使用亚马逊 Comprehend Medical 以及医疗 LLMs 保健和生命科学

AWS 规范性指导

AWS 规范性指导: 使用亚马逊 Comprehend Medical 以及医疗 LLMs 保健和生命科学

Table of Contents

使用亚马逊 Comprehend Medical 以及医疗 LLMs 保健和生命科学

亚马逊 Web Services（???贡献者）

2025 年 12 月（文档历史记录）

概述

医疗数据量的不断增加以及对高效、准确处理的需求推动了自然语言处理 (NLP) 以及人工智能和机器学习 (AI/ML) 技术的采用。预训练的分类器模型和大型语言模型 (LLMs) 已成为各种医学 NLP 任务的强大工具，包括临床问答、报告摘要和洞察生成。但是，由于医学术语、特定领域的知识和监管要求的复杂性，医疗保健和生命科学领域面临着独特的挑战。要有效地使用预训练的分类器或在这个领域 LLMs 中使用预训练的分类器，需要一种精心设计的方法，将这些模型的优势与特定领域的资源和技术相结合。

医疗保健和生命科学领域的行业实践传统上依赖基于规则的系统、手动编码和专家评审流程。这些系统和流程既耗时又容易出错。人工智能和自然语言处理技术（例如A mazon Comprehend Medical和 Amazon Bedrock 中的基础模型）的集成为处理医疗数据提供了高效且可扩展的解决方案，同时提高了准确性和一致性。

本指南探讨了 Amazon Comprehend Medical LLMs 以及医疗行业中智能自动化的用途。它概述了简化医疗编码、患者信息提取和记录摘要流程的最佳实践、挑战和实用方法。通过使用 Amazon Comprehend Medical LLMs 和，医疗机构可以将运营效率提升到新的水平，降低成本，并有可能改善患者护理。

该指南详细介绍了医疗保健领域的独特注意事项，例如理解医学术语、使用特定领域以及解决 LLMs系统的局限性。 AI/ML 它为医疗保健 IT 经理、架构师和技术主管提供了一条全面的决策路径，帮助他们评估组织就绪情况，评估实施选项， AWS 服务 并使用适当的工具成功实现自动化。

通过遵循本指南中概述的指导方针和最佳实践，医疗保健组织可以利用 AI/ML 技术的力量，同时应对医疗领域的复杂性。这种方法支持遵守道德和监管准则，并促进在医疗保健中负责任地使用人工智能系统。它旨在生成准确和私密的见解。

目标受众

本指南适用于希望为医疗数据分析和自动化实施人工智能驱动的自然语言处理解决方案的技术利益相关者、架构师、技术主管和决策者。

目标

医疗保健和生命科学组织可以通过使用 Amazon Comprehend Medical 和. LLMs 这些结果通常包括提高运营效率、降低成本和改善患者护理。本节概述了关键业务目标以及实施本指南中概述的战略和最佳实践的相关好处。

以下是各组织通过实施本指南中的指导方针和最佳做法可以实现的一些目标：

- 缩短开发时间 — 本指南的最终目标是减少开发时间和成本，减少技术债务，并减少POC可能导致的项目失败。通过了解关键 AI/ML 服务（例如Amazon Comprehend Medical），以及将法学硕士学位用于医疗保健任务的优势和局限性，企业可以缩短上市时间，提高实现业务目标的速度。
- 提取信息以自动执行医疗编码任务 — 患者就诊后，编码专家和提供者可以从医学文本中提取见解，例如主观、客观、评估和计划 (SOAP) 笔记。这可以减少手动记录的工作量，并帮助提供者专注于患者的需求。通过将 Amazon Comprehend Medical 的实体识别功能 LLMs与之相结合，组织可以从患者记录、临床记录和其他医疗保健数据源中提取相关的医疗信息。这可以最大限度地减少人为错误并促进一致的做法。
- 汇总患者记录和临床文档 — 自动汇总患者病史、治疗计划和医疗结果可以为医疗保健提供者节省宝贵的时间。 LLMs 可以帮助生成全面和结构化的临床文档。你可以通过 Amazon Comprehend Medical 获取更多背景信息，使用医学领域 LLM，或者使用医疗数据微调 LLM。这些方法可以帮助提供准确的摘要，并确保文档符合合规要求和标准。
- 支持临床决策和患者护理 — 通过在 Amazon Comprehend Med ical 中使用本体链接并 LLMs使用，提供者可以回答医疗问题或寻求有关患者护理的建议。这使医疗保健专业人员能够做出明智的决定，从而改善患者的预后并降低医疗失误的风险。

用于医疗保健和生命科学的生成式人工智能和自然语言处理方法

自然语言处理 (NLP) 是一种机器学习技术，它使计算机能够解释、操纵和理解人类语言。医疗保健和生命科学组织拥有来自患者记录的大量数据。他们可以使用 NLP 软件自动处理这些数据。例如，他们可以将 NLP 与生成式 AI 相结合，以简化医疗编码、提取患者信息并汇总记录。

根据您要执行的 NLP 任务，不同的架构可能最适合您的用例。本指南介绍了以下针对医疗保健和生命科学应用的生成式 AI 和 NLP 选项： AWS

- 使用亚马逊 Comprehend Medical— 了解如何独立使用 Amazon Comprehend Medical，无需将其与大型语言模型 (LLM) 集成。
- 将 Amazon Comprehend Medical 与大型语言模型相结合— 了解如何在检索增强生成 (RAG) 架构中将 Amazon Comprehend Medical 与法学硕士学位相结合。
- 在医疗保健和生命科学用例中使用大型语言模型— 通过使用经过微调的法学硕士学位或RAG架构，学习如何将法学硕士学位用于医疗保健和生命科学应用。

使用亚马逊 Comprehend Medical

Amazon Comprehend Medical 可以检测并返回非结构化临床文本中的有用信息，例如医生记录、出院摘要、测试结果和病例记录。 AWS 服务 它使用自然语言处理 (NLP) 模型来检测实体。实体是指医疗信息的文本引用，例如医疗状况、药物或受保护的健康信息 (PHI)。

> **Important**
>
> Amazon Comprehend Medical 并不代替专业的医学意见、诊断或治疗。Amazon Comprehend Medical 提供置信度分数，该分数表明对检测到的实体的准确性的信心。为您的使用案例确定正确的置信度阈值，并在需要高准确度的情况下使用高置信度阈值。对于某些使用案例，结果应由经过适当培训的人力审核人员进行审核和验证。例如，只有经过训练有素的医学专家审查准确性和进行合理的医学判断后，Amazon Comprehend Medical 才能在患者护理场景中使用。

你可以通过、AWS CLI() 或通过 AWS 管理控制台，访问 Amazon Comprehend Med AWS Command Line Interface ical。 AWS SDKs AWS SDKs 它们适用于各种编程语言和平台，例如

Java、Python、Ruby、.NET、iOS 和 Android。您可以使用通过您的客户端应用程序 SDKs 以编程方式访问 Amazon Comprehend Medical。

本节回顾了亚马逊 Comprehend Medical 的主要功能。它还讨论了与大型语言模型 (LLM) 相比，使用此服务的优势。

亚马逊 Comprehend Medical 能力

Amazon Comprehend Medical APIs 提供近乎实时的批量推理。它们 APIs 可以通过使用医疗实体识别和识别实体关系来摄取医学文本，为医学自然语言处理任务提供结果。您可以对单个文件进行分析，也可以对存储在亚马逊简单存储服务 (Amazon S3) 存储桶中的多个文件进行批量分析。Amazon Comprehend Medical 提供以下文本分析 API 操作，用于同步实体检测：

- 检测实体-检测一般医学类别，例如解剖学、医疗状况、PHI 类别、程序和时间表达式。
- 检测 PHI — 检测特定实体，例如年龄、日期、姓名和类似的个人信息。

Amazon Comprehend Medical 还包括多个 API 操作，您可以使用这些操作对临床文件进行批量文本分析。要详细了解如何使用这些 API 操作，请参阅批量文本分析 APIs。

使用 Amazon Comprehend Medical 检测临床文本中的实体，并将这些实体与标准化医学本体中的概念关联起来，包括、ICD-10-CM 和 SNOMED CT RxNorm 知识库。您可以对单个文件进行分析，也可以对存储在 Amazon S3 存储桶中的大型文档或多个文件进行批量分析。Amazon Comprehend Medical 提供以下本体链接 API 操作：

- Infer ICD10 CM — Infer ICD10 CM 操作可检测潜在的疾病，并将其与 2019 年版《国际疾病分类》第 10 修订版《临床修改》（ICD-10-CM）中的代码关联起来。对于检测到的每种潜在医学状况，Amazon Comprehend Medical 都会列出匹配的 ICD-10-CM 代码和描述。结果中列出的医学状况包括置信度分数，该分数表明 Amazon Comprehend Medical 对实体与结果中的概念匹配准确性的信心程度。
- InferRxNorm— 该InferRxNorm手术将患者记录中列出的药物识别为实体。它将实体链接到国家医学图书馆 RxNorm 数据库中的概念标识符 (rxCUI)。每个 RxCUI 都是独一无二的，具有不同的强度和剂量形式。结果中列出的药物包括置信度分数，这表明Amazon Comprehend Medical对与知识库中的概念相匹配的实体的准确性有信心。 RxNormAmazon Comprehend Medical 根据置信CUIs 度分数按降序列出了可能与其检测到的每种药物相匹配的顶级处方药物。
- InfersnomedCT — InfersnomedCT 操作将可能的医学概念识别为实体，并将其与2021-03年版本的《系统化医学命名法，临床术语》（SNOMEDCT）中的代码联系起来。SNOMED CT 提供了全面的医学概念词汇，包括医学状况和解剖学，以及医学检查、治疗和手术。对于每个匹配的概念

ID，Amazon Comprehend Medical 会返回排名前五的医学概念，每个概念都有置信度分数和情境信息，例如相关特征和属性。然后，当与 SNOMED CT 多层次结构一起使用时，SNOMED CT 概念 IDs 可用于构建患者临床数据，用于医学编码、报告或临床分析。

有关更多信息，请参阅 Amazon Compre hend Medical APIs 文档中的文本分析 APIs和本体论链接。

亚马逊 Comprehend Medical 的用例

作为一项独立服务，Amazon Comprehend Medical 可能会解决贵组织的使用案例。Amazon Comprehend Medical 可以执行以下任务：

- 帮助在患者记录中进行医疗编码
- 检测受保护的健康信息 (PHI) 数据
- 验证药物，包括剂量、频率和形式等属性

对于大多数医疗机构来说，Amazon Comprehend Medical 的结果是可以理解的。但是，如果您有以下限制，则可能需要考虑其他选择：

- 不同的实体定义 — 例如，您对药物实体的定义可能会有所不同。FREQUENCY对于频率，Amazon Comprehend Medical 会根据需要进行预测，但您的组织可能会使用 pro re nata（PRN）一词。
- 大量结果 — 例如，患者笔记通常包含多个症状和关键字，这些症状和关键字映射到多个 ICD-10-CM 代码。但是，有几个关键词不适用于诊断。在这种情况下，提供者必须评估许多 ICD-10-CM 实体及其置信度分数，这需要手动处理时间。
- 自定义实体或 NLP 任务 — 例如，提供者可能想要提取 PRN 证据，例如按需服用。由于无法通过亚马逊 Comprehend Medical 购买，因此需要使用不同的型号。 AI/ML 如果 NLP 任务不在实体识别范围内，例如摘要、问答和情感分析，则需要不同的 AI/ML 解决方案。

将 Amazon Comprehend Medical 与大型语言模型相结合

NEJM AI在2024年进行的一项研究表明，使用带有零镜头提示的法学硕士学位来完成医疗编码任务通常会导致表现不佳。将 Amazon Comprehend Medical 与法学硕士学位一起使用可以帮助缓解这些绩效问题。Amazon Comprehend Medical 的结果为执行自然语言处理任务的法学硕士提供了有用的背景。例如，提供从 Amazon Comprehend Medical 到大型语言模型的背景信息可以帮助您：

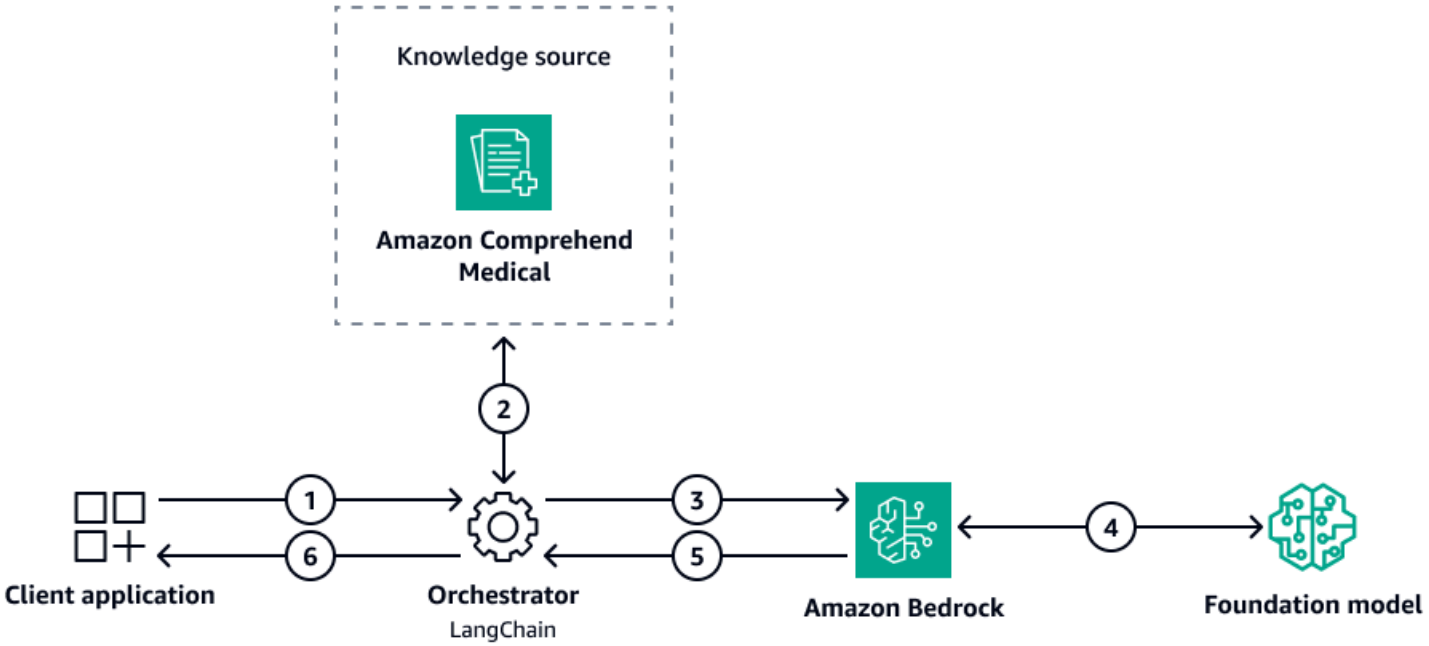
- 使用亚马逊 Comprehend Medical 的初始结果作为法学硕士学位的背景，提高实体选择的准确性
- 实现自定义实体识别、摘要、问答和其他用例

本节介绍如何使用检索增强生成 (RAG) 方法将 Amazon Comprehend Medical 与法学硕士学位相结合。检索增强生成 (RAG) 是一种生成式人工智能技术，其中法学硕士在生成响应之前引用其训练数据源之外的权威数据源。有关更多信息，请参阅什么是 RAG。

为了说明这种方法，本节使用了与 ICD-10-CM 相关的医疗（诊断）编码示例。它包括示例架构和快速工程模板，可帮助您加快创新。它还包括在 RAG 工作流程中使用 Amazon Comprehend Medical 的最佳实践。

基于 RAG 的架构，采用亚马逊 Comprehend Medical

下图说明了一种从患者记录中识别 ICD-10-CM 诊断代码的 RAG 方法。它使用亚马逊 Comprehend Medical 作为知识来源。在 RAG 方法中，检索方法通常从包含适用知识的矢量数据库中检索信息。该架构不是向量数据库，而是使用 Amazon Comprehend Medical 来执行检索任务。协调员将患者病历信息发送到 Amazon Comprehend Medical 并检索 ICD-10-CM 代码信息。协调器通过 Amazon Bedrock 将此上下文发送到下游基础模型 (LLM)。LLM 使用 ICD-10-CM 代码信息生成响应，并将该响应发送回客户端应用程序。



该图显示了以下 RAG 工作流程：

1. 客户端应用程序将患者笔记作为查询发送给协调员。这些患者笔记的一个例子可能是：“患者是X医生的71岁女性患者。该患者昨晚被送往急诊室，有大约7天至8天的腹痛史，这种病史一直持续存在。她没有明显的发烧或发冷，也没有黄疸病史。患者否认最近有任何明显的体重减轻。”
2. 协调器使用 Amazon Comprehend Medical 检索与查询中的医疗信息相关的 ICD-10-CM 代码。它使用 Infer ICD10 CM API 从患者笔记中提取和推断 ICD-10-CM 代码。

3. 协调器构造一个提示，其中包括提示模板、原始查询和从 Amazon Comprehend Medical 检索到的 ICD-10-CM 代码。它将此增强的背景信息发送给 Amazon Bedrock。
4. Amazon Bedrock 处理输入并使用基础模型生成响应，其中包含 ICD-10-CM 代码及其相应查询证据。生成的响应包括已识别的 ICD-10-CM 代码和患者记录中支持每种代码的证据。以下为示例响应：

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock 将生成的响应发送给协调器。
6. 协调器将响应发送回客户端应用程序，用户可以在其中查看响应。

在 RAG 工作流程中使用 Amazon Comprehend Medical 的用例

Amazon Comprehend Medical 可以执行特定的 NLP 任务。有关更多信息，请参阅 Amazon Comprehend Medical 的用例。

您可能需要将 Amazon Comprehend Medical 集成到 RAG 工作流程中，以用于高级用例，例如：

- 通过将提取的医疗实体与患者记录中的情境信息相结合，生成详细的临床摘要
- 使用带有本体关联信息的提取实体进行代码分配，从而自动对复杂案例进行医学编码
- 使用提取的医疗实体，自动从非结构化文本创建结构化临床笔记
- 根据提取的药物名称和属性分析药物副作用
- 开发智能临床支持系统，将提取的医疗信息与 up-to-date研究和指南相结合

在 RAG 工作流程中使用 Amazon Comprehend Medical 的最佳实践

在将 Amazon Comprehend Medical 成绩整合为法学硕士学位的提示时，必须遵循最佳实践。这可以提高性能和准确性。以下是主要建议：

- 了解 Amazon Comprehend Medical 置信度分数 — Amazon Comprehend Medical 为每个检测到的实体和本体关联提供置信度分数。理解这些分数的含义并为您的特定用例设定适当的阈值至关重要。置信度分数有助于过滤掉低可信度的实体，从而减少噪音并提高法学硕士输入的质量。
- 在@@ 即时工程中使用信心分数 — 在起草法学硕士学位的提示时，可以考虑将 Amazon Comprehend Medical 置信度分数作为其他背景。这有助于法学硕士根据实体的置信度对实体进行优先级排序或权衡，从而有可能提高产出的质量。
- 使用实况数据评估 Amazon Comprehend Medical 的结果 — 实况数据是已知真实的信息。它可以用来验证 AI/ML 应用程序是否产生了准确的结果。在将 Amazon Comprehend Medical 结果整合到您的法学硕士工作流程之前，请根据具有代表性的数据样本评估该服务的性能。将结果与事实注释进行比较，以确定潜在的差异或需要改进的领域。此评估可帮助您了解亚马逊 Comprehend Medical 在您的用例中的优势和局限性。
- 策略性地选择相关信息 — Amazon Comprehend Medical 可以提供大量信息，但并非所有信息都可能与您的任务相关。仔细选择与您的用例最相关的实体、属性和元数据。向 LLM 提供过多的无关信息可能会带来噪音，并可能降低性能。
- 调整实体定义 — 确保 Amazon Comprehend Medical 使用的实体和属性的定义与您的解释一致。如果存在差异，可以考虑向法学硕士提供更多背景信息或澄清，以弥合亚马逊 Comprehend Medical 的产出与您的要求之间的差距。如果 Amazon Comprehend Medical 实体未达到您的期望，您可以通过在提示中包含其他说明（和可能的示例）来实现自定义实体检测。
- 提供特定领域的知识 — 虽然 Amazon Comprehend Medical 提供了宝贵的医疗信息，但它可能无法捕捉到您特定领域的所有细微差别。考虑在 Amazon Comprehend Medical 结果中补充其他特定领域的知识来源，例如本体论、术语或专家精心策划的数据集。这为法学硕士提供了更全面的背景信息。
- 遵守道德和监管准则 — 在处理医疗数据时，必须遵守道德原则和监管准则，例如与数据隐私、安全和在医疗保健中负责任地使用人工智能系统相关的道德原则和监管准则。确保您的实施符合相关法律和行业最佳实践。

通过遵循这些最佳实践， AI/ML 从业者可以有效地利用Amazon Comprehend Medical和的优势。LLMs对于医疗 NLP 任务，这些最佳实践有助于降低潜在风险并提高绩效。

为亚马逊 Comprehend Medical 环境进行及时的工程设计

P@@ rompt Engineering 是设计和完善提示的过程，以指导生成式 AI 解决方案生成所需的输出。您可以选择最合适的格式、短语、单词和符号，以引导 AI 与用户进行更有意义的互动。

根据您执行的 API 操作，Amazon Comprehend Medical 会返回检测到的实体、本体代码和描述以及置信度分数。当您的解决方案调用目标 LLM 时，这些结果将成为提示中的上下文。您必须设计提示以在提示模板中呈现上下文。

> **Note**
>
> 本节中的示例提示遵循了 Anthropic 的指导。如果您使用的是其他法学硕士提供商，请遵循该提供商的建议。

通常，您需要在提示中同时插入原始医学文本和亚马逊 Comprehend Medical 结果。以下是常见的提示结构：

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>

<prompt_instructions>
prompt instructions
</prompt_instructions>
```

本节提供了将 Amazon Comprehend Medical 结果作为以下常见医疗 NLP 任务的即时上下文的策略：

- 筛选亚马逊 Comprehend Medical 结果
- 使用亚马逊 Comprehend Medical 扩展医疗 NLP 任务 Amazon Medical
- 使用亚马逊 Comprehend Medical 安装护栏 Amazon Comprehend Medical

筛选亚马逊 Comprehend Medical 结果

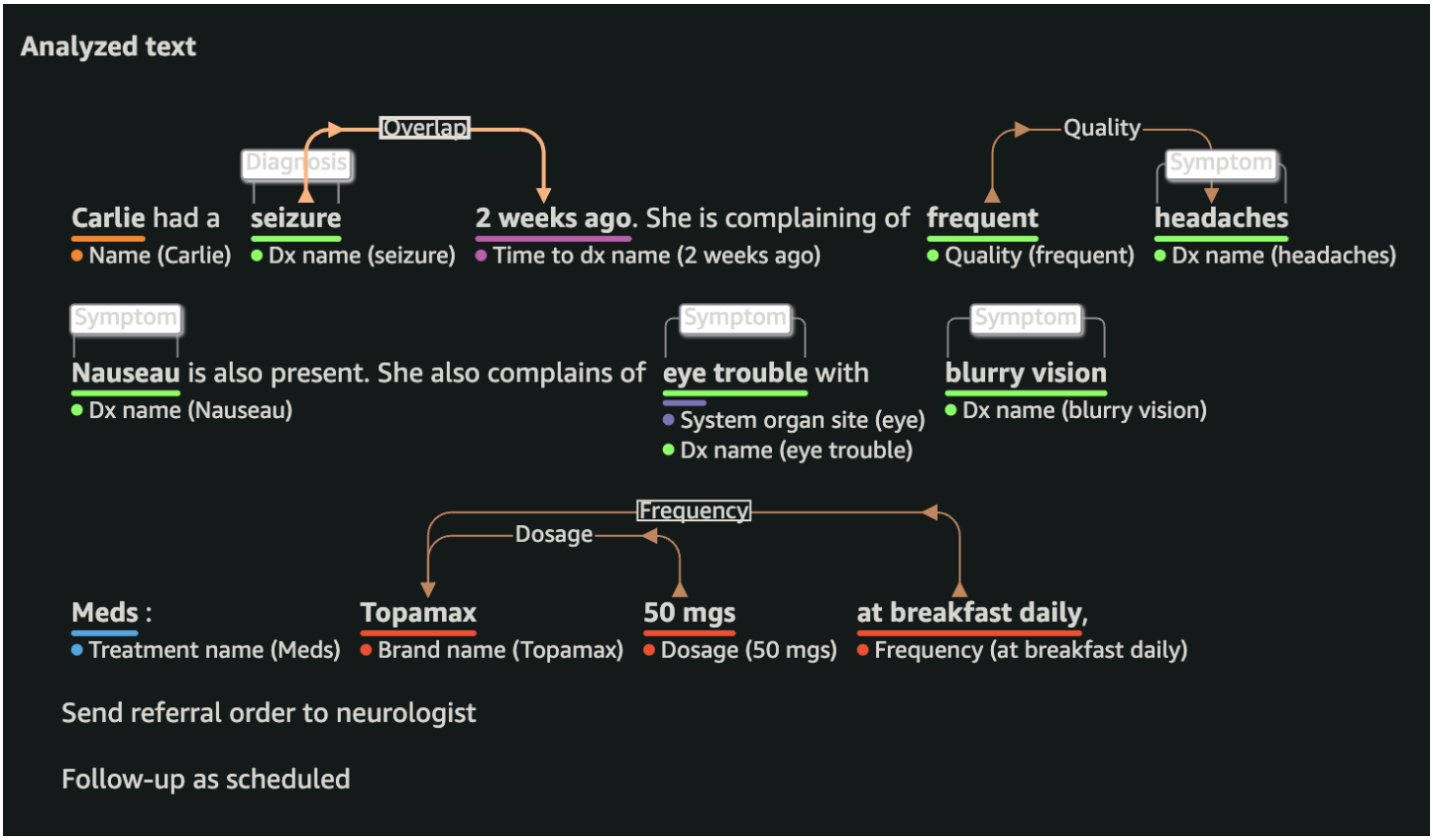
Amazon Comprehend Medical 通常会提供大量信息。您可能需要减少医疗专业人员必须审查的结果数量。在这种情况下，您可以使用 LLM 来筛选这些结果。Amazon Comprehend Medical 实体包括置信度分数，您可以在设计提示时将其用作筛选机制。

以下是患者笔记示例：

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
```

```
Follow-up as scheduled
```

在这份患者记录中，Amazon Comprehend Medical 检测到以下实体。



这些实体链接到以下用于癫痫发作和头痛的 ICD-10-CM 代码。

类别	ICD-10-CM 代码	ICD-10-CM 描述	置信度分数
发作	R56.9	未指明的抽搐	0.8348
发作	G40.909	癫痫，未指明，不可治愈，无癫痫持续状态	0.5424
发作	R56.00	单纯性发热性惊厥	0.4937
发作	G40.09	其他癫痫发作	0.4397
发作	G40.409	其他全身性癫痫和癫痫综合征，非难治性，无癫痫持续状态	0.4138

头痛	R51	头痛	0.4067
头痛	R51.9	头痛，未指明	0.3844
头痛	G44.52	新的每日持续性头痛 (NDPH)	0.3005
头痛	G44	其他头痛综合症	0.2670
头痛	G44.8	其他指定的头痛综合征	0.2542

您可以将 ICD-10-CM 代码传递到提示符中以提高 LLM 精度。为了减少噪音，您可以使用亚马逊 Comprehend Medical 结果中包含的置信度分数筛选 ICD-10-CM 代码。以下是仅包含置信度分数高于 0.4 的 ICD-10-CM 代码的提示示例：

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>

<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
```

```
      <score>0.43966275453567505</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.6460834741592407</score>
    </code>
  </entity>
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
      <score>0.6780954599380493</score>
    </code>
    <code>
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
```

```
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

使用亚马逊 Comprehend Medical 扩展医疗 NLP 任务 Amazon Medical

在处理医学文本时，Amazon Comprehend Medical 的背景信息可以帮助法学硕士选择更好的代币。在此示例中，您希望将诊断症状与药物相匹配。您还想查找与医学检查相关的文本，例如与血液检查相关的术语。你可以使用 Amazon Comprehend Medical 来检测实体和药物名称。在这种情况下，你可以使用 DetectEntitiesV2 和 Amazon Comprehen InferRxNorm APIs d Medical。

以下是患者笔记示例：

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
```

为了重点介绍诊断代码，提示中仅使用与 wit MEDICAL_CONDITION h typ DX_NAME e 相关的实体。由于不相关，其他元数据被排除在外。对于药物实体，包括药物名称和提取的属性。由于无关紧要，不包括来自亚马逊 Comprehend Medical 的其他药物实体元数据。以下是使用筛选后的 Amazon Comprehend Medical 结果的提示示例。提示侧重于具有该DX_NAME类型的MEDICAL_CONDITION实体。此提示旨在更精确地将诊断代码与药物联系起来，并更精确地提取医疗命令测试：

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
```

```
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
```

```
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

使用亚马逊 Comprehend Medical 安装护栏 Amazon Comprehend Medical

在使用生成的响应之前，你可以使用 LLM 和 Amazon Comprehend Medical 来创建护栏。您可以对未经修改或经过后处理的医学文本运行此工作流程。用例包括处理受保护的健康信息 (PHI)、检测幻觉或实施用于发布结果的自定义策略。例如，您可以使用 Amazon Comprehend Medical 中的上下文来识别 PHI 数据，然后使用 LLM 删除该 PHI 数据。

以下是患者记录中包括 PHI 的信息示例：

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

以下是包含亚马逊 Comprehend Medical 结果作为背景的示例提示：

```
<original_text>
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
```

```
</instructions>
```

在医疗保健和生命科学用例中使用大型语言模型

这描述了如何在医疗保健和生命科学应用中使用大型语言模型 (LLMs)。某些用例需要使用大型语言模型来实现生成式 AI 功能。即使是大多数也存在优点和局限性 state-of-the-art LLMs，本节中的建议旨在帮助您实现目标结果。

考虑到领域知识和可用训练数据等因素，您可以使用决策路径来确定适合您的用例的法学硕士解决方案。此外，本节还讨论了流行的预训练医学 LLMs 和最佳实践，供其选择和使用。它还讨论了复杂、高性能的解决方案和更简单、更低成本的方法之间的权衡。

法学硕士学位的用例

Amazon Comprehend Medical 可以执行特定的 NLP 任务。有关更多信息，请参阅 亚马逊 Comprehend Medical 的用例。

高级医疗保健和生命科学用例可能需要法学硕士的逻辑和生成人工智能能力，例如：

- 对自定义医疗实体或文本类别进行分类
- 回答临床问题
- 汇总医疗报告
- 从医疗信息中生成和检测见解

定制方法

了解如何实施 LLMs 至关重要。 LLMs 通常使用数十亿个参数进行训练，包括来自多个领域的训练数据。该培训使法学硕士能够解决大多数一般性任务。但是，当需要特定领域的知识时，往往会出现挑战。医疗保健和生命科学领域知识的例子包括诊所代码、医学术语和生成准确答案所需的健康信息。因此，在这些用例中按原样使用 LLM（在不补充领域知识的情况下进行零点提示）可能会导致结果不准确。您可以使用几种流行的方法来克服这一挑战：即时工程、检索增强生成 (RAG) 和微调。

提示工程

P@@ rompt Engin eering 是指导生成式 AI 解决方案通过调整 LLM 的输入来创建所需输出的过程。通过制定具有相关背景的精确提示，可以引导模型完成需要推理的专业医疗保健任务。有效的即时工程可以显著提高医疗保健用例的模型性能，而无需修改模型。有关提示工程的更多信息，请参阅使用

Amazon Bedrock 实现高级提示工程（AWS 博客文章）。Few-shot chain-of-thought 提示和提示是你可以在提示工程中使用的技术。

少样本提示

Few-shot prompting 是一种技术，在让 LLM 执行类似的任务之前，你可以向 LLM 提供一些所需输入输出的示例。在医疗保健领域，这种方法对于专业任务（例如医疗实体识别或临床记录摘要）特别有价值。通过在提示中包含 3-5 个高质量的示例，可以显著提高模型对医学术语和特定领域模式的理解。有关少镜头提示的示例，请参阅 LLMs Amazon Bedrock 中的 Few-shot 提示工程和微调（博客文章）。AWS

例如，当您从临床记录中提取药物剂量时，可以提供不同符号样式的示例，以帮助模型识别医疗保健专业人员记录处方的方式的差异。当使用标准化文档格式或数据中存在一致模式时，这种方法特别有效。

Chain-of-thought 提示

Chain-of-thought (CoT) 提示引导法学硕士完成 step-by-step推理过程。这使得它对于复杂的医疗决策支持和诊断推理任务非常有价值。通过明确指示模型在分析临床情景时 “分步思考”，您可以提高其遵循医学推理方案并减少诊断错误的能力。

当临床推理需要多个逻辑步骤（例如鉴别诊断或治疗计划）时，这种技术非常出色。但是，在处理模型训练数据之外的高度专业化的医学知识或需要绝对精确的重症监护决策时，这种方法存在局限性。

在这些情况下，将 CoT 与另一种方法结合使用可以产生更好的结果。一种选择是将 CoT 与自一致性提示相结合。有关更多信息，请参阅 Amazon Bedrock 上使用自一致性提示增强生成语言模型的性能（AWS 博客文章）。另一种选择是将推理框架（例如 ReAct 提示）与RAG结合使用。有关更多信息，请参阅使用 RAG 和 ReAct提示开发基于 AI 聊天的高级生成式 AI 助手（AWS 规范性指导）。

检索增强生成

检索增强生成 (RAG) 是一种生成式人工智能技术，其中法学硕士在生成响应之前引用其训练数据源之外的权威数据源。RAG 系统可以从知识来源检索医学本体信息（例如国际疾病分类、国家药物档案和医学主题标题）。这为法学硕士提供了额外的背景信息，以支持医学自然语言处理任务。

如将 Amazon Comprehend Medical 与大型语言模型相结合本节所述，您可以使用 RAG 方法从 Amazon Comprehend Medical 检索上下文。其他常用知识来源包括存储在数据库服务中的医疗领域数据，例如亚马逊 OpenSearch 服务、Amazon Kendra或Amazon Aurora。从这些知识源中提取信息可能会影响检索性能，尤其是使用矢量数据库的语义查询。

存储和检索特定领域知识的另一种选择是在您的 RAG 工作流程中使用 A mazon Q Bus iness。Amazon Q Business 可以为内部文档存储库或面向公众的网站（例如 ICD-10 数据的

cms.gov）编制索引。然后，Amazon Q Business 可以从这些来源提取相关信息，然后再将您的查询传递给 LLM。

有多种方法可以构建自定义 RAG 工作流程。例如，有许多方法可以从知识源检索数据。为简单起见，我们建议使用常用的检索方法，即使用矢量数据库（例如 Amazon S OpenSearch ervice）将知识存储为嵌入内容。这要求您使用嵌入模型（例如句子转换器）来为查询和存储在矢量数据库中的知识生成嵌入式。

有关完全托管和自定义 RAG 方法的更多信息，请参阅中的检索增强生成选项和架构。 AWS

微调

微调现有模型包括获取 LLM，例如 Amazon Titan、Mistral 或 Llama 模型，然后根据您的自定义数据调整模型。有多种微调技术，其中大多数只涉及修改几个参数，而不是修改模型中的所有参数。这称为参数效率微调 (PEFT)。欲了解更多信息，请参阅上的 Hugging Face PEF T GitHub。

以下是两个常见的用例，你可以选择微调法学硕士以完成医疗 NLP 任务：

- 生成任务 — 基于解码器的模型执行生成式 AI 任务。 AI/ML 从业者使用实况数据来微调现有的法学硕士。例如，你可以使用公共医疗问答数据集 MedQuAD 来训练 LLM。当你调用对经过微调的 LLM 的查询时，你不需要 RAG 方法来为 LLM 提供额外的上下文。
- 嵌入 — 基于编码器的模型通过将文本转换为数值向量来生成嵌入。这些基于编码器的模型通常称为嵌入模型。句子转换器模型是一种针对句子进行了优化的特定类型的嵌入模型。目标是根据输入文本生成嵌入内容。然后，嵌入用于语义分析或检索任务。要微调嵌入模型，您必须拥有可以用作训练数据的医学知识语料库，例如文档。这是通过基于相似度或情感的文本对来微调句子转换器模型来实现的。有关更多信息，请参阅 Hugg ing Face 上的 “使用句子转换器 v3 训练和微调嵌入模型”。

您可以使用 Amazon G SageMaker round Trut h 来构建带有标签的高质量训练数据集。您可以使用从 Ground Truth 输出的已标注数据集来训练自己的模型。您也可以将输出用作 Amazon A SageMaker I 模型的训练数据集。有关命名实体识别、单标签文本分类和多标签文本分类的更多信息，请参阅 Amazon A SageMaker I 文档中的使用 G round Truth 进行文本标注。

有关微调的更多信息，请参阅本指南微调医疗保健中的大型语言模型中的。

选择法学硕士

Amazon Bedrock 是评估高 LLMs绩效的推荐起点。有关更多信息，请参阅 Amazon Bedrock 中支持的基础模型。您可以在 Amazon Bedrock 中使用模型评估任务来比较多个输出的输出，然后选择最适合

您的用例的模型。有关更多信息，请参阅 Amazon Bedrock 文档中的使用 Amazon Bedrock 评估选择性能最佳的模型。

LLMs 有些人接受的医学领域数据培训有限。如果你的用例需要微调 Amazon Bedrock 不支持的 LLM 或 LLM，可以考虑使用 Amazon AI。 SageMaker 在 SageMaker AI 中，你可以使用经过微调的 LLM，也可以选择经过医学领域数据训练的自定义 LLM。

下表列出了接受 LLMs 过医学领域数据训练的热门课程。

LLM	任务	知识	架构
BioBert	信息检索、文本分类和命名实体识别	摘要 PubMed、来自的全文文章和一般领域知识 PubMedCentral	编码器
临床伯特	信息检索、文本分类和命名实体识别	大型多中心数据集以及来自电子健康记录 (EHR) 系统的超过 300 万份患者记录	编码器
clinicalGPT	摘要、问答和文本生成	广泛而多样的医疗数据集，包括病历、特定领域的知识和多轮对话咨询	解码器
GatorTron-OG	摘要、问答、文本生成和信息检索	临床笔记和生物医学文献	编码器
Med-bert	信息检索、文本分类和命名实体识别	包含医学文本、临床笔记、研究论文和医疗保健相关文档的大型数据集	编码器
Med-Palm	用于医疗目的的问答	医学和生物医学文本数据集	解码器
MedalPaca	问答和医学对话任务	各种医学文本，包括医学抽认卡、wiki 和对话数据集等资源	解码器

BiomedBert	信息检索、文本分类和命名实体识别	独家摘要 PubMed 和全文文章来自 PubMedCentral	编码器
BioMedLM	摘要、问答和文本生成	来自 PubMed 知识来源的生物医学文献	解码器

以下是使用预先训练的医疗 LLMs服务的最佳实践：

- 了解训练数据及其与您的医学 NLP 任务的相关性。
- 确定法学硕士架构及其用途。编码器适用于嵌入和 NLP 任务。解码器用于生成任务。
- 评估托管预先培训的医学法学硕士学位所需的基础设施、性能和成本要求。
- 如果需要微调，请确保训练数据的地面真相或知识准确。确保屏蔽或删除任何个人身份信息 (PII) 或受保护的健康信息 (PHI)。

在知识或预期用例方面，现实世界 LLMs 中的医学 NLP 任务可能与预先训练的任务有所不同。如果特定领域的 LLM 不符合您的评估基准，则可以使用自己的数据集对 LLM 进行微调，也可以训练新的基础模型。培训新的基础模型是一项雄心勃勃且往往代价高昂的任务。对于大多数用例，我们建议对现有模型进行微调。

当你使用或微调经过预先训练的医学法学硕士时，解决基础设施、安全和护栏问题很重要。

Infrastructure

与使用 Amazon Bedrock 进行按需或批量推理相比，托管预先训练的医学 LLM（通常来自 Hugging Face）需要大量资源。要托管预先训练的医学 LLM，通常使用在亚马逊弹性计算云 (Amazon EC2) 实例上运行的 Amazon A SageMaker I 映像，其中包含一个或多个 GPUs实例，例如用于加速计算的 ml.g5 实例或用于加速计算的 ml.inf2 实例。 AWS Inferentia这是因为 LLMs 消耗大量内存和磁盘空间。

安全和护栏

根据您的业务合规要求，可以考虑使用亚马逊 Comprehend 和 Amazon Comprehend Medical 来屏蔽或删除培训数据中的个人身份信息 (PII) 和受保护的健康信息 (PHI)。这有助于防止 LLM 在生成响应时使用机密数据。

我们建议您考虑和评估生成式 AI 应用中的偏见、公平性和幻觉。无论您使用的是先前存在的 LLM 还是微调的 LLM，都要实施护栏以防止有害反应。护栏是您可以根据生成式人工智能应用程序要求和负责任的人工智能政策进行定制的保障措施。例如，你可以使用 Amazon Bedrock Guardrails。

微调医疗保健中的大型语言模型

本节中描述的微调方法支持遵守道德和监管准则，并促进在医疗保健中负责任地使用人工智能系统。它旨在生成准确、私密的见解。生成式人工智能正在彻底改变医疗保健的交付，但在准确性至关重要且合规性不可谈判的临床环境中， off-the-shelf模型往往不够完善。使用特定领域数据微调基础模型弥合了这一差距。它可以帮助您创建讲医学语言的人工智能系统，同时遵守严格的监管标准。但是，成功的微调之路需要仔细研究医疗保健面临的独特挑战：保护敏感数据，用可衡量的结果证明人工智能投资的合理性，以及在快速变化的医疗环境中保持临床相关性。

当轻量化方法达到极限时，微调就成为一项战略投资。预计精度、延迟或运营效率的提高将抵消所需的巨额计算和工程成本。重要的是要记住，基础模型的进展速度很快，因此经过微调的模型的优势可能要持续到下一个主要模型的发布。

本节将讨论以下两个来自 AWS 医疗保健客户的高影响力用例作为讨论的基础：

- 临床决策支持系统 — 通过了解复杂患者病史和不断演变的指导方针的模型，提高诊断准确性。微调可以帮助模型深入了解复杂的患者病史并整合专门的指导方针。这有可能减少模型预测错误。但是，您需要权衡这些收益与大型敏感数据集的培训成本和高风险临床应用所需的基础架构。提高的准确性和情境感知能力是否证明投资是合理的，尤其是在频繁发布新车型的情况下？
- 医疗文件分析 — 自动处理临床记录、影像报告和保险文件，同时保持《健康保险便携性和责任法案》(HIPAA) 的合规性。在这里，微调可以使模型更有效地处理独特的格式、专门的缩写和监管要求。回报通常体现在减少人工审查时间和提高合规性上。尽管如此，还是必须评估这些改进是否足够大，足以保证微调资源。确定及时的工程和工作流程协调能否满足您的需求。

这些真实场景说明了从最初的实验到模型部署的微调过程，同时满足了医疗保健在每个阶段的独特需求。

估算成本和投资回报

以下是微调法学硕士学位时必须考虑的成本因素：

- 模型大小 — 较大的模型微调成本更高
- 数据集大小-计算成本和时间会随着数据集的大小而增加，以便进行微调
- 微调策略 — 与完整参数更新相比，具有参数效率的方法可以降低成本

在计算投资回报率 (ROI) 时，请考虑所选指标（例如准确性）的改善乘以请求量（模型的使用频率）以及新版本超越模型之前的预期持续时间。

另外，请考虑基础法学硕士的寿命。每隔 6-12 个月就会出现新的基础模型。如果您的罕见病检测器需要8个月的时间进行微调和验证，那么在较新的型号缩小差距之前，您可能只能获得4个月的卓越性能。

通过计算用例的成本、投资回报率和潜在寿命，您可以做出以数据为导向的决策。例如，如果微调您的临床决策支持模型可以显著减少每年成千上万例病例的诊断错误，那么投资可能会很快得到回报。相反，如果仅凭即时工程就能使您的文档分析工作流程接近目标精度，那么明智的做法是推迟微调，直到下一代模型问世。

微调不 one-size-fits-all是。如果您决定进行微调，则正确的方法取决于您的用例、数据和资源。

选择微调策略

在确定微调是适合您的医疗保健用例的正确方法之后，下一步就是选择最合适的微调策略。有几种方法可供选择。对于医疗保健应用，每种方法都有明显的优势和权衡取舍。这些方法之间的选择取决于您的具体目标、可用数据和资源限制。

培训目标

领域自适应预训练 (DAPT) 是一种无人监督的方法，它涉及在大量特定领域的、未加标签的文本（例如数百万份医疗文档）上对模型进行预训练。这种方法非常适合提高模型理解放射科医生、神经科医生和其他专业提供者使用的医学专业缩写和术语的能力。但是，DAPT 需要大量数据，并且不涉及特定的任务输出。

监督微调 (SFT) 使用结构化输入输出示例，教导模型遵循明确的指令。这种方法非常适合医学文档分析工作流程，例如文档摘要或临床编码。指令调谐是 SFT 的一种常见形式，其中模型根据示例进行训练，这些示例包括与所需输出配对的显式指令。这增强了模型理解和遵循不同用户提示的能力。该技术在医疗保健环境中特别有价值，因为它使用特定的临床示例来训练模型。主要缺点是它需要精心标记的示例。此外，经过微调的模型可能会难以应对没有示例的边缘情况。有关使用 Amazon Jumpstar SageMaker t 进行微调的说明，请参阅使用 Amazon Jumpstart 对 FLAN T5 XL 进行指令微调（博 SageMaker 客文章）。AWS

通过@@ 人工反馈进行强化学习 (RLHF) 可根据专家反馈和偏好优化模型行为。使用根据人类偏好和方法（例如近端策略优化 (PPO) 或直接偏好优化 (D PO)）进行训练的奖励模型来优化模型，同时防止破坏性更新。RLHF 非常适合使产出与临床指南保持一致，并确保建议保持在批准的方案之内。这种方法需要大量的临床医生时间来获得反馈，并且涉及复杂的培训流程。但是，RLHF在医疗保健领域特别

有价值，因为它可以帮助医学专家塑造人工智能系统的沟通方式并提出建议。例如，临床医生可以提供反馈，以确保模型保持适当的床边方式，知道何时表达不确定性，并保持在临床指南范围内。PPO 等技术可根据专家反馈迭代优化模型行为，同时限制参数更新以保留核心医学知识。这使模型能够以对患者友好的语言传达复杂的诊断，同时仍能标记严重病情以立即就医。这对于医疗保健至关重要，因为准确性和沟通方式都很重要。有关 RLHF 的更多信息，请参阅利用人类或人工智能反馈的强化学习对大型语言模型进行微调（AWS 博客文章）。

实现方法

完整参数更新涉及在训练期间更新所有模型参数。这种方法最适合需要深度整合患者病史、实验室结果和不断演变的指南的临床决策支持系统。缺点包括计算成本高，如果您的数据集不大且不多样化，则存在过度拟合的风险。

参数高效微调 (PEFT) 方法仅更新参数的子集，以防止过度拟合或语言能力的灾难性损失。类型包括低等级自适应 (LoRa)、适配器和前缀调整。PEFT 方法的计算成本更低，训练速度更快，并且非常适合诸如根据新医院的方案或术语调整临床决策支持模型之类的实验。主要限制是与完整参数更新相比，性能可能会降低。

有关微调方法的更多信息，请参阅 Amazon A SageMaker I 上的高级微调方法（AWS 博客文章）。

构建微调数据集

微调数据集的质量和多样性对于模型性能、安全性和偏见预防至关重要。以下是构建此数据集时需要考虑的三个关键领域：

- 基于微调方法的音量
- 来自领域专家的数据注释
- 数据集的多样性

如下表所示，微调的数据集大小要求因所执行的微调类型而异。

微调策略	数据集大小
适应领域的预训练	100,000 多个域名文本
有监督的微调	超过 10,000 个带标签的货币对
通过人工反馈进行强化学习	1,000 多个专家偏好组合

您可以使用 AWS GlueAmazon EMR 和 A mazon SageMaker Data Wrangler 自动执行数据提取和转换过程，以整理您拥有的数据集。如果您无法整理足够大的数据集，则可以直接发现数据集并将其下载到您的直 AWS 账户 通AWS Data Exchange中。在使用任何第三方数据集之前，请咨询您的法律顾问。

具有领域知识的专家注释者，例如医生、生物学家和化学家，应参与数据整理过程，以便将医学和生物数据的细微差别纳入模型输出中。Amazon G SageMaker round Trut h 提供了一个低代码用户界面，供专家对数据集进行注释。

代表人口的数据集对于医疗保健和生命科学微调用例以防止偏见并反映现实世界的结果至关重要。AWS Glue 交互式会话或 Amazon SageMaker 笔记本实例提供了一种使用兼容 Jupyter 的笔记本来迭代探索数据集和微调转换的强大方法。交互式会话使您能够在本地环境中使用多种常用的集成开发环境 (IDEs)。或者，您可以通过使用 AWS Glue 我们的 Amazon SageMaker Studio 笔记本电脑 AWS 管理控制台。

微调模型

AWS 提供诸如 Amazon A SageMaker I 和 Amaz on Bedrock 之类的服务，这些服务对于成功进行微调至关重要。

SageMaker AI 是一项完全托管的机器学习服务，可帮助开发人员和数据科学家快速构建、训练和部署机器学习模型。用于微调的 SageMaker AI 的三个有用功能包括：

- SageMaker训练 — 一项完全托管的机器学习功能，可帮助您高效地大规模训练各种模型
- SageMaker JumpStart— 一种建立在 SageMaker 训练作业之上的功能，可为机器学习任务提供预训练模型、内置算法和解决方案模板
- SageMaker HyperPod— 专门构建的基础架构解决方案，用于基础模型的分布式训练 LLMs

Amazon Bedrock 是一项完全托管的服务，可通过 API 提供对高性能基础模型的访问，并具有内置的安全、隐私和可扩展功能。该服务提供了微调多个可用的基础模型的功能。有关更多信息，请参阅 Amazon Bedrock 文档中的支持模型和区域以进行微调和继续预训练。

在使用任一服务进行微调过程时，请考虑基本模型、微调策略和基础架构。

基本型号选择

诸如Anthropic Claude、Meta Llama和Amazon Nova之类的封闭源模型在托管合规的情况下提供了强劲的 out-of-the-box性能，但将微调灵活性限制在提供商支持的选项（例如像Amazon Bedrock这样的托管选项）上。 APIs 这限制了可定制性，特别是对于受监管的医疗保健用例。相比之下，诸如 Meta

Llama 之类的开源模型提供了对 Amazon SageMaker AI 服务的完全控制和灵活性，当您需要根据自己的特定数据或工作流程要求自定义、审计或深度调整模型时，它们是理想的选择。

微调策略

简单的指令调整可以通过 Amazon Bedrock 模型定制或亚马逊 SageMaker JumpStart来处理。复杂的 PEFT 方法，例如 LoRa 或适配器，需要在 Amazon Bedrock 中使用 SageMaker训练作业或自定义微调功能。支持超大型模型的分布式训练 SageMaker HyperPod。

基础设施规模和控制

诸如 Amazon Bedrock 之类的完全托管服务可最大限度地减少基础设施管理，非常适合优先考虑易用性和合规性的组织。半托管选项（例如）可提供一定的灵活性 SageMaker JumpStart，同时降低复杂性。这些选项适用于快速原型设计或使用预先构建的工作流程。T SageMaker raining 作业附带完全控制和自定义 HyperPod，尽管这些任务需要更多的专业知识，并且在您需要扩展大型数据集或需要自定义管道时最好。

监控经过微调的模型

在医疗保健和生命科学领域，监控法学硕士微调需要跟踪多个关键绩效指标。准确性提供了基准测量，但这必须与精度和召回率相平衡，尤其是在错误分类会带来严重后果的应用中。F1-score 有助于解决医疗数据集中可能常见的类别失衡问题。有关更多信息，请参阅本指南中的评估 LLMs 医疗保健和生命科学应用。

校准指标可帮助您确保模型的置信水平与现实世界的概率相匹配。公平性指标可以帮助您发现不同患者人口统计的潜在偏见。

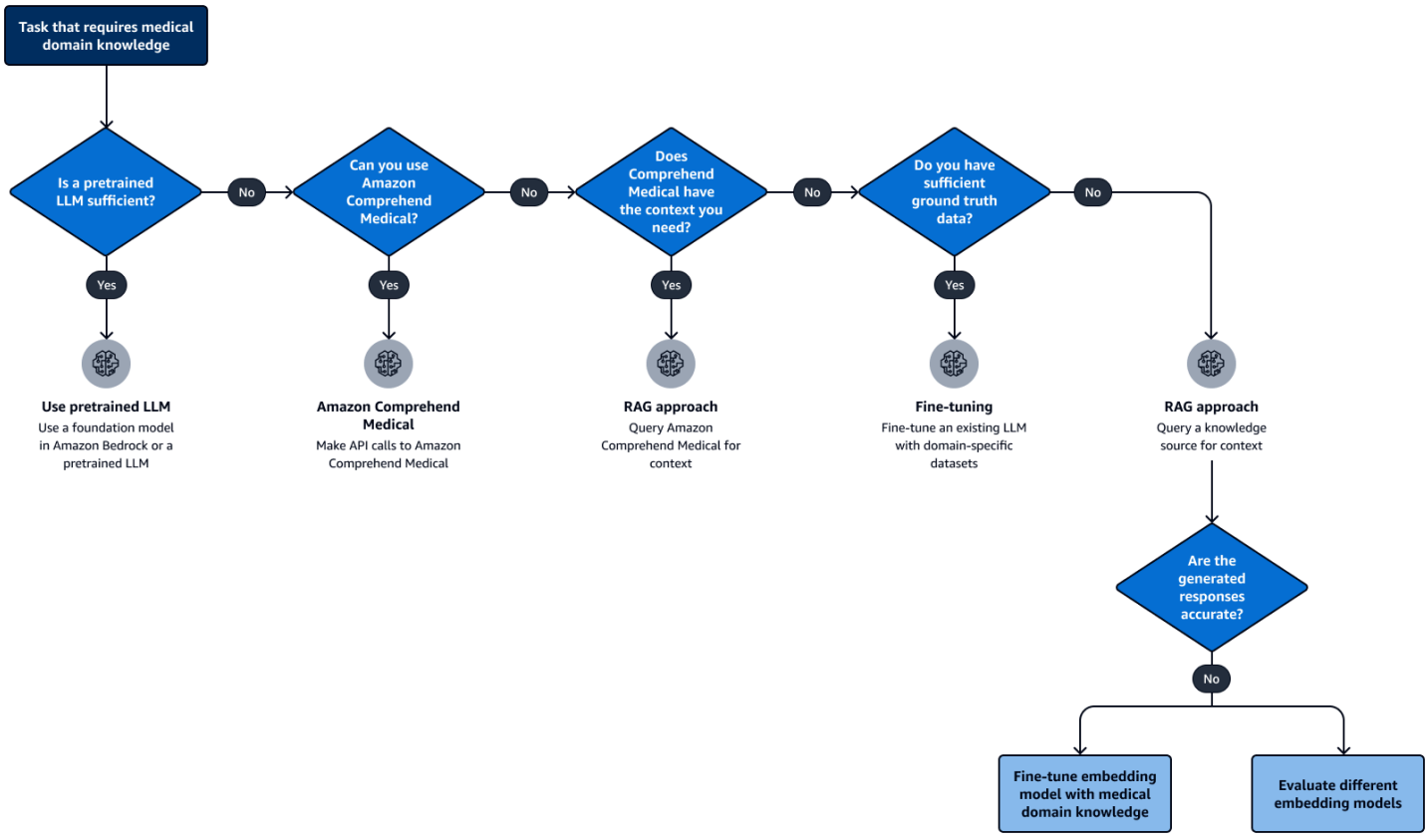
MLflow是一种开源解决方案，可以帮助您跟踪微调实验。 MLflow 在 Amazon A SageMaker I 中原生支持，它可以帮助您直观地比较训练运行中的指标。为了在 Amazon Bedrock 上进行微调任务，指标会流式传输 CloudWatch 到亚马逊，这样你就可以在控制台中可视化这些指标。 CloudWatch

为医疗保健和生命科学选择自然语言处理方法

本用于医疗保健和生命科学的生成式人工智能和自然语言处理方法节介绍解决医疗保健和生命科学应用的自然语言处理 (NLP) 任务的以下方法：

- 使用亚马逊 Comprehend Medical
- 在检索增强生成 (RAG) 工作流程中将 Amazon Comprehend Medical 与法学硕士学位相结合
- 使用经过微调的 LLM
- 使用 RAG 工作流程

通过评估医疗领域任务 LLMs 的已知局限性和您的用例，您可以选择哪种方法最适合您的任务。以下决策树可以帮助您为医疗自然语言处理任务选择法学硕士方法：



下图显示了如下工作流：

1. 对于医疗保健和生命科学用例，请确定自然语言处理任务是否需要特定的领域知识。根据需要，与主题专家协调 (SMEs)。

2. 如果您可以使用通用法学硕士学位或已在医学数据集上训练过的模型，请使用 Amazon Bedrock 中可用的基础模型或经过预训练的 LLM。有关更多信息，请参阅本指南中的选择法学硕士。
3. 如果 Amazon Comprehend Medical 的实体检测和本体关联功能可以解决您的用例，请使用 Amazon Comprehend Medical。 APIs有关更多信息，请参阅本指南中的使用亚马逊 Comprehend Medical。
4. 有时，Amazon Comprehend Medical 有所需的上下文，但不支持你的用例。例如，您可能需要不同的实体定义、接收大量结果、需要自定义实体或需要自定义 NLP 任务。如果是这样的话，请使用 RAG 方法向 Amazon Comprehend Medical 查询背景信息。有关更多信息，请参阅本指南中的将 Amazon Comprehend Medical 与大型语言模型相结合。
5. 如果您有足够数量的地面实况数据，请对现有的 LLM 进行微调。有关更多信息，请参阅本指南中的定制方法。
6. 如果其他方法不能满足你的 NLP 医疗任务目标，请实施 RAG 解决方案。有关更多信息，请参阅本指南中的定制方法。
7. 实施 RAG 解决方案后，评估生成的响应是否准确。有关更多信息，请参阅本指南中的评估 LLMs 医疗保健和生命科学应用。通常从 Amazon Titan 文本嵌入模型或通用句子转换器模型（例如 All-minilm-L6-v2）开始。但是，由于缺乏领域背景，这些模型可能无法捕捉文本中的医学术语。如有必要，请考虑进行以下调整：
   a. 评估其他嵌入模型
   b. 使用特定领域的数据集微调嵌入模型

业务成熟度注意事项

在为医疗保健和生命科学应用调整法学硕士解决方案时，业务成熟度至关重要。这些组织在实施时面临不同程度的复杂性 LLMs，具体取决于其接受标准。通常，缺乏 AI/ML 资源的组织会投资承包商支持，以构建法学硕士解决方案。在这些情况下，了解以下权衡非常重要：

- 高性能，成本高，维护成本高 — 您可能需要复杂的解决方案，包括微调或定制， LLMs 以满足严格的性能标准。但是，这会带来更高的成本和维护要求。您可能需要雇用专业资源或与承包商合作来维护这些复杂的解决方案。这可能会减慢开发速度。
- 性能良好，成本低，维护成本低 — 或者，您可能会发现诸如Amazon Bedrock或Amazon Comprehend Medical之类的服务提供了可接受的性能。尽管这些 LLMs 或方法可能提供完美的结果，但这些解决方案通常可以提供一致、高质量的结果。这些解决方案降低了成本，减少了维护负担。这可以加速开发。

如果更简单、成本更低的方法可以持续提供符合验收标准的高质量结果，请考虑提高性能是否值得在成本、维护和时间上进行权衡。但是，如果较简单的解决方案远未达到目标性能，并且您的组织缺乏复杂解决方案及其维护需求的投资能力，则可以考虑将 AI/ML 开发推迟到有更多资源或替代解决方案可用之前。

此外，对于任何依赖法学硕士学位的医疗自然语言处理解决方案，我们建议您进行持续的监测和评估。评估用户在一段时间内的反馈，并定期进行评估，以确保解决方案继续满足您的业务目标。

评估 LLMs 医疗保健和生命科学应用

本节全面概述了在医疗保健和生命科学用例中评估大型语言模型 (LLMs) 的要求和注意事项。

重要的是要使用实况数据和中小企业反馈来减轻偏见，并验证法学硕士生成的响应的准确性。本节介绍收集和整理训练和测试数据的最佳实践。它还可以帮助您实施防护措施并衡量数据偏见和公平性。它还讨论了常见的医学自然语言处理 (NLP) 任务，例如文本分类、命名实体识别和文本生成，及其相关的评估指标。

它还提供了在训练实验阶段和后期制作阶段进行法学硕士评估的工作流程。模型监控和 LLM 操作是该评估过程的重要组成部分。

医疗 NLP 任务的训练和测试数据

医疗 NLP 任务通常使用医学语料库（例如 PubMed）或患者信息（例如临床患者就诊记录）来分类、总结和生成见解。医务人员，例如医生、医疗保健管理人员或技术人员，其专业知识和观点各不相同。由于这些医务人员之间的主观性，较小的培训和测试数据集会带来偏见的风险。为了降低这种风险，我们建议采用以下最佳实践：

- 使用预训练的 LLM 解决方案时，请确保您有足够数量的测试数据。测试数据应与实际医疗数据非常相似。根据任务的不同，记录的范围可能从 20 到 100 多条不等。
- 在微调法学硕士学位时，请从各种目标医学领域收集足够数量的带标签（事实真相）SMEs 的记录。一般的起点是至少 100 张高质量的唱片。但是，考虑到任务的复杂性以及您的准确性验收标准，可能需要更多记录。
- 如果您的医疗用例需要，请实施防护措施并衡量数据的偏见和公平性。例如，请确保法学硕士学位防止由于患者的种族特征而导致的误诊。有关更多信息，请参阅本指南的安全和护栏部分。

许多人工智能研发公司，例如Anthropic，已经在其基础模型中实施了护栏以避免毒性。您可以使用毒性检测来检查输入提示和来自的输出响应 LLMs。有关更多信息，请参阅 Amazon Comprehend 文档中的毒性检测和亚马逊 Bedrock 文档中的护栏。

在任何生成式人工智能任务中，都有产生幻觉的风险。您可以通过执行 NLP 任务（例如分类）来降低这种风险。您还可以使用更高级的技术，例如文本相似度量度。 BertScore是一种常用的文本相似度量标准。有关可用于缓解幻觉的技术的更多信息，请参阅大型语言模型中幻觉缓解技术的综合调查。

医疗 NLP 任务的指标

在为训练和测试建立基本真相数据和中小企业提供的标签后，您可以创建可量化的指标。通过定性流程（例如压力测试和审查法学硕士学位）来检查质量有助于快速开发。但是，指标充当量基准，支持 Future LLM的运营，并充当每个生产版本的性能基准。

了解医疗任务至关重要。指标通常映射到以下常规 NLP 任务之一：

- 文本分类 — LLM 根据输入提示和提供的上下文将文本分类为一个或多个预定义类别。一个例子是使用疼痛量表对疼痛类别进行分类。文本分类指标的示例包括：
  - 准确性
  - 精度，也称为宏观精度
  - 召回，也称为宏调用
  - F1 分数，也称为宏观 F1 分数
  - 重击损失
- 命名实体识别 (NER) — 也称为文本提取，命名实体识别是将非结构化文本中提及的命名实体定位和分类为预定义类别的过程。一个例子是从患者记录中提取药物名称。NER 指标的示例包括：
  - 准确性
  - 精度
  - 召回率
  - F1 分数
  - 重击损失
- 生成 — LLM 通过处理提示和提供的上下文来生成新文本。生成包括摘要任务或问答任务。生成指标的示例包括：
  - 以召回为导向的要点评估底层研究 (ROUGE)
  - 显式翻译评估指标 ORdering (METEOR)
  - 双语评估研究中 (BLEU)（用于翻译）
  - 字符串距离，也称为余弦相似度

关于医疗保健和生命科学用例的常见问题

以下是与使用 Amazon Comprehend Medical 或执行医疗 LLMs NLP 任务相关的常见问题。

我该如何在亚马逊 Comprehend Medical 和法学硕士学位之间做出选择？

如果您的任务是在医学文本中检测医疗实体，请查看 Amazon Comprehend Medical 文档，以了解可以提取哪些医疗实体，以及是否有任何本体可以解决您的使用案例。如果没有，请考虑使用法学硕士。有关更多信息，请参阅本指南中的亚马逊 Comprehend Medical 的用例和法学硕士学位的用例。

如何向法学硕士提供亚马逊 Comprehend Medical 成绩？

你可以将 Amazon Comprehend Medical 成绩作为背景纳入法学硕士提示中。这为法学硕士提供了额外的医学知识和术语。提供的上下文可以提高法学硕士在实体识别、总结或问答等任务上的表现。该指南提供了几个示例，说明如何使用 Amazon Comprehend Medical 结果来构造提示。有关更多信息，请参阅本指南中的将 Amazon Comprehend Medical 与大型语言模型相结合。

使用 Amazon Comprehend Medical 时，有哪些最佳实践？ LLMs

我们建议使用 Amazon Comprehend Medical 置信度分数来筛选提示中的实体或确定其优先顺序。评估其在您的特定数据上的表现并验证实体定义是否符合您的要求也很重要。将Amazon Comprehend Medical与特定领域的知识来源相结合，可以进一步提高法学硕士的绩效。有关更多信息，请参阅本指南中的在 RAG 工作流程中使用 Amazon Comprehend Medical 的最佳实践。

我应该使用经过预先培训的医学法学硕士学位还是针对我的医疗保健用例微调普通法学硕士学位？

该决定取决于您的具体要求和高质量训练数据的可用性。经过预先培训的医疗人员 LLMs 可以提供一个良好的起点。但是，您可能仍需要使用特定于域的数据对其进行微调。如果您有足够的标签数据，则微调通用法学硕士学位可能是一个可行的选择。有关更多信息，请参阅本指南中的选择法学硕士和为医疗保健和生命科学选择自然语言处理方法。

如何评估医疗自然语言处理任务 LLMs 的表现？

对于文本分类和命名实体识别任务，我们建议使用定量指标，例如准确性、精确度、召回率和 F1 分数。你可以使用 ROUGE 和 METEOR 来完成文本生成任务。重要的是要由主题专家标记可靠的地面实况数据，并实施一段时间内监测模型性能的流程。有关更多信息，请参阅本指南中的评估 LLMs 医疗保健和生命科学应用。

高复杂性和低复杂性法学硕士解决方案之间有什么权衡取舍？

微调 LLM 或构建自定义 LLM 是非常复杂的解决方案。这些方法可以提高性能，但会带来更高的成本和维护要求。更简单的解决方案，例如使用预训练 LLMs 或 Amazon Comprehend Medical，可能会以更低的成本和更快的开发周期提供可接受的性能。但是，对于某些用例，这些方法可能无法满足严格的精度要求。有关更多信息，请参阅本指南中的业务成熟度注意事项。

后续步骤和资源

本指南可帮助您用于 AWS 服务 在生产环境中为实际应用自动执行医疗 NLP 和生成式 AI 任务。它描述了如何使用 Amazon Comprehend Medical、Amazon Bedrock 支持 LLMs 、预先培训的医疗或 LLMs 微调来实现您的 LLMs医疗保健和生命科学业务目标。本指南描述了以下方法的优点和局限性：

- 独立使用亚马逊 Comprehend Medical
- 向法学硕士提供亚马逊 Comprehend Medical 成绩
- 在检索增强生成 (RAG) 方法中使用预先训练的普通法学硕士或医学法学硕士
- 微调普通法学硕士或医学法学硕士

使用本指南中的决策树和业务成熟度注意事项，根据组织 AI/ML 的成熟度级别在这些方法之间进行选择。尽管 Amazon Comprehend Medical 和 A LLMs mazon Bedrock 提供了强大的功能，但只有当你正确实施和评估它们时，它们才会成功。使用本指南中描述的评估信息和指标来验证您的解决方案的性能。

对于后续步骤，我们建议医疗保健 IT 经理、架构师和技术主管与 AI/ML 从业人员合作，确定他们的 NLP 医疗任务。使用本指南选择开发路径，然后使用相应的 AWS 服务 和功能在上成功实施自动化解决方案 AWS。

AWS 资源

- 亚马逊 Comprehend Medical 文档：
  - 开发人员指南
  - API 引用
- Amazon Bedrock 文档
  - 亚马逊 Bedrock 模型评估
  - 在 Amazon Bedrock 中进行微调
- 在 Amazon A SageMaker I 中微调模型
- 亚马逊 G SageMaker round Truth
- 亚马逊 Comprehend 毒性检测
- AWS 医疗保健能力合作伙伴

其他 资源

- 打开 Medical-LLM 排行榜
- 医疗保健大型语言模型调查：从数据、技术和应用到问责制和道德
- 大型语言模型是糟糕的医疗编码人员——医疗代码查询的基准
- 从初学者到专家：将医学知识建模为一般知识 LLMs

贡献者

编写

- Joe King， AWS 高级数据科学家
- Ankith Ede，解决方案架构师 AWS
- Clement Perrot， AWS 高级生成式人工智能策略师
- Jillian Forde，高级解决方案架构师 AWS
- 拉杰什·西塔拉曼，高级交付顾问 AWS
- 罗斯·克莱托， AWS 首席应用科学家
- Shivesh Ummat，解决方案架构师 AWS

审阅

- Dilshad Raihan Akkam Veettil，高级数据科学家 AWS
- 约瑟夫·科廷汉姆， AWS 深度学习架构师

技术写作

- Lilly AbouHarb， AWS 资深技术撰稿人

文档历史记录

下表介绍了本指南的一些重要更改。如果您希望收到有关未来更新的通知，可以订阅 RSS 源。

变更	说明	日期
新章节	我们在医疗保健部分和提示工程部分添加了微调大型语言模型。	2025年12月5日
初次发布	—	2024 年 12 月 16 日

AWS 规范性指导词汇表

以下是 AWS 规范性指导提供的策略、指南和模式中的常用术语。若要推荐词条，请使用术语表末尾的提供反馈链接。

数字

7 R

将应用程序迁移到云中的 7 种常见迁移策略。这些策略以 Gartner 于 2011 年确定的 5 R 为基础，包括以下内容：

- 重构/重新架构：充分利用云原生功能来提高敏捷性、性能和可扩展性，以迁移应用程序并修改其架构。这通常涉及到移植操作系统和数据库。示例：将本地 Oracle 数据库迁移到 Amazon Aurora PostgreSQL 兼容版。
- 更换平台：将应用程序迁移到云中，并进行一定程度的优化，以利用云功能。示例：将本地 Oracle 数据库迁移到 AWS 云中的 Amazon Relational Database Service（Amazon RDS）for Oracle。
- 重新购买：转换到其他产品，通常是从传统许可转向 SaaS 模式。示例：将客户关系管理（CRM）系统迁移到 Salesforce.com。
- 重新托管（直接迁移）：将应用程序迁移到云中，无需进行任何更改即可利用云功能。示例：将本地 Oracle 数据库迁移到 AWS 云中 EC2 实例上的 Oracle。
- 重新放置（虚拟机监控器级直接迁移）：将基础设施迁移到云中，无需购买新硬件、重写应用程序或修改现有操作。您将服务器从本地平台迁移到同一平台的云服务中。示例：将Microsoft Hyper-V应用程序迁移到 AWS。
- 保留（重访）：将应用程序保留在源环境中。其中可能包括需要进行重大重构的应用程序，并且您希望将工作推迟到以后，以及您希望保留的遗留应用程序，因为迁移它们没有商业上的理由。
- 停用：停用或删除源环境中不再需要的应用程序。

A

ABAC

请参阅基于属性的访问控制。

抽象服务

请参阅托管服务。

ACID

请参阅原子性、一致性、隔离性、持久性。

主动-主动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步（通过使用双向复制工具或双写操作），两个数据库都在迁移期间处理来自连接应用程序的事务。这种方法支持小批量、可控的迁移，而不需要一次性割接。它比主动-被动迁移更灵活，但工作量更大。

主动-被动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步，但在将数据复制到目标数据库时，只有源数据库处理来自连接应用程序的事务。目标数据库在迁移期间不接受任何事务。

聚合函数

一种 SQL 函数，它对一组行进行操作并计算该组的单个返回值。聚合函数的示例包括 SUM 和 MAX。

AI

请参阅人工智能。

AIOps

请参阅人工智能运营。

匿名化

永久删除数据集中个人信息的过程。匿名化可以帮助保护个人隐私。匿名化数据不再被视为个人数据。

反模式

一种用于解决反复出现的问题的常用解决方案，而在这类问题中，此解决方案适得其反、无效或不如替代方案有效。

应用程序控制

一种安全方法，仅允许使用经批准的应用程序，以帮助保护系统免受恶意软件的侵害。

应用程序组合

有关组织使用的每个应用程序的详细信息的集合，包括构建和维护该应用程序的成本及其业务价值。这些信息是产品组合发现和分析过程的关键，有助于识别需要进行迁移、现代化和优化的应用程序并确定其优先级。

人工智能（AI）

计算机科学领域致力于使用计算技术执行通常与人类相关的认知功能，例如学习、解决问题和识别模式。有关更多信息，请参阅什么是人工智能？

人工智能操作 (AIOps)

使用机器学习技术解决运营问题、减少运营事故和人为干预以及提高服务质量的过程。有关如何在 AIOps AWS 迁移策略中使用的更多信息，请参阅操作集成指南。

非对称加密

一种加密算法，使用一对密钥，一个公钥用于加密，一个私钥用于解密。您可以共享公钥，因为它不用于解密，但对私钥的访问应受到严格限制。

原子性、一致性、隔离性、持久性（ACID）

一组软件属性，即使在出现错误、电源故障或其他问题的情况下，也能保证数据库的数据有效性和操作可靠性。

基于属性的访问权限控制（ABAC）

根据用户属性（如部门、工作角色和团队名称）创建精细访问权限的做法。有关更多信息，请参阅 AWS Identity and Access Management (I AM) 文档 AWS中的 AB AC。

权威数据来源

存储主要数据版本的位置，被认为是最可靠的信息源。您可以将数据从权威数据来源复制到其他位置，以便处理或修改数据，例如对数据进行匿名化、编辑或假名化。

可用区

中的一个不同位置 AWS 区域 ，不受其他可用区域故障的影响，并向同一区域中的其他可用区提供低成本、低延迟的网络连接。

AWS 云采用框架 (AWS CAF)

该框架包含指导方针和最佳实践 AWS ，可帮助组织制定高效且有效的计划，以成功迁移到云端。AWS CAF将指导分为六个重点领域，称为视角：业务、人员、治理、平台、安全和运营。业务、人员和治理角度侧重于业务技能和流程；平台、安全和运营角度侧重于技术技能和流程。例如，人

员角度针对的是负责人力资源（HR）、人员配置职能和人员管理的利益相关者。从这个角度来看，AWS CAF 为人员发展、培训和沟通提供了指导，以帮助组织为成功采用云做好准备。有关更多信息，请参阅 AWS CAF 网站和 AWS CAF 白皮书。

AWS 工作负载资格框架 (AWS WQF)

一种评估数据库迁移工作负载、推荐迁移策略和提供工作估算的工具。 AWS WQF 包含在 AWS Schema Conversion Tool (AWS SCT) 中。它用来分析数据库架构和代码对象、应用程序代码、依赖关系和性能特征，并提供评测报告。

B

恶意机器人

一种旨在扰乱或伤害个人或组织的机器人。

BCP

请参阅业务连续性计划。

行为图

一段时间内资源行为和交互的统一交互式视图。您可以使用 Amazon Detective 的行为图来检查失败的登录尝试、可疑的 API 调用和类似的操作。有关更多信息，请参阅 Detective 文档中的行为图中的数据。

大端序系统

一个先存储最高有效字节的系统。另请参阅字节顺序。

二进制分类

一种预测二进制结果（两个可能的类别之一）的过程。例如，您的 ML 模型可能需要预测诸如“该电子邮件是否为垃圾邮件？”或“这个产品是书还是汽车？”之类的问题

bloom 筛选条件

一种概率性、内存高效的数据结构，用于测试元素是否为集合的成员。

蓝/绿部署

一种部署策略，您可以创建两个独立但完全相同的环境。在一个环境中运行当前应用程序版本（蓝色），在另一个环境中运行新应用程序版本（绿色）。此策略可帮助您在影响最小的情况下快速回滚。

自动程序

一种通过互联网运行自动任务并模拟人类活动或交互的软件应用程序。有些机器人是有用或有益的，例如在互联网上索引信息的 Web 爬网程序。还有一些被称为恶意机器人的机器人，其目的是扰乱或伤害个人或组织。

僵尸网络

被恶意软件感染并受单方（称为僵尸网络控制者或僵尸网络操作者）控制的僵尸网络。僵尸网络是最著名的扩展机器人及其影响力的机制。

分支

代码存储库的一个包含区域。在存储库中创建的第一个分支是主分支。您可以从现有分支创建新分支，然后在新分支中开发功能或修复错误。为构建功能而创建的分支通常称为功能分支。当功能可以发布时，将功能分支合并回主分支。有关更多信息，请参阅关于分支（GitHub 文档）。

紧急（break-glass）访问

在特殊情况下，通过批准的流程，用户 AWS 账户 可以快速访问他们通常没有访问权限的内容。有关更多信息，请参阅 AWS Well-Architected Guidance 中的 Implement break-glass procedures 指示器。

棕地策略

您环境中的现有基础设施。在为系统架构采用棕地策略时，您需要围绕当前系统和基础设施的限制来设计架构。如果您正在扩展现有基础设施，则可以将棕地策略和全新策略混合。

缓冲区缓存

存储最常访问的数据的内存区域。

业务能力

企业如何创造价值（例如，销售、客户服务或营销）。微服务架构和开发决策可以由业务能力驱动。有关更多信息，请参阅在 AWS上运行容器化微服务白皮书中的围绕业务能力进行组织部分。

业务连续性计划（BCP）

一项计划，旨在应对大规模迁移等破坏性事件对运营的潜在影响，并使企业能够快速恢复运营。

C

CAF

请参阅 AWS 云采用框架。

金丝雀部署

缓慢而渐进地向最终用户发布版本。当您确信无误后，即可部署新版本，并完全替换当前版本。

CCoE

请参阅云卓越中心。

CDC

请参阅更改数据捕获。

更改数据捕获（CDC）

跟踪数据来源（如数据库表）的更改并记录有关更改的元数据的过程。您可以将 CDC 用于各种目的，例如审计或复制目标系统中的更改以保持同步。

混沌工程

故意引入故障或破坏性事件来测试系统的韧性。您可以使用 AWS Fault Injection Service (AWS FIS) 来执行实验，对您的 AWS 工作负载施加压力并评估其响应。

CI/CD

请参阅持续集成和持续交付。

分类

一种有助于生成预测的分类流程。分类问题的 ML 模型预测离散值。离散值始终彼此不同。例如，一个模型可能需要评估图像中是否有汽车。

客户端加密

在目标 AWS 服务 收到数据之前，对数据进行本地加密。

云卓越中心 (CCoE)

一个多学科团队，负责推动整个组织的云采用工作，包括开发云最佳实践、调动资源、制定迁移时间表、领导组织完成大规模转型。有关更多信息，请参阅 AWS 云 企业战略博客上的 CCoE 帖子。

云计算

通常用于远程数据存储和 IoT 设备管理的云技术。云计算通常连接到边缘计算技术。

云运营模型

在 IT 组织中，一种用于构建、完善和优化一个或多个云环境的运营模型。有关更多信息，请参阅构建您的云运营模型。

云采用阶段

组织迁移到 AWS 云中时通常会经历四个阶段：

- 项目 - 出于概念验证和学习目的，开展一些与云相关的项目
- 基础 — 进行基础投资以扩大云采用率（例如，创建着陆区、定义 CCo E、建立运营模型）
- 迁移 - 迁移单个应用程序
- 重塑 - 优化产品和服务，在云中创新

Stephen Orban在 AWS 云 企业战略博客的博客文章《云优先之旅和采用阶段》中定义了这些阶段。有关它们与 AWS 迁移策略的关系的信息，请参阅迁移准备指南。

CMDB

请参阅配置管理数据库。

代码存储库

通过版本控制过程存储和更新源代码和其他资产（如文档、示例和脚本）的位置。常见的云存储库包括 GitHub 或 Bitbucket Cloud。每个版本的代码都称为一个分支。在微服务结构中，每个存储库都专门用于一个功能。单个 CI/CD 管线可以使用多个存储库。

冷缓存

一种空的、填充不足或包含过时或不相关数据的缓冲区缓存。这会影响性能，因为数据库实例必须从主内存或磁盘读取，这比从缓冲区缓存读取要慢。

冷数据

很少访问的数据，且通常是历史数据。查询此类数据时，通常可以接受慢速查询。将这些数据转移到性能较低且成本更低的存储层或类别可以降低成本。

计算机视觉（CV）

一种 AI 领域，它使用机器学习来分析和提取数字图像和视频等视觉格式中的信息。例如，Amazon SageMaker AI 为 CV 提供了图像处理算法。

配置偏移

对于工作负载而言，一种偏离预期状态的配置更改。这可能会导致工作负载变得不合规，且通常是渐进的，不是故意的。

配置管理数据库（CMDB）

一种存储库，用于存储和管理有关数据库及其 IT 环境的信息，包括硬件和软件组件及其配置。您通常在迁移的产品组合发现和分析阶段使用来自 CMDB 的数据。

合规性包

一系列 AWS Config 规则和补救措施，您可以汇编这些规则和补救措施，以自定义您的合规性和安全性检查。您可以使用 YAML 模板将一致性包作为单个实体部署在 AWS 账户 和区域或整个组织中。有关更多信息，请参阅 AWS Config 文档中的一致性包。

持续集成和持续交付（CI/CD）

自动执行软件发布过程的源代码、构建、测试、暂存和生产阶段的过程。 CI/CD 通常被描述为管道。 CI/CD 可以帮助您实现流程自动化、提高生产力、提高代码质量和更快地交付。有关更多信息，请参阅持续交付的优势。CD 也可以表示持续部署。有关更多信息，请参阅持续交付与持续部署。

CV

请参阅计算机视觉。

D

静态数据

网络中静止的数据，例如存储中的数据。

数据分类

根据网络中数据的关键性和敏感性对其进行识别和分类的过程。它是任何网络安全风险管理策略的关键组成部分，因为它可以帮助您确定对数据的适当保护和保留控制。数据分类是 Well-Architecte AWS d Framework 中安全支柱的一个组成部分。有关详细信息，请参阅数据分类。

数据漂移

生产数据与用来训练机器学习模型的数据之间的有意义差异，或者输入数据随时间推移的有意义变化。数据漂移可能降低机器学习模型预测的整体质量、准确性和公平性。

传输中数据

在网络中主动移动的数据，例如在网络资源之间移动的数据。

数据网格

一种架构框架，可提供分布式、去中心化的数据所有权以及集中式管理和治理。

数据最少化

仅收集并处理绝对必要数据的原则。在中进行数据最小化 AWS 云 可以降低隐私风险、成本和分析碳足迹。

数据边界

AWS 环境中的一组预防性防护措施，可帮助确保只有可信身份才能访问来自预期网络的可信资源。有关更多信息，请参阅在上构建数据边界。 AWS

数据预处理

将原始数据转换为 ML 模型易于解析的格式。预处理数据可能意味着删除某些列或行，并处理缺失、不一致或重复的值。

数据溯源

在数据的整个生命周期跟踪其来源和历史的过程，例如数据如何生成、传输和存储。

数据主体

正在收集和处理其数据的个人。

数据仓库

一种支持商业智能（例如分析）的数据管理系统。数据仓库通常包含大量历史数据，通常用于查询和分析。

数据库定义语言（DDL）

在数据库中创建或修改表和对象结构的语句或命令。

数据库操作语言（DML）

在数据库中修改（插入、更新和删除）信息的语句或命令。

DDL

请参阅数据库定义语言。

深度融合

组合多个深度学习模型进行预测。您可以使用深度融合来获得更准确的预测或估算预测中的不确定性。

深度学习

一个 ML 子字段使用多层人工神经网络来识别输入数据和感兴趣的目标变量之间的映射。

defense-in-depth

一种信息安全方法，经过深思熟虑，在整个计算机网络中分层实施一系列安全机制和控制措施，以保护网络及其中数据的机密性、完整性和可用性。当你采用这种策略时 AWS，你会在 AWS

Organizations 结构的不同层面添加多个控件来帮助保护资源。例如，一种 defense-in-depth方法可以结合多因素身份验证、网络分段和加密。

委派管理员

在中 AWS Organizations，兼容的服务可以注册 AWS 成员帐户来管理组织的帐户并管理该服务的权限。此账户被称为该服务的委托管理员。有关更多信息和兼容服务列表，请参阅 AWS Organizations 文档中使用 AWS Organizations的服务。

部署

使应用程序、新功能或代码修复在目标环境中可用的过程。部署涉及在代码库中实现更改，然后在应用程序的环境中构建和运行该代码库。

开发环境

请参阅环境。

侦测性控制

一种安全控制，在事件发生后进行检测、记录日志和发出提醒。这些控制是第二道防线，提醒您注意绕过现有预防性控制的安全事件。有关更多信息，请参阅在 AWS上实施安全控制中的侦测性控制。

开发价值流映射（DVSM）

用于识别对软件开发生命周期中的速度和质量产生不利影响的限制因素并确定其优先级的流程。DVSM 扩展了最初为精益生产实践设计的价值流映射流程。其重点关注在软件开发过程中创造和转移价值所需的步骤和团队。

数字孪生

真实世界系统的虚拟再现，如建筑物、工厂、工业设备或生产线。数字孪生支持预测性维护、远程监控和生产优化。

维度表

星型架构中的一种较小的表，其中包含事实表中定量数据的数据属性。维度表属性通常是文本字段或行为类似于文本的离散数字。这些属性通常用于查询约束、筛选和结果集标注。

灾难

阻止工作负载或系统在其主要部署位置实现其业务目标的事件。这些事件可能是自然灾害、技术故障或人为操作的结果，例如无意的配置错误或恶意软件攻击。

灾难恢复（DR）

您用来最大程度地减少由灾难造成的停机时间和数据丢失的策略和流程。有关更多信息，请参阅 Well-Architected Fr ame AWS work 中的 “工作负载灾难恢复：云端 AWS 恢复”。

DML

请参阅数据库操作语言。

领域驱动设计

一种开发复杂软件系统的方法，通过将其组件连接到每个组件所服务的不断发展的领域或核心业务目标。Eric Evans 在其著作领域驱动设计：软件核心复杂性应对之道（Boston: Addison-Wesley Professional, 2003）中介绍了这一概念。有关如何将领域驱动设计与 strangler fig 模式结合使用的信息，请参阅使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET（ASMX）Web 服务现代化。

DR

请参阅灾难恢复。

偏差检测

跟踪与基准配置的偏差。例如，您可以使用 AWS CloudFormation 来检测系统资源中的偏差，也可以使用 AWS Control Tower 来检测着陆区中可能影响监管要求合规性的变化。

DVSM

请参阅开发价值流映射。

E

EDA

请参阅探索性数据分析。

EDI

请参阅电子数据交换。

边缘计算

该技术可提高位于 IoT 网络边缘的智能设备的计算能力。与云计算比较时，边缘计算可以减少通信延迟并缩短响应时间。

电子数据交换（EDI）

组织之间业务文件的自动交换。有关更多信息，请参阅什么是电子数据交换。

加密

一种将人类可读的纯文本数据转换为加密文字的计算流程。

加密密钥

由加密算法生成的随机位的加密字符串。密钥的长度可能有所不同，而且每个密钥都设计为不可预测且唯一。

字节顺序

字节在计算机内存中的存储顺序。大端序系统先存储最高有效字节。小端序系统先存储最低有效字节。

端点

请参阅服务端点。

端点服务

一种可以在虚拟私有云（VPC）中托管，与其他用户共享的服务。您可以使用其他 AWS 账户 或 AWS Identity and Access Management (IAM) 委托人创建终端节点服务， AWS PrivateLink 并向其授予权限。这些账户或主体可通过创建接口 VPC 端点来私密地连接到您的端点服务。有关更多信息，请参阅 Amazon Virtual Private Cloud（Amazon VPC）文档中的创建端点服务。

企业资源规划（ERP）

一种自动化和管理企业关键业务流程（例如会计、MES 和项目管理）的系统。

信封加密

用另一个加密密钥对加密密钥进行加密的过程。有关更多信息，请参阅 AWS Key Management Service (AWS KMS) 文档中的信封加密。

环境

正在运行的应用程序的实例。以下是云计算中常见的环境类型：

- 开发环境 — 正在运行的应用程序的实例，只有负责维护应用程序的核心团队才能使用。开发环境用于测试更改，然后再将其提升到上层环境。这类环境有时称为测试环境。
- 下层环境 — 应用程序的所有开发环境，比如用于初始构建和测试的环境。

- 生产环境 — 最终用户可以访问的正在运行的应用程序的实例。在 CI/CD 管道中，生产环境是最后一个部署环境。
- 上层环境 — 除核心开发团队以外的用户可以访问的所有环境。这可能包括生产环境、预生产环境和用户验收测试环境。

epic

在敏捷方法学中，有助于组织工作和确定优先级的功能类别。epics 提供了对需求和实施任务的总体描述。例如， AWS CAF 安全史诗包括身份和访问管理、侦探控制、基础设施安全、数据保护和事件响应。有关 AWS 迁移策略中 epics 的更多信息，请参阅计划实施指南。

ERP

请参阅企业资源规划。

探索性数据分析（EDA）

分析数据集以了解其主要特征的过程。您收集或汇总数据，并进行初步调查，以发现模式、检测异常并检查假定情况。EDA 通过计算汇总统计数据和创建数据可视化得以执行。

F

事实表

星型架构中的中心表。它存储有关业务运营的定量数据。通常，事实表包含两种类型的列：包含度量的列和包含维度表外键的列。

快速失效机制

一种使用频繁且增量式的测试来缩短开发生命周期的理念。这是敏捷方法的关键部分。

故障隔离边界

在中 AWS 云，诸如可用区 AWS 区域、控制平面或数据平面之类的边界，它限制了故障的影响并有助于提高工作负载的弹性。有关更多信息，请参阅 AWS 故障隔离边界。

功能分支

请参阅分支。

特征

您用来进行预测的输入数据。例如，在制造环境中，特征可能是定期从生产线捕获的图像。

特征重要性

特征对于模型预测的重要性。这通常表示为数值分数，可以通过各种技术进行计算，例如 Shapley 加法解释（SHAP）和积分梯度。有关更多信息，请参阅使用机器学习模型的可解释性 AWS。

功能转换

为 ML 流程优化数据，包括使用其他来源丰富数据、扩展值或从单个数据字段中提取多组信息。这使得 ML 模型能从数据中获益。例如，如果您将“2021-05-27 00:15:37”日期分解为“2021”、“五月”、“星期四”和“15”，则可以帮助学习与不同数据成分相关的算法学习精细模式。

少样本提示

在要求 LLM 执行类似任务之前，先向其提供少量示例，以演示任务和预期输出。此技术是上下文内学习的一种应用，其中模型可以从提示中嵌入的示例（样本）中学习。对于需要特定格式、推理或领域知识的任务，少样本提示可能非常有效。另请参阅零样本提示。

FGAC

请参阅精细访问控制。

精细访问控制（FGAC）

使用多个条件允许或拒绝访问请求。

快闪迁移

一种数据库迁移方法，通过更改数据捕获使用连续数据复制，在极短的时间内迁移数据，而非使用分阶段方法。目标是将停机时间降至最低。

FM

请参阅基础模型。

基础模型（FM）

一个大型深度学习神经网络，一直在广义和未标记数据的大量数据集上进行训练。 FMs 能够执行各种各样的一般任务，例如理解语言、生成文本和图像以及用自然语言进行对话。有关更多信息，请参阅什么是基础模型。

G

生成式人工智能

AI 模型的一个子集，这些模型已经过大量数据训练，可以使用简单的文本提示来创建新的内容和构件，例如图像、视频、文本和音频。有关更多信息，请参阅什么是生成式人工智能。

地理阻止

请参阅地理限制。

地理限制（地理阻止）

在 Amazon 中 CloudFront，一种阻止特定国家/地区的用户访问内容分发的选项。您可以使用允许列表或阻止列表来指定已批准和已禁止的国家/地区。有关更多信息，请参阅 CloudFront 文档中的限制内容的地理分布。

GitFlow 工作流程

一种方法，在这种方法中，下层和上层环境在源代码存储库中使用不同的分支。Gitflow 工作流程被认为是传统的工作流程，而基于中继的工作流程则是现代的、首选的方法。

黄金映像

系统或软件的快照，用作部署该系统或软件的新实例的模板。例如，在制造业中，黄金映像可用于在多个设备上预调配软件，并有助于提高设备制造操作的速度、可扩展性和生产效率。

全新策略

在新环境中缺少现有基础设施。在对系统架构采用全新策略时，您可以选择所有新技术，而不受对现有基础设施（也称为棕地）兼容性的限制。如果您正在扩展现有基础设施，则可以将棕地策略和全新策略混合。

防护机制

帮助管理各组织单位的资源、策略和合规性的高级规则 (OUs)。预防性防护机制会执行策略以确保符合合规性标准。它们是使用服务控制策略和 IAM 权限边界实现的。侦测性护栏会检测策略违规和合规性问题，并生成提醒以进行修复。它们通过使用 AWS Config、、Amazon、 AWS Security Hub CSPM GuardDuty AWS Trusted Advisor、Amazon Inspector 和自定义 AWS Lambda 支票来实现。

H

HA

请参阅高可用性。

异构数据库迁移

将源数据库迁移到使用不同数据库引擎的目标数据库（例如，从 Oracle 迁移到 Amazon Aurora）。异构迁移通常是重新架构工作的一部分，而转换架构可能是一项复杂的任务。AWS 提供了 AWS SCT 来帮助实现架构转换。

高可用性（HA）

在遇到挑战或灾难时，工作负载无需干预即可连续运行的能力。HA 系统旨在自动进行故障转移、持续提供良好性能，并以最小的性能影响处理不同负载和故障。

历史数据库现代化

一种用于实现运营技术（OT）系统现代化和升级以更好满足制造业需求的方法。历史数据库是一种用于收集和存储工厂中各种来源数据的数据库。

保留数据

从用于训练机器学习模型的数据集中保留的一部分标注的历史数据。通过将模型预测与保留数据进行比较，您可以使用保留数据来评估模型性能。

同构数据库迁移

将源数据库迁移到共享同一数据库引擎的目标数据库（例如，从 Microsoft SQL Server 迁移到 Amazon RDS for SQL Server）。同构迁移通常是更换主机或更换平台工作的一部分。您可以使用本机数据库实用程序来迁移架构。

热数据

经常访问的数据，例如实时数据或近期的转化数据。这些数据通常需要高性能存储层或存储类别才能提供快速的查询响应。

修补程序

针对生产环境中关键问题的紧急修复。由于其紧迫性，修补程序通常是在典型的 DevOps 发布工作流程之外进行的。

hypercare 周期

割接之后，迁移团队立即管理和监控云中迁移的应用程序以解决任何问题的时间段。通常，这个周期持续 1-4 天。在 hypercare 周期结束时，迁移团队通常会将应用程序的责任移交给云运营团队。

我

IaC

请参阅基础设施即代码。

基于身份的策略

附加到一个或多个 IAM 委托人的策略，用于定义他们在 AWS 云 环境中的权限。

空闲应用程序

90 天内平均 CPU 和内存使用率在 5% 到 20% 之间的应用程序。在迁移项目中，通常会停用这些应用程序或将其保留在本地。

IIoT

请参阅工业物联网。

不可变基础设施

一种模型，可为生产工作负载部署新的基础设施，而不是更新、修补或修改现有基础设施。不可变基础设施本质上比可变基础设施更一致、更可靠、更可预测。有关更多信息，请参阅 AWS Well-Architected Framework 中的使用不可变基础设施进行部署最佳实践。

入站（入口）VPC

在 AWS 多账户架构中，一种接受、检查和路由来自应用程序外部的网络连接的 VPC。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

增量迁移

一种割接策略，在这种策略中，您可以将应用程序分成小部分进行迁移，而不是一次性完整割接。例如，您最初可能只将几个微服务或用户迁移到新系统。在确认一切正常后，您可以逐步迁移其他微服务或用户，直到停用遗留系统。这种策略降低了大规模迁移带来的风险。

工业 4.0

该术语由 Klaus Schwab 在 2016 年提出，指的是通过连接、实时数据、自动化、分析和 AI/ML 的进步来实现制造流程的现代化。

基础设施

应用程序环境中包含的所有资源和资产。

基础设施即代码（IaC）

通过一组配置文件预调配和管理应用程序基础设施的过程。IaC 旨在帮助您集中管理基础设施、实现资源标准化和快速扩展，使新环境具有可重复性、可靠性和一致性。

工业物联网 (IIoT)

在工业领域使用联网的传感器和设备，例如制造业、能源、汽车、医疗保健、生命科学和农业。有关更多信息，请参阅制定工业物联网 (IIoT) 数字化转型战略。

检查 VPC

在 AWS 多账户架构中，一种集中式 VPC，用于管理对 VPCs（相同或不同 AWS 区域）、互联网和本地网络之间的网络流量的检查。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

物联网（IoT）

由带有嵌入式传感器或处理器的连接物理对象组成的网络，这些传感器或处理器通过互联网或本地通信网络与其他设备和系统进行通信。有关更多信息，请参阅什么是 IoT？

可解释性

它是机器学习模型的一种特征，描述了人类可以理解模型的预测如何取决于其输入的程度。有关更多信息，请参阅使用机器学习模型的可解释性 AWS。

物联网

请参阅物联网。

IT 信息库（ITIL）

提供 IT 服务并使这些服务符合业务要求的一套最佳实践。ITIL 是 ITSM 的基础。

IT 服务管理（ITSM）

为组织设计、实施、管理和支持 IT 服务的相关活动。有关将云运营与 ITSM 工具集成的信息，请参阅运营集成指南。

ITIL

请参阅 IT 信息库。

ITSM

请参阅 IT 服务管理。

L

基于标签的访问控制（LBAC）

强制访问控制（MAC）的一种实施方式，其中明确为用户和数据本身分配了安全标签值。用户安全标签和数据安全标签之间的交集决定了用户可以看到哪些行和列。

登录区

landing zone 是一个架构精良的多账户 AWS 环境，具有可扩展性和安全性。这是一个起点，您的组织可以从这里放心地在安全和基础设施环境中快速启动和部署工作负载和应用程序。有关登录区的更多信息，请参阅设置安全且可扩展的多账户 AWS 环境。

大语言模型（LLM）

一种基于大量数据进行预训练的深度学习 AI 模型。LLM 可以执行多项任务，例如回答问题、总结文档、将文本翻译成其他语言以及完成句子。有关更多信息，请参阅什么是 LLMs。

大规模迁移

迁移 300 台或更多服务器。

LBAC

请参阅基于标签的访问控制。

最低权限

授予执行任务所需的最低权限的最佳安全实践。有关更多信息，请参阅 IAM 文档中的应用最低权限许可。

直接迁移

请参阅 7 R。

小端序系统

一个先存储最低有效字节的系统。另请参阅字节顺序。

LLM

请参阅大型语言模型。

下层环境

请参阅环境。

M

机器学习（ML）

一种使用算法和技术进行模式识别和学习的人工智能。ML 对记录的数据（例如物联网（IoT）数据）进行分析和学习，以生成基于模式的统计模型。有关更多信息，请参阅机器学习。

主分支

请参阅分支。

恶意软件

旨在危害计算机安全或隐私的软件。恶意软件可能会破坏计算机系统、泄露敏感信息或获得未经授权的访问权限。恶意软件的示例包括病毒、蠕虫、勒索软件、木马、间谍软件和键盘记录器。

托管式服务

AWS 服务 它 AWS 运行基础设施层、操作系统和平台，您可以访问端点来存储和检索数据。Amazon Simple Storage Service（Amazon S3）和 Amazon DynamoDB 就是托管服务的示例。这些服务也称为抽象服务。

制造执行系统（MES）

一种软件系统，用于跟踪、监控、记录和控制在车间将原材料转化为成品的生产过程。

MAP

请参阅迁移加速计划。

机制

一个完整的过程，您可以在其中创建工具，推动工具的采用，然后检查结果以进行调整。机制是一种在运作过程中自我强化和改善的循环。有关更多信息，请参阅在 Well-Architect AWS ed 框架中构建机制。

成员账户

AWS 账户 除属于组织中的管理账户之外的所有账户 AWS Organizations。一个账户一次只能是一个组织的成员。

MES

请参阅制造执行系统。

消息队列遥测传输（MQTT）

一种基于发布/订阅模式的轻量级 machine-to-machine (M2M) 通信协议，适用于资源受限的物联网设备。

微服务

一种小型的独立服务，通过明确的定义进行通信 APIs ，通常由小型的独立团队拥有。例如，保险系统可能包括映射到业务能力（如销售或营销）或子域（如购买、理赔或分析）的微服务。微服务

的好处包括敏捷、灵活扩展、易于部署、可重复使用的代码和恢复能力。有关更多信息，请参阅使用 AWS 无服务器服务集成微服务。

微服务架构

一种使用独立组件构建应用程序的方法，这些组件将每个应用程序进程作为微服务运行。这些微服务使用轻量级通过定义明确的接口进行通信。 APIs该架构中的每个微服务都可以更新、部署和扩展，以满足对应用程序特定功能的需求。有关更多信息，请参阅在上实现微服务。 AWS

迁移加速计划（MAP）

AWS 该计划提供咨询支持、培训和服务，以帮助组织为迁移到云奠定坚实的运营基础，并帮助抵消迁移的初始成本。MAP 提供了一种以系统的方式执行遗留迁移的迁移方法，以及一套用于自动执行和加速常见迁移场景的工具。

大规模迁移

将大部分应用程序组合分波迁移到云中的过程，在每一波中以更快的速度迁移更多应用程序。本阶段使用从早期阶段获得的最佳实践和经验教训，实施由团队、工具和流程组成的迁移工厂，通过自动化和敏捷交付简化工作负载的迁移。这是 AWS 迁移策略的第三阶段。

迁移工厂

跨职能团队，通过自动化、敏捷的方法简化工作负载迁移。迁移工厂团队通常包括运营、业务分析师和所有者、迁移工程师、开发人员和冲刺 DevOps 领域的专业人员。20% 到 50% 的企业应用程序组合由可通过工厂方法优化的重复模式组成。有关更多信息，请参阅本内容集中有关迁移工厂的讨论和云迁移工厂指南。

迁移元数据

有关完成迁移所需的应用程序和服务器的信息。每种迁移模式都需要一套不同的迁移元数据。迁移元数据的示例包括目标子网、安全组和 AWS 账户。

迁移模式

一种可重复的迁移任务，详细列出了迁移策略、迁移目标以及所使用的迁移应用程序或服务。示例：使用 AWS 应用程序迁移服务重新托管向 Amazon EC2 的迁移。

迁移组合评测（MPA）

一种在线工具，提供了用于验证迁移到 AWS 云的业务案例的信息。MPA 提供了详细的组合评测（服务器规模调整、定价、TCO 比较、迁移成本分析）以及迁移计划（应用程序数据分析和数据收集、应用程序分组、迁移优先级排序和波次规划）。所有 AWS 顾问和 APN 合作伙伴顾问均可免费使用 MPA 工具（需要登录）。

迁移准备情况评测（MRA）

使用 AWS CAF 深入了解组织的云就绪状态、确定优势和劣势以及制定行动计划以缩小已发现差距的过程。有关更多信息，请参阅迁移准备指南。MRA 是 AWS 迁移策略的第一阶段。

迁移策略

将工作负载迁移到 AWS 云的方法。有关更多信息，请参见术语表中的 7 R 词条，以及动员您的组织以加快大规模迁移。

ML

请参阅机器学习。

现代化

将过时的（原有的或单体）应用程序及其基础设施转变为云中敏捷、弹性和高度可用的系统，以降低成本、提高效率和利用创新。有关更多信息，请参阅在 AWS 云中实现应用程序现代化的策略。

现代化准备情况评估

一种评估方式，有助于确定组织应用程序的现代化准备情况；确定收益、风险和依赖关系；确定组织能够在多大程度上支持这些应用程序的未来状态。评估结果是目标架构的蓝图、详细说明现代化进程发展阶段和里程碑的路线图以及解决已发现差距的行动计划。有关更多信息，请参阅在 AWS 云中评估应用程序的现代化准备情况。

单体应用程序（单体式）

作为具有紧密耦合进程的单个服务运行的应用程序。单体应用程序有几个缺点。如果某个应用程序功能的需求激增，则必须扩展整个架构。随着代码库的增长，添加或改进单体应用程序的功能也会变得更加复杂。若要解决这些问题，可以使用微服务架构。有关更多信息，请参阅将单体分解为微服务。

MPA

请参阅迁移组合评测。

MQTT

请参阅消息队列遥测传输。

多分类器

一种帮助为多个类别生成预测（预测两个以上结果之一）的过程。例如，ML 模型可能会询问“这个产品是书、汽车还是手机？”或“此客户最感兴趣什么类别的产品？”

可变基础设施

一种用于更新和修改生产工作负载的现有基础设施的模型。为了提高一致性、可靠性和可预测性，Well-Architect AWS ed Framework 建议使用不可变基础设施作为最佳实践。

O

OAC

请参阅来源访问控制。

OAI

请参阅来源访问身份。

OCM

请参阅组织变革管理。

离线迁移

一种迁移方法，在这种方法中，源工作负载会在迁移过程中停止运行。这种方法会延长停机时间，通常用于小型非关键工作负载。

OI

请参阅运营集成。

OLA

请参阅运营级别协议。

在线迁移

一种迁移方法，在这种方法中，源工作负载无需离线即可复制到目标系统。在迁移过程中，连接工作负载的应用程序可以继续运行。这种方法的停机时间为零或最短，通常用于关键生产工作负载。

OPC-UA

请参阅开放流程通信 – 统一架构。

开放流程通信 – 统一架构（OPC-UA）

一种用于工业自动化的 machine-to-machine（M2M）通信协议。OPC-UA 提供了一个包含数据加密、身份验证和授权方案的互操作性标准。

运营级别协议（OLA）

一项协议，阐明了 IT 职能部门承诺相互交付的内容，以支持服务水平协议（SLA）。

运营准备情况审查（ORR）

一份问题核对清单和关联的最佳实践，可帮助您了解、评估、预防或缩小事件和可能的故障的范围。有关更多信息，请参阅 AWS Well-Architected Framework 中的运营准备情况审查（ORR）。

运营技术（OT）

与物理环境配合使用以控制工业运营、设备和基础设施的硬件和软件系统。在制造业中，OT 和信息技术（IT）系统的集成是工业 4.0 转型的关键重点。

运营整合（OI）

在云中实现运营现代化的过程，包括就绪计划、自动化和集成。有关更多信息，请参阅运营整合指南。

组织跟踪

由 AWS CloudTrail 此创建的跟踪记录组织 AWS 账户 中所有人的所有事件 AWS Organizations。该跟踪是在每个 AWS 账户 中创建的，属于组织的一部分，并跟踪每个账户的活动。有关更多信息，请参阅 CloudTrail文档中的为组织创建跟踪。

组织变革管理（OCM）

一个从人员、文化和领导力角度管理重大、颠覆性业务转型的框架。OCM 通过加快变革采用、解决过渡问题以及推动文化和组织变革，帮助组织为新系统和战略做好准备和过渡。在 AWS 迁移策略中，该框架被称为人员加速，因为云采用项目需要变更的速度。有关更多信息，请参阅 OCM 指南。

来源访问控制（OAC）

在中 CloudFront，一个增强的选项，用于限制访问以保护您的亚马逊简单存储服务 (Amazon S3) 内容。OAC 全部支持所有 S3 存储桶 AWS 区域、使用 AWS KMS (SSE-KMS) 进行服务器端加密，以及对 S3 存储桶的动态PUT和DELETE请求。

来源访问身份（OAI）

在中 CloudFront，一个用于限制访问权限以保护您的 Amazon S3 内容的选项。当您使用 OAI 时，CloudFront 会创建一个 Amazon S3 可以对其进行身份验证的委托人。经过身份验证的委托人只能通过特定 CloudFront 分配访问 S3 存储桶中的内容。另请参阅 OAC，其中提供了更精细和增强的访问控制。

ORR

请参阅运营准备情况审查。

OT

请参阅运营技术。

出站（出口）VPC

在 AWS 多账户架构中，一种处理从应用程序内部启动的网络连接的 VPC。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

P

权限边界

附加到 IAM 主体的 IAM 管理策略，用于设置用户或角色可以拥有的最大权限。有关更多信息，请参阅 IAM 文档中的权限边界。

个人身份信息（PII）

直接查看其他相关数据或与之配对时可用于合理推断个人身份的信息。PII 的示例包括姓名、地址和联系信息。

PII

请参阅个人身份信息。

playbook

一套预定义的步骤，用于捕获与迁移相关的工作，例如在云中交付核心运营功能。playbook 可以采用脚本、自动化运行手册的形式，也可以是操作现代化环境所需的流程或步骤的摘要。

PLC

请参阅可编程逻辑控制器。

PLM

请参阅产品生命周期管理。

policy

一个对象，可以定义权限（请参阅基于身份的策略）、指定访问条件（请参阅基于资源的策略）或定义 AWS Organizations 的组织中所有账户的最大权限（请参阅服务控制策略）。

多语言持久性

根据数据访问模式和其他要求，独立选择微服务的数据存储技术。如果您的微服务采用相同的数据存储技术，它们可能会遇到实现难题或性能不佳。如果微服务使用最适合其需求的数据存储，则可以更轻松地实现微服务，并获得更好的性能和可扩展性。

组合评测

一个发现、分析和确定应用程序组合优先级以规划迁移的过程。有关更多信息，请参阅评估迁移准备情况。

谓词

返回 `true` 或 `false` 的查询条件，通常位于 `WHERE` 子句中。

谓词下推

一种数据库查询优化技术，可在传输之前筛选查询中的数据。这将减少从关系数据库检索和处理的数据量，并提高查询性能。

预防性控制

一种安全控制，旨在防止事件发生。这些控制是第一道防线，帮助防止未经授权的访问或对网络的意外更改。有关更多信息，请参阅在 AWS上实施安全控制中的预防性控制。

主体

中 AWS 可以执行操作和访问资源的实体。此实体通常是 IAM 角色的根用户或用户。 AWS 账户有关更多信息，请参阅 IAM 文档中角色术语和概念中的主体。

隐私设计

一种在整个开发过程中都考虑隐私的系统工程方法。

私有托管区

一个容器，其中包含有关您希望 Amazon Route 53 如何响应针对一个或多个 VPCs域名及其子域名的 DNS 查询的信息。有关更多信息，请参阅 Route 53 文档中的私有托管区的使用。

主动控制

一种安全控制，旨在防止部署不合规资源。这些控制会在资源预置之前对其进行扫描。如果资源与控制不兼容，则不会预置它。有关更多信息，请参阅 AWS Control Tower 文档中的控制参考指南，并参见在上实施安全控制中的主动控制 AWS。

产品生命周期管理（PLM）

对产品在其整个生命周期内的数据和流程的管理，从设计、开发和发布，到增长和成熟，再到衰退和淘汰。

生产环境

请参阅环境。

可编程逻辑控制器（PLC）

在制造业中，一种高度可靠、适应性强的计算机，用于监控机器并实现制造过程自动化。

提示串接

使用一个 LLM 提示的输出作为下一个提示的输入，以生成更好的响应。该技术用于将复杂的任务分解为子任务，或者迭代地完善或扩展初步响应。它有助于提高模型响应的准确性和相关性，并允许获得更精细的个性化结果。

假名化

用占位符值替换数据集中个人标识符的过程。假名化可以帮助保护个人隐私。假名化数据仍被视为个人数据。

publish/subscribe (pub/sub)

一种支持微服务间异步通信的模式，可提高可扩展性和响应能力。例如，在基于微服务的 MES 中，微服务可以将事件消息发布到其他微服务可以订阅的频道。系统可以在不更改发布服务的情况下添加新的微服务。

Q

查询计划

一系列用于访问 SQL 关系数据库系统中的数据的步骤，类似于指令。

查询计划回归

当数据库服务优化程序选择的最佳计划不如数据库环境发生特定变化之前时。这可能是由统计数据、约束、环境设置、查询参数绑定更改和数据库引擎更新造成的。

R

RACI 矩阵

请参阅责任、问责、咨询和知情（RACI）。

RAG

请参阅检索增强生成。

勒索软件

一种恶意软件，旨在阻止对计算机系统或数据的访问，直到付款为止。

RASCI 矩阵

请参阅责任、问责、咨询和知情（RACI）。

RCAC

请参阅行列访问控制。

只读副本

用于只读目的的数据库副本。您可以将查询路由到只读副本，以减轻主数据库的负载。

重新架构

请参阅 7 R。

恢复点目标（RPO）

自上一个数据恢复点以来可接受的最长时间。这决定了从上一个恢复点到服务中断之间可接受的数据丢失情况。

恢复时间目标（RTO）

服务中断和服务恢复之间可接受的最大延迟。

重构

请参阅 7 R。

Region

地理区域内的 AWS 资源集合。每一个 AWS 区域 都相互隔离，彼此独立，以提供容错、稳定性和弹性。有关更多信息，请参阅指定您的账户可以使用的 AWS 区域。

回归

一种预测数值的 ML 技术。例如，要解决“这套房子的售价是多少？”的问题 ML 模型可以使用线性回归模型，根据房屋的已知事实（如建筑面积）来预测房屋的销售价格。

重新托管

请参阅 7 R。

版本

在部署过程中，推动生产环境变更的行为。

重新放置

请参阅 7 R。

更换平台

请参阅 7 R。

重新购买

请参阅 7 R。

韧性

应用程序抵御中断或从中断中恢复的能力。在 AWS 云中规划韧性时，高可用性和灾难恢复是常见的考虑因素。有关更多信息，请参阅 AWS 云 韧性。

基于资源的策略

一种附加到资源的策略，例如 AmazonS3 存储桶、端点或加密密钥。此类策略指定了允许哪些主体访问、支持的操作以及必须满足的任何其他条件。

责任、问责、咨询和知情（RACI）矩阵

定义参与迁移活动和云运营的所有各方的角色和责任的矩阵。矩阵名称源自矩阵中定义的责任类型：负责（R）、问责（A）、咨询（C）和知情（I）。支持（S）类型是可选的。如果包括支持，则该矩阵称为 RASCI 矩阵，如果将其排除在外，则称为 RACI 矩阵。

响应性控制

一种安全控制，旨在推动对不良事件或偏离安全基线的情况进行修复。有关更多信息，请参阅在AWS上实施安全控制中的响应性控制。

保留

请参阅 7 R。

停用

请参阅 7 R。

检索增强生成（RAG）

一种生成式人工智能技术，其中 LLM 在生成响应之前引用其训练数据来源之外的权威数据来源。例如，RAG 模型可以对组织的知识库或自定义数据执行语义搜索。有关更多信息，请参阅什么是 RAG。

轮换

定期更新密钥以使攻击者更难访问凭证的过程。

行列访问控制（RCAC）

使用已定义访问规则的基本、灵活的 SQL 表达式。RCAC 由行权限和列掩码组成。

RPO

请参阅恢复点目标。

RTO

请参阅恢复时间目标。

运行手册

执行特定任务所需的一套手动或自动程序。它们通常是为了简化重复性操作或高错误率的程序而设计的。

S

SAML 2.0

许多身份提供商 (IdPs) 使用的开放标准。此功能支持联合单点登录 (SSO)，因此用户无需在 IAM 中为组织中的所有人创建用户即可登录 AWS 管理控制台 或调用 AWS API 操作。有关基于 SAML 2.0 的联合身份验证的更多信息，请参阅 IAM 文档中的关于基于 SAML 2.0 的联合身份验证。

SCADA

请参阅监督控制和数据采集。

SCP

请参阅服务控制策略。

机密密钥

在中 AWS Secrets Manager，您以加密形式存储的机密或受限信息，例如密码或用户凭证。它由密钥值及其元数据组成。密钥值可以是二进制、单个字符串或多个字符串。有关更多信息，请参阅 Secrets Manager 文档中的什么是 Amazon Secrets Manager 密钥？。

安全设计

一种在整个开发过程中都考虑安全的系统工程方法。

安全控制

一种技术或管理防护机制，可防止、检测或降低威胁行为体利用安全漏洞的能力。安全控制有以下四种类型：预防性、检测性、响应性和主动性。

安全固化

缩小攻击面，使其更能抵御攻击的过程。这可能包括删除不再需要的资源、实施授予最低权限的最佳安全实践或停用配置文件中不必要的功能等操作。

安全信息和事件管理（SIEM）系统

结合了安全信息管理（SIM）和安全事件管理（SEM）系统的工具和服务。SIEM 系统会收集、监控和分析来自服务器、网络、设备和其他来源的数据，以检测威胁和安全漏洞，并生成警报。

安全响应自动化

一种预定义的程序化操作，旨在自动响应或修复安全事件。这些自动化可作为侦探或响应式安全控制措施，帮助您实施 AWS 安全最佳实践。自动响应操作的示例包括修改 VPC 安全组、修补 Amazon EC2 实例或轮换凭证。

服务器端加密

由接收数据的人在目的地对数据 AWS 服务 进行加密。

服务控制策略（SCP）

一种策略，用于集中控制组织中所有账户的权限 AWS Organizations。 SCPs 定义防护措施或限制管理员可以委托给用户或角色的操作。您可以使用 SCPs 允许列表或拒绝列表来指定允许或禁止哪些服务或操作。有关更多信息，请参阅 AWS Organizations 文档中的服务控制策略。

服务端点

的入口点的 URL AWS 服务。您可以使用端点，通过编程方式连接到目标服务。有关更多信息，请参阅 AWS 一般参考 中的 AWS 服务 端点。

服务水平协议（SLA）

一份协议，阐明了 IT 团队承诺向客户交付的内容，比如服务正常运行时间和性能。

服务水平指示器（SLI）

对服务性能方面的衡量，例如错误率、可用性或吞吐量。

服务水平目标（SLO）

代表服务运行状况的目标指标，由服务水平指示器衡量。

责任共担模式

描述您在云安全与合规方面共同承担 AWS 的责任的模型。 AWS 负责云的安全，而您则负责云中的安全。有关更多信息，请参阅责任共担模式。

SIEM

请参阅安全信息和事件管理系统。

单点故障（SPOF）

应用程序的单个关键组件出现故障，可能会中断系统。

SLA

请参阅服务水平协议。

SLI

请参阅服务水平指示器。

SLO

请参阅服务水平目标。

split-and-seed 模型

一种扩展和加速现代化项目的模式。随着新功能和产品发布的定义，核心团队会拆分以创建新的产品团队。这有助于扩展组织的能力和服务，提高开发人员的工作效率，支持快速创新。有关更多信息，请参阅在 AWS 云中实现应用程序现代化的分阶段方法。

SPOF

请参阅单点故障。

星型架构

一种数据库组织结构，它使用一个大型事实表来存储事务数据或测量数据，并使用一个或多个较小的维度表来存储数据属性。此结构专为在数据仓库中使用或用于商业智能目的而设计。

strangler fig 模式

一种通过逐步重写和替换系统功能直至可以停用原有的系统来实现单体系统现代化的方法。这种模式用无花果藤作为类比，这种藤蔓成长为一棵树，最终战胜并取代了宿主。该模式是由 Martin Fowler 提出的，作为重写单体系统时管理风险的一种方法。有关如何应用此模式的示例，请参阅使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET（ASMX）Web 服务现代化。

子网

您的 VPC 内的一个 IP 地址范围。子网必须位于单个可用区中。

监督控制和数据采集（SCADA）

在制造业中，一种使用硬件和软件来监控实物资产和生产操作的系统。

对称加密

一种加密算法，它使用相同的密钥来加密和解密数据。

综合测试

以模拟用户交互的方式测试系统，以检测潜在问题或监控性能。你可以使用 Amazon S CloudWatch ynthetics 来创建这些测试。

系统提示

一种为 LLM 提供上下文、说明或准则以指导其行为的技术。系统提示有助于设置上下文并制定与用户交互的规则。

T

标签

键值对，用作组织资源的元数据。 AWS 标签有助于您管理、识别、组织、搜索和筛选 资源。有关更多信息，请参阅标记您的 AWS 资源。

目标变量

您在监督式 ML 中尝试预测的值。这也被称为结果变量。例如，在制造环境中，目标变量可能是产品缺陷。

任务列表

一种通过运行手册用于跟踪进度的工具。任务列表包含运行手册的概述和要完成的常规任务列表。对于每项常规任务，它包括预计所需时间、所有者和进度。

测试环境

请参阅环境。

训练

为您的 ML 模型提供学习数据。训练数据必须包含正确答案。学习算法在训练数据中查找将输入数据属性映射到目标（您希望预测的答案）的模式。然后输出捕获这些模式的 ML 模型。然后，您可以使用 ML 模型对不知道目标的新数据进行预测。

中转网关

一个网络传输中心，可用于将您的网络 VPCs 和本地网络互连。有关更多信息，请参阅 AWS Transit Gateway 文档中的什么是公交网关。

基于中继的工作流程

一种方法，开发人员在功能分支中本地构建和测试功能，然后将这些更改合并到主分支中。然后，按顺序将主分支构建到开发、预生产和生产环境。

可信访问权限

向您指定的服务授予权限，该服务可代表您在其账户中执行任务。 AWS Organizations 当需要服务相关的角色时，受信任的服务会在每个账户中创建一个角色，为您执行管理任务。有关更多信息，请参阅 AWS Organizations 文档中的AWS Organizations 与其他 AWS 服务一起使用。

优化

更改训练过程的各个方面，以提高 ML 模型的准确性。例如，您可以通过生成标签集、添加标签，并在不同的设置下多次重复这些步骤来优化模型，从而训练 ML 模型。

双披萨团队

一个小 DevOps 团队，你可以用两个披萨来喂食。双披萨团队的规模可确保在软件开发过程中充分协作。

U

不确定性

这一概念指的是不精确、不完整或未知的信息，这些信息可能会破坏预测式 ML 模型的可靠性。不确定性有两种类型：认知不确定性是由有限的、不完整的数据造成的，而偶然不确定性是由数据中固有的噪声和随机性导致的。

无差别任务

也称为繁重工作，即创建和运行应用程序所必需的工作，但不能为最终用户提供直接价值或竞争优势。无差别任务的示例包括采购、维护和容量规划。

上层环境

请参阅环境。

V

vacuum 操作

一种数据库维护操作，包括在增量更新后进行清理，以回收存储空间并提高性能。

版本控制

跟踪更改的过程和工具，例如存储库中源代码的更改。

VPC 对等连接

两者之间的连接 VPCs ，允许您使用私有 IP 地址路由流量。有关更多信息，请参阅 Amazon VPC 文档中的什么是 VPC 对等连接。

漏洞

损害系统安全的软件缺陷或硬件缺陷。

W

热缓存

一种包含经常访问的当前相关数据的缓冲区缓存。数据库实例可以从缓冲区缓存读取，这比从主内存或磁盘读取要快。

暖数据

不常访问的数据。查询此类数据时，通常可以接受中速查询。

窗口函数

一种对与当前记录有某种关联的一组行执行计算的 SQL 函数。窗口函数对于处理任务很有用，例如计算移动平均值或根据当前行的相对位置访问行的值。

工作负载

一系列资源和代码，它们可以提供商业价值，如面向客户的应用程序或后端过程。

工作流

迁移项目中负责一组特定任务的职能小组。每个工作流都是独立的，但支持项目中的其他工作流。例如，组合工作流负责确定应用程序的优先级、波次规划和收集迁移元数据。组合工作流将这些资产交付给迁移工作流，然后迁移服务器和应用程序。

WORM

请参阅一次写入多次读取。

WQF

请参阅 AWS 工作负载资格鉴定框架。

一次写入多次读取（WORM）

一种存储模型，可一次写入数据并防止数据被删除或修改。授权用户可以根据需要多次读取数据，但无法对其进行更改。此数据存储基础设施被认为不可变。

Z

零日漏洞利用

一种利用零日漏洞的攻击，通常为恶意软件。

零日漏洞

生产系统中不可避免的缺陷或漏洞。威胁主体可能利用这种类型的漏洞攻击系统。开发人员经常因攻击而意识到该漏洞。

零样本提示

为 LLM 提供执行任务的说明，但没有可以帮助指导的示例（样本）。LLM 必须使用预先训练的知识来处理任务。零样本提示的有效性取决于任务的复杂性和提示的质量。另请参阅少样本提示。

僵尸应用程序

平均 CPU 和内存使用率低于 5% 的应用程序。在迁移项目中，通常会停用这些应用程序。