发的 AI 解决方案，帮助员工分析 ESG 数据，建立数据模式并起草 ESG 税务透明度报告；基于微软 Azure 开发的 AI 知识平台，加速为客户创建专门的解决方案。

成功亮点

由 AI 生成的“虚拟助手”，创建全新的客户服务模式，帮助税务专业人员提高效率；

基于 Azure OpenAI 服务开发的 AI 解决方案，为毕马威提供了新的创收机会；

加快为客户创建专门的解决方案，助力提升客户的竞争优势和盈利能力，同时获取企业隐私、道德和安全保障。

06 零售行业 CarMax

CarMax 是美国最大的二手车零售商，其服务覆盖全国，且在任意时间段都有超过 45,000 辆汽车可供顾客使用。帮助顾客进行购前研究，为消费者提供诚信和透明的购车体验是 CarMax 的首要任务。基于微软 Azure OpenAI 服务，CarMax 通过真实用户评论，打破信息壁垒，为众多客户提供更好的购车体验。

企业需求

1 提升客户体验，打破二手车市场信息不对称的挑战，让选购过程更清晰、简单

2 随着汽车库存不断增加，CarMax 需进一步优化 OpenAI 模型部署，以应对规模扩大带来的性能压力

使用 Azure OpenAI 服务

CarMax 使用 Azure OpenAI 服务的 GPT-3 模型，在短短几个月内生成了大量原创内容，从 5000 个汽车页面的真实用户评论中提取出最易于阅读和理解的亮点摘要，并根据 CarMax 的需求微调，快速为买家提供有价值、可用的内容。同时，CarMax 将其 OpenAI 负载迁移到微软 Azure，采用 Azure OpenAI 服务，获得微软智能云 Azure 内置的企业级功能，如安全性、合规性和区域可用性。

成功亮点

简化了其汽车搜索页面文本摘要的创建流程，且经编辑审查，生成的摘要批准率达到 80%；

改善顾客的购车体验，帮助 CarMax 网站在搜索引擎的排名不断攀升；

进一步优化 CarMax 的 AI 部署，获取更强性能和企业级安全性。


Microsoft Azure