AWS Certified Generative AI Developer - Professional (AIP-C01)
考试指南

简介

AWS Certified Generative AI Developer - Professional (AIP-C01) 考试面向担任 GenAI 开发人员角色的人员。考试旨在检验考生将基础模型 (FM) 有效地集成到应用程序和业务工作流中的能力。此认证要求考生展示所掌握的实践知识，运用 AWS 技术在生产环境中实施 GenAI 解决方案。

考试还验证考生完成以下任务的能力：

- 使用向量存储、检索增强生成 (RAG)、知识库和其他 GenAI 能架构，设计和实施解决方案。
- 将基础模型集成到应用程序和业务工作流中。
- 运用提示工程和管理技术。
- 实施代理式 AI 解决方案。
- 优化 GenAI 应用程序，来优化成本，提高性能和商业价值。
- 实施安全和监管措施以及负责任 AI 实践。
- 监控、优化 GenAI 应用程序以及进行故障排除。
- 评估基础模型的质量和责任。

目标考生说明

考生应满足下列要求：具备至少 2 年在 AWS 上构建生产级应用程序或使用开源技术的经验；具备常规 AI/ML 或数据工程经验；以及 1 年实施 GenAI 解决方案的实践经验。

AWS 知识推荐

考生应掌握以下 AWS 知识：

- 拥有使用 AWS 计算、存储和联网服务的经验
- 了解 AWS 安全最佳实践和身份管理
- 拥有使用 AWS 部署和基础设施即代码 (IaC) 工具的经验
- 熟悉 AWS 监控和可观测性服务
- 了解 AWS 成本优化原则

超出目标考生考试范围的工作任务

下表列出了不要求目标考生能够完成的相关工作任务。此列表并非详尽无遗。以下任务超出考试范围：

- 模型开发和训练
- 高级 ML 技术
- 数据工程和特征工程

有关考试中可能出现的技术和概念的列表、考试范围内的 AWS 服务和功能的列表，以及超出考试范围的 AWS 服务和功能的列表，请参阅附录。

考试内容

题型

考试包含以下一种或多种题型：

- **单选题：**有一个正确答案和三个错误答案（干扰项）。
- **多选题：**在五个或更多答案选项中有两个或更多正确答案。您必须选对所有正确答案才能得分。
- **排序题：**列出完成指定任务可能需要的 3 至 5 个答案。您必须选择正确的答案并按正确的顺序排列答案，才能得分。
- **配对题：**采用一列提示和一列答案的形式，提示列有 3 至 7 条提示。您必须将所有答案与提示正确匹配才能得分。

未回答的试题计为回答不正确。猜答不扣分。本考试包括 65 道计分试题，这些试题将影响您的分数。[1]

[1] 不适用于测试版考试。您可以在 AWS Certification 网站上找到有关测试版考试一般说明的更多信息。

不计分内容

本考试包括 10 道不计分试题，这些试题不影响您的分数。AWS 收集这些不计分试题的答题情况并进行评估，以便未来将这些试题作为计分试题。在考试中不会标明不计分试题。

考试结果

AWS Certified Generative AI Developer - Professional (AIP-C01) 考试结果分为及格和不及格两种。本考试按照 AWS 专业人员根据认证行业最佳实践和准则制定的最低标准进行评分。

您的考试结果换算分数为 100-1,000 分。最低及格分数为 750 分。您的分数表明您的总体考试答题情况以及是否通过考试。标准分模型有助于将难易程度可能略有不同的多种考试形式中的分数进行公平处理。

您的成绩单可能包含一个分类表，其中列出您在每个部分的考试成绩。本考试采用补偿评分模型，这意味着您无需在每个部分都达到及格分数。您只需通过整体考试即可。

考试的每个部分具有特定的权重，因此，某些部分的试题比其他部分多。分类表包含常规信息，用于突出显示您的强项和弱项。在解读各个部分的反馈时，请务必小心谨慎。

内容大纲

本考试指南介绍考试的权重、内容领域、任务和所需掌握的技能，并未列出考试的全部内容。

考试中考查的内容领域和相应的权重如下：

- 内容领域 1：基础模型集成、数据管理和合规性（计分内容的 31%）
- 内容领域 2：实施和集成（计分内容的 26%）
- 内容领域 3：AI 安全、保障与监管（计分内容的 20%）
- 内容领域 4：GenAI 应用程序的运营效率和优化（计分内容的 12%）
- 内容领域 5：测试、验证和故障排除（计分内容的 11%）

内容领域 1：基础模型集成、数据管理和合规性

任务 1.1：分析需求，设计 GenAI 解决方案。

技能 1.1.1：根据具体业务需求和技术限制，创建全面的架构设计（例如，使用合适的基础模型、集成模式、部署策略）。
技能 1.1.2：开发技术概念验证实施方案，来验证可行性、性能特征和商业价值，然后进行全面部署（例如，使用 Amazon Bedrock）。
技能 1.1.3：创建标准化技术组件，确保跨多个部署场景中实现一致的实施（例如，使用 AWS Well-Architected Framework、AWS WA 工具生成式人工智能剖析）。

任务 1.2：选择并配置基础模型。

技能 1.2.1：评估和选择基础模型，确保很好地契合特定的业务使用案例和技术要求（例如，使用性能基准、能力分析、限制评估等）。
技能 1.2.2：创建灵活的架构模式，可以动态选择模型和切换提供商，而且无需修改代码（例如，使用 AWS Lambda、Amazon API Gateway、AWS AppConfig）。
技能 1.2.3：设计具备弹性的 AI 系统，可确保服务中断期间的持续运行（例如，使用 AWS Step Functions 断路器模式，针对在有限区域提供的模型使用 Amazon Bedrock 跨区域推理，跨区域模型部署，优雅降级策略）。
技能 1.2.4：实施基础模型自定义部署和生命周期管理（例如，使用 Amazon SageMaker 人工智能，部署特定于领域的经过微调的模型；低秩自适应 [LoRA] 等参数-效率自适应技术和用于模型部署的适配器；用于版本控制和部署自定义模型的 SageMaker 模型注册表；用于更新模型的自动部署管道；失败部署的回滚策略；管理生命周期来停用和替换模型）。

任务 1.3：实施数据验证和处理管道，用于基础模型的使用。

技能 1.3.1：创建全面的数据验证工作流，确保数据符合质量标准以便供基础模型使用（例如，使用 AWS Glue 数据质量自动监测功能、SageMaker Data Wrangler、自定义 Lambda 函数、Amazon CloudWatch 指标）。
技能 1.3.2：创建数据处理工作流来处理复杂的数据类型，包括文本、图像、音频和表格数据等，这些数据在用于基础模型时有专门的处理要求（例如，使用 Amazon Bedrock 多模态模型、SageMaker Processing、AWS Transcribe、高级多模态管道架构）。
技能 1.3.3：根据特定于模型的要求，为基础模型推理设置输入数据的格式（例如，为 Amazon Bedrock API 请求使用 JSON 格式，适用于 SageMaker 人工智能终端节点的结构化数据准备，为基于对话的应用程序使用对话格式）。

技能 1.3.4：改善输入数据质量以提高基础模型回复的质量和一致性（例如，使用 Amazon Bedrock 重新格式化文本，使用 Amazon Comprehend 提取实体，使用 Lambda 函数对数据进行标准化）。

任务 1.4：设计和实施向量存储解决方案。

技能 1.4.1：创建专门用于基础模型增强的高级向量数据库架构，实现超越传统搜索功能的高效语义检索（例如，使用 Amazon Bedrock 知识库实现分层组织；使用带 Neural 插件的 Amazon OpenSearch Service 用于 Amazon Bedrock 集成，来进行基于主题的分割；Amazon RDS 与 Amazon S3 文档存储库结合使用；Amazon DynamoDB 与向量数据库结合使用来处理元数据和嵌入）。
技能 1.4.2：开发全面的元数据框架，提高基础模型交互的搜索查准率和上下文感知能力（例如，使用 S3 对象元数据作为文档时间戳，用于创作者信息的自定义属性，使用标记系统进行领域分类）。
技能 1.4.3：实施高性能向量数据库架构，来大规模优化语义搜索性能以便用于基础模型检索（例如，使用 OpenSearch 分片策略，针对专用领域的多索引方法，分层索引技术）。
技能 1.4.4：使用 AWS 服务创建集成组件来连接资源（例如，文档管理系统，知识库，用于 GenAI 应用程序中全面数据集成的内部 wiki）。
技能 1.4.5：设计和部署数据维护系统，确保向量存储包含当前的准确信息，以便用于基础模型增强（例如，使用递增更新机制、实时更改检测系统、自动同步工作流、定期刷新管道）。

任务 1.5：设计用于增强基础模型的检索机制。

技能 1.5.1：开发高效的文档分割方法来优化检索性能，用于基础模型上下文增强（例如，使用 Amazon Bedrock 分块功能，使用 Lambda 函数实施固定大小分块，根据内容结构采用分层式分块的自定义处理）。
技能 1.5.2：选择和配置优化的嵌入解决方案，为语义搜索高效地创建向量表示（例如，使用基于维度和领域拟合的 Amazon Titan 嵌入，评估 Amazon Bedrock 嵌入模型的性能特征，使用 Lambda 函数批量生成嵌入）。
技能 1.5.3：部署和配置向量搜索解决方案，启用语义搜索功能用于增强基础模型（例如，使用具有向量搜索功能的 OpenSearch Service，带有 pgvector 扩展的 Amazon Aurora，具有托管向量存储功能的 Amazon Bedrock 知识库）。
技能 1.5.4：创建高级搜索架构，提高针对基础模型上下文检索到的信息的相关性和准确率（例如，使用 OpenSearch 进行语义搜索、结合使用关键字和向量的混合搜索，Amazon Bedrock 重排器模型）。

技能 1.5.5：开发先进的查询处理系统来提高检索有效性和结果质量，以便用于增强基础模型（例如，使用 Amazon Bedrock 进行查询扩展，使用 Lambda 函数进行查询分解，使用 Step Functions 进行查询转换）。
技能 1.5.6： 创建一致的访问机制，实现与基础模型的无缝集成（例如，使用函数调用接口进行向量搜索，使用模型上下文协议 [MCP] 客户端进行向量查询，使用标准化 API 模式进行检索增强）。

任务 1.6：为基础模型交互实施提示工程策略和监管措施。

技能 1.6.1：创建高效的模型指令框架来控制基础模型的行为和输出（例如，使用 Amazon Bedrock 提示管理器强制实施角色定义，使用 Amazon Bedrock 防护机制强制实施负责任 AI 指导原则，使用模板配置对回复进行格式化处理）。
技能 1.6.2：构建交互式 AI 系统用于维护上下文并改善用户与基础模型的互动（例如，Step Functions 可用于澄清工作流，Amazon Comprehend 可用于意图识别，DynamoDB 可用于对话历史记录存储）。
技能 1.6.3：实施全面的提示管理和监管系统，对基础模型操作进行监督来确保实现操作一致性（例如，使用 Amazon Bedrock 提示管理器创建参数化模板和审批工作流，使用 Amazon S3 存储模板存储库，使用 AWS CloudTrail 跟踪使用情况，使用 Amazon CloudWatch Logs 记录访问活动）。
技能 1.6.4：开发质量保证系统，确保基础模型的提示有效性和可靠性（例如，使用 Lambda 函数验证预期输出，使用 Step Functions 测试边缘用例，使用 CloudWatch 测试提示回归）。
技能 1.6.5：增强基础模型性能，以迭代方式完善提示，并运用基本提示技巧之外的技术来提高回复质量（例如，使用结构化输入组件、输出格式规范、思维链指令模式、反馈循环）。
技能 1.6.6：设计复杂的提示系统，利用基础模型来处理复杂的任务（例如，使用 Amazon Bedrock 提示工作流管理器执行顺序提示链，基于模型回复的有条件分支，可重复使用的提示组件，集成式预处理和后处理步骤）。

内容领域 2： 实施和集成

任务 2.1：实施代理式 AI 解决方案和工具集成。

技能 2.1.1：开发智能化的自治系统，具备相应的内存和状态管理功能（例如，将 Strands Agents 与 AWS Agent Squad 结合用于多代理系统，使用 MCP 进行代理与工具的交互）。
技能 2.1.2：创建高级问题解决系统，赋予基础模型按照结构化推理步骤来分析和解决复杂问题的能力（例如，使用 Step Functions 实施 ReAct 模式和思维链推理方法）。
技能 2.1.3：开发具有保障措施的 AI 工作流，确保实现受控的基础模型行为（例如，使用 Step Functions 实施筛选停用词条件，使用 Lambda 函数实施超时机制，使用 IAM 策略强制实施资源边界，使用断路器来防范故障）。
技能 2.1.4：创建先进的模型协调系统，用于优化多种功能的性能（例如，使用专用基础模型执行复杂任务，使用自定义聚合逻辑进行模型组合，模型选择框架）。
技能 2.1.5：开发协作式 AI 系统，利用人员的专业知识增强基础模型能力（例如，使用 Step Functions 协调审核和审批流程，使用 API Gateway 实施反馈收集机制，人工增强模式）。
技能 2.1.6：实施智能工具集成来扩展基础模型功能，确保可靠的工具运行（例如，使用 Strands API 实施自定义行为，标准化函数定义，使用 Lambda 函数实施错误处理和参数验证方法）。
技能 2.1.7：开发模型扩展框架来增强基础模型能力（例如，使用 Lambda 函数实施无状态 MCP 服务器，提供轻量级工具访问；使用 Amazon ECS 实施 MCP 服务器，提供复杂的工具，使用 MCP 客户端库来确保一致的访问模式）。

任务 2.2：实施模型部署策略。

技能 2.2.1：根据具体的应用程序需求和性能要求部署基础模型（例如，使用 Lambda 函数进行按需调用，Amazon Bedrock 预置吞吐量配置，使用 SageMaker 人工智能终端节点实施混合解决方案）。
技能 2.2.2：部署不同于传统 ML 部署的基础模型解决方案，解决大型语言模型 (LLM) 的独特挑战（例如，实施基于容器的部署模型；针对内存要求、GPU 利用率和词元处理能力进行优化；遵循专门的模型加载策略）。
技能 2.2.3：开发优化的基础模型部署方法，平衡 GenAI 工作负载的性能和资源需求（例如，选择合适的模型，针对特定任务使用较小的预训练模型，使用基于 API 的模型级联来执行常规查询）。

任务 2.3：设计和实施企业集成架构。

技能 2.3.1：创建企业连接解决方案，将基础模型功能无缝融入到现有的企业环境中（例如，对旧式系统使用基于 API 的集成，使用事件驱动型架构实施松耦合，数据同步模式）。
技能 2.3.2：开发集成式 AI 功能，使用 GenAI 功能增强现有应用程序（例如，使用 API Gateway 实施微服务集成，将 Lambda 函数用作 Webhook 处理程序，使用 Amazon EventBridge 实施事件驱动型集成）。
技能 2.3.3：创建安全访问框架，确保实施了适当的安全控制措施（例如，在基础模型服务和企业系统之间使用身份联合验证，为模型和数据访问使用基于角色的访问控制，对基础模型采用最低权限 API 访问）。
技能 2.3.4：开发跨环境的 AI 解决方案，确保在各司法管辖区中的数据合规性，同时实现基础模型访问控制（例如，使用 AWS Outposts 进行本地数据集成，使用 AWS Wavelength 进行边缘部署，云服务与本地资源之间的安全路由）。
技能 2.3.5：实施 CI/CD 管道和 GenAI 网关架构，在企业环境中实施安全且合规的使用模式（例如，使用 AWS CodePipeline、AWS CodeBuild；为持续部署使用自动化测试框架；使用安全扫描和回滚支持来测试 GenAI 组件；集中式抽象层、可观测性和控制机制）。

任务 2.4：实施基础模型 API 集成。

技能 2.4.1：创建灵活的模型交互系统（例如，使用 Amazon Bedrock API 管理来自各种计算环境的同步请求，使用特定于语言的 AWS SDK 和 Amazon SQS 进行异步处理，使用 API Gateway 为自定义 API 客户端提供请求验证）。
技能 2.4.2：开发实时 AI 交互系统，提供来自基础模型的即时反馈（例如，使用 Amazon Bedrock 流式传输 API 进行递增式回复交付，使用 WebSocket 或服务器发送的事件实时生成文本，使用 API Gateway 实施分块传输编码）。
技能 2.4.3：创建具有弹性的基础模型系统来确保可靠运行（例如，使用 AWS SDK 进行指数回退，使用 API Gateway 管理速率限制，利用回退机制实现优雅降级，使用 AWS X-Ray 提供跨服务边界的可观测性）。
技能 2.4.4：开发智能模型路由系统用以优化模型选择（例如，使用应用程序代码实施静态路由配置；使用 Step Functions 基于内容动态路由到专用基础模型；基于指标的智能模型路由；使用 API Gateway 通过请求转换实现路由逻辑）。

任务 2.5：实施应用程序集成模式和开发工具。

技能 2.5.1：创建基础模型 API 接口，满足 GenAI 工作负载的特定要求（例如，使用 API Gateway 处理流式回复，词元限制管理，处理模型超时的重试策略）。
技能 2.5.2：开发易于访问的 AI 接口来加快基础模型的采用和集成（例如，使用 AWS Amplify 开发声明式 UI 组件，适用于 API 优先开发方法的 OpenAPI 规范，使用 Amazon Bedrock 提示工作流管理器开发无代码工作流构建器）。
技能 2.5.3：创建业务系统增强功能（例如，使用 Lambda 函数实施客户关系管理 [CRM] 增强功能，使用 Step Functions 编排文档处理系统，使用 Amazon Q 企业版数据来源提供内部知识工具，使用 Amazon Bedrock 数据自动化来管理自动数据处理工作流）。
技能 2.5.4：提高开发人员工作效率，加快 GenAI 应用程序开发工作流的速度（例如，使用 Amazon Q 开发者版生成和重构代码，API 辅助代码建议，AI 组件测试，性能优化）。
技能 2.5.5：开发高级 GenAI 应用程序来实施先进的 AI 功能（例如，使用 Strands Agents 和 AWS Agent Squad 实施 AWS 云原生的编排，使用 Step Functions 编排代理设计模式，使用 Amazon Bedrock 管理提示链模式）。
技能 2.5.6：提高基础模型应用程序的故障排除效率（例如，使用 CloudWatch Logs Insights 分析提示和回复，使用 X-Ray 跟踪 FM API 调用，使用 Amazon Q 开发者版实施特定于 GenAI 的错误模式识别）。

内容领域 3：AI 安全、保障与监管

任务 3.1：实施输入和输出安全控制措施。

技能 3.1.1：开发全面的内容安全系统，用于防止有害的用户信息输入到基础模型中（例如，使用 Amazon Bedrock 防护机制筛选内容，使用 Step Functions 和 Lambda 函数实施自定义审核工作流，实时验证机制）。
技能 3.1.2：创建内容安全框架来防止有害输出（例如，使用 Amazon Bedrock 防护机制筛选回复，使用专门的基础模型评估进行内容审核和毒性检测，使用文本到 SQL 的转换来确保获得确定性结果）。

技能 3.1.3：开发准确率验证系统来减少基础模型回复中的幻觉（例如，使用 Amazon Bedrock 知识库确定回复依据和进行事实核查，通过置信度评分和语义相似度搜索进行验证，使用 JSON 架构强制实现结构化输出）。
技能 3.1.4：创建深度防御安全系统来提供全面的保护，防止基础模型滥用（例如，使用 Amazon Comprehend 开发预处理筛选条件，使用 Amazon Bedrock 实施基于模型的防护机制，使用 Lambda 函数执行后处理验证，使用 API Gateway 实施 API 响应筛选）。
技能 3.1.5：实施高级威胁检测功能，防范对抗输入和安全漏洞（例如，使用提示注入和越狱检测机制，输入清理和内容筛选条件，安全分类器，自动对抗测试工作流）。

任务 3.2：实施数据安全和隐私控制措施。

技能 3.2.1：开发受保护的 AI 环境，确保基础模型部署的全面安全性（例如，使用 VPC 端点隔离网络，使用 IAM 策略强制执行安全数据访问模式，使用 AWS Lake Formation 提供精细的数据访问，使用 CloudWatch 监控数据访问）。
技能 3.2.2：开发隐私保护系统，在基础模型交互期间保护敏感信息（例如，使用 Amazon Comprehend 和 Amazon Macie 检测个人身份信息 [PII]，Amazon Bedrock 原生的数据隐私功能，使用 Amazon Bedrock 防护机制筛选输出，使用 Amazon S3 生命周期配置来实施数据留存策略）。
技能 3.2.3：创建注重保护隐私的 AI 系统，在保护用户隐私的同时，确保基础模型的实用性和有效性（例如，使用数据掩蔽技术，Amazon Comprehend PII 检测，敏感信息匿名化策略、Amazon Bedrock 防护机制）。

任务 3.3：实施 AI 监管和合规性机制。

技能 3.3.1：制定合规性框架，确保基础模型部署的监管合规（例如，使用 SageMaker 人工智能开发编程式模型卡片，使用 AWS Glue 自动跟踪数据血统，使用元数据标记进行系统化数据来源归因，使用 CloudWatch Logs 收集全面的决策日志）。
技能 3.3.2：实施数据来源跟踪措施，维护 GenAI 应用程序的可追溯性（例如，使用 AWS Glue Data Catalog 注册数据来源，使用元数据标记对基础模型生成的内容进行来源归因，使用 CloudTrail 进行审计日志记录）。
技能 3.3.3：创建企业监管系统，确保以一致的方式监督基础模型实施（例如，使用符合企业策略、监管要求和负责任 AI 原则的综合性框架）。
技能 3.3.4：实施持续监控和高级治理控制措施，支持安全审计，做好监管准备工作（例如，使用自动化功能，检测滥用、偏差和策略违规行为；监控偏差偏移；自动报警和补救工作流；词元级编辑；响应日志记录、AI 输出策略筛选条件）。

任务 3.4：实施负责任 AI 原则。

技能 3.4.1：开发透明的 AI 系统用于基础模型输出（例如，使用推理显示提供面向用户的解释，使用 CloudWatch 收集置信度指标并量化不确定性，提供用于来源归因的证据，使用 Amazon Bedrock 代理追踪提供推理跟踪）。
技能 3.4.2：应用公平性评估来确保基础模型输出中没有偏见（例如，使用 CloudWatch 中的预定义公平性指标、Amazon Bedrock 提示管理器和 Amazon Bedrock 提示工作流管理器，执行系统化 A/B 测试；使用 LLM-asa-Judge 解决方案执行自动模型评估）。
技能 3.4.3：开发符合策略要求的 AI 系统，确保遵循负责任 AI 实践（例如，根据策略要求使用 Amazon Bedrock 防护机制，使用模型卡片记录基础模型限制，使用 Lambda 函数自动执行合规性检查）。

内容领域 4：GenAI 应用程序的运营效率和优化

任务 4.1：实施成本优化和资源效率策略。

技能 4.1.1：开发词元效率系统来降低基础模型成本，同时保持模型有效性（例如，使用词元估算和跟踪、上下文窗口优化、响应大小控制、提示压缩、上下文修剪、回复限制）。
技能 4.1.2：创建经济高效的模型选择框架（例如，使用成本-能力权衡评估，基于查询复杂度的分级基础模型使用，推理成本与回复质量的平衡，性价比测量，高效的推理模式）。
技能 4.1.3：开发高性能基础模型系统，充分提高 GenAI 工作负载的资源利用率和吞吐量（例如，使用批处理策略、容量规划、使用率监控、弹性伸缩配置、预置吞吐量优化）。
技能 4.1.4：创建智能缓存系统，通过避免不必要的基础模型调用来降低成本并缩短响应时间（例如，使用语义缓存、结果指纹识别、边缘缓存、确定性请求哈希、提示缓存）。

任务 4.2：优化应用程序性能。

技能 4.2.1： 创建响应灵敏的 AI 系统，解决延迟与成本的权衡问题，并改善基础模型的用户体验（例如，使用预计算来执行可预测的查询，为注重时间的应用程序使用延迟优化的 Amazon Bedrock 模型，为复杂工作流使用并行请求，回复流式传输，性能基准测试）。

技能 4.2.2：增强检索性能，提高检索信息的相关性和速度，用以增强基础模型上下文（例如，使用索引优化、查询预处理、带自定义评分的混合搜索实施方案）。

技能 4.2.3：实施基础模型吞吐量优化，解决 GenAI 工作负载的特定吞吐量挑战（例如，使用词元处理优化、批量推理策略、并行模型调用管理）。

技能 4.2.4：增强基础模型性能，针对特定 GenAI 使用案例实现优化结果（例如，使用特定于模型的参数配置，通过 A/B 测试评估改进，适当的温度处理，以及根据要求进行 top-k/top-p 选择）。

技能 4.2.5：创建专用于基础模型工作负载的高效资源分配系统（例如，根据词元处理需求使用容量规划，对提示和完成模式的资源利用率进行监控，针对 GenAI 流量模式进行优化的弹性伸缩配置）。

技能 4.2.6：优化 GenAI 工作流的基础模型系统性能（例如，针对提示-完成模式使用 API 调用分析，针对检索增强的向量数据库查询优化，特定于 LLM 推理的缩短延迟技术，高效的服务通信模式）。

任务 4.3：为 GenAI 应用程序实施监控系统。

技能 4.3.1：创建全面的可观测性系统，提供对基础模型应用程序性能的全面监测能力（例如，使用运营指标、性能跟踪、基础模型交互跟踪、业务影响指标和自定义控制面板）。

技能 4.3.2： 施全面的 GenAI 监控系统，主动识别问题，评估特定于基础模型实施的关键绩效指标（例如，使用 CloudWatch 跟踪词元使用情况、提示有效性、幻觉率和回复质量；针对词元爆发模式和回复偏差的异常检测；使用 Amazon Bedrock 模型调用日志进行详细的请求和回复分析、确定性能基准、成本异常检测）。

技能 4.3.3：开发集成的可观测性解决方案，为基础模型应用程序提供可用于指导操作的见解（例如，使用运营指标控制面板，业务影响力可视化，合规性监控，取证可追溯性和审计日志记录，用户互动跟踪、模型行为模式跟踪）。

技能 4.3.4：创建工具性能框架来确保为基础模型优化工具操作和利用率（例如，使用调用模式跟踪，收集性能指标，工具调用可观测性和多代理协调跟踪，用于异常检测的使用率基准）。

技能 4.3.5：创建向量存储操作管理系统，确保优化向量存储操作和可靠性，用于增强基础模型（例如，对向量数据库使用性能监控，自动化索引优化例程，数据质量验证流程）。
技能 4.3.6：开发特定于基础模型的故障排除框架，用于识别传统 ML 系统中没有的 GenAI 独有故障模式（例如，使用黄金数据集检测幻觉，使用输出差异对比技术开展回复一致性分析，通过推理路径跟踪来识别逻辑错误，专用可观测性管道）。

内容领域 5：测试、验证和故障排除

任务 5.1：为 GenAI 实施评估系统。

技能 5.1.1：制定全面的评估框架，在传统 ML 的评估方法之外，利用新方法来评估基础模型输出的质量和有效性（例如，使用相关性、事实准确率、一致性和流畅性指标）。
技能 5.1.2：创建系统化模型评估系统来确定优化配置（例如，使用 Amazon Bedrock 模型评估、A/B 测试和基础模型的金丝雀测试，多模型评估，通过成本性能分析来衡量词元效率，延迟-质量比率和业务成果）。
技能 5.1.3：开发以用户为中心的评估机制，根据用户体验持续改进基础模型性能（例如，使用反馈界面，针对模型输出的评级系统，用于评估回复质量的注释工作流）。
技能 5.1.4：创建系统化质量保证流程，维护一致的基础模型性能标准（例如，使用持续评估工作流，对模型输出进行回归测试，部署自动化质量控制机制）。
技能 5.1.5：开发全面的评估系统，确保从多个角度对基础模型输出进行全面评估（例如，使用 RAG 评估，使用 LLM-asa-Judge 技术进行自动质量评估，人员反馈收集界面）。
技能 5.1.6：实施检索质量测试，评估和优化检索组件来增强基础模型（例如，使用相关性评分，上下文匹配验证，检索延迟测量）。
技能 5.1.7：开发代理性能框架，确保代理高效且正确地执行任务（例如，任务完成率量度，工具使用有效性评估，Amazon Bedrock 代理评估，多步骤工作流中的推理质量评估）。
技能 5.1.8：创建全面的报告系统，高效地向利益攸关方传递绩效指标和见解，用于实施基础模型（例如，使用可视化工具，自动报告机制，模型比较可视化）。
技能 5.1.9：创建部署验证系统，用于在基础模型更新期间保持可靠性（例如，使用合成用户工作流，针对幻觉率和语义漂移的 AI 特定输出验证，通过自动质量检查确保回复一致性）。

任务 5.2：对 GenAI 应用程序进行故障排除。

技能 5.2.1：解决内容处理问题，确保在基础模型交互中完善地处理必要的信息（例如，使用上下文窗口溢出诊断、动态分块策略、提示设计优化、与截断相关的错误分析）。
技能 5.2.2：诊断和解决基础模型集成问题，识别并修复特定于 GenAI 服务的 API 集成问题（例如，使用错误日志记录、请求验证、回复分析）。
技能 5.2.3：对提示工程问题进行故障排除，利用基本提示调整之外的方法，提高基础模型的回复质量和一致性（例如，使用提示测试框架、版本比较、系统化改进）。
技能 5.2.4：对检索系统问题进行故障排除，识别并解决影响信息检索有效性的问题，来增强基础模型（例如，使用模型回复相关性分析、嵌入质量诊断、漂移监控、向量化问题解决、分块和预处理补救、向量搜索性能优化）。
技能 5.2.5：对提示维护问题进行故障排除，来持续改进基础模型交互的性能（例如，使用模板测试和 CloudWatch Logs 来诊断提示混淆，使用 X-Ray 实施提示可观测性管道，使用架构验证来检测格式不一致的情况，使用系统化提示优化工作流）。

附录

考试中可能出现的技术和概念

下表包含考试中可能出现的技术和概念。此列表并非详尽无遗，并且可能会更改。表中项目的顺序和位置并不表明它们在考试中的相对权重或重要性：

- 检索增强生成 (RAG)
- 向量数据库和嵌入
- 提示工程和管理
- 基础模型 (FM) 集成
- 代理式 AI 系统
- 负责任 AI 实践
- 内容安全和审核
- 模型评估和验证
- AI 工作负载的成本优化
- AI 应用程序的性能优化
- AI 系统的监控和可观测性
- AI 应用程序的安全和监管
- API 设计和集成模式
- 事件驱动型架构
- 无服务器计算
- 容器编排
- 基础设施即代码 (IaC)
- AI 应用程序的 CI/CD
- 混合云架构
- 企业系统集成

考试中提及的 AWS 服务

AWS Certification 使用广为人知的 AWS 服务名的官方简称，这些简称包含缩写或附加说明信息，能够减少本次考试中在阅读方面的负担。例如，Amazon Simple Notification Service (Amazon SNS) 在考试中以 Amazon SNS 的形式出现。

考试中的帮助功能（适用于所有问题）包含 AWS 服务简称以及相应全名的列表。

您可以在 AWS Certification 网站上查阅 AWS 服务名，获取在考试中以简称形式显示的服务列表。列表上显示但超出考试范围的任何服务都不会出现在考试中。

注意： 并非每个缩写在考试中都有完整拼写或者在“帮助”功能中可用。某些 AWS 服务的官方全名包含从不展开的缩写（例如，Amazon API Gateway、Amazon EMR）。考试中还可能包含目标受众应该知道的其他缩写。

考试范围内的 AWS 服务和功能

下表列出了考试范围内的 AWS 服务和功能。此列表并非详尽无遗，并且可能会更改。AWS 产品/服务的类别与产品/服务的主要功能一致：

分析：

- Amazon Athena
- Amazon EMR
- AWS Glue
- Amazon Kinesis
- Amazon OpenSearch Service
- Amazon QuickSight
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

应用程序集成：

- Amazon AppFlow
- AWS AppConfig
- Amazon EventBridge
- Amazon SNS
- Amazon SQS
- AWS Step Functions

计算：

- AWS App Runner
- Amazon EC2
- AWS Lambda
- AWS Lambda@Edge
- AWS Outposts
- AWS Wavelength

容器：

- Amazon ECR
- Amazon ECS
- Amazon EKS
- AWS Fargate

客户参与：

- Amazon Connect

数据库：

- Amazon Aurora
- Amazon DocumentDB
- Amazon DynamoDB
- Amazon DynamoDB Streams
- Amazon ElastiCache
- Amazon Neptune
- Amazon RDS

开发工具：

- AWS Amplify
- AWS CDK
- AWS CLI
- AWS CloudFormation
- AWS CodeArtifact
- AWS CodeBuild
- AWS CodeDeploy

- AWS CodePipeline
- AWS 工具和 SDK
- AWS X-Ray

机器学习：

- Amazon Augmented AI
- Amazon Bedrock
- Amazon Bedrock AgentCore
- Amazon Bedrock 知识库
- Amazon Bedrock 提示管理器
- Amazon Bedrock 提示工作流管理器
- Amazon Comprehend
- Amazon Kendra
- Amazon Lex
- Amazon Q 企业版
- Amazon Q 企业版应用程序
- Amazon Q 开发者版
- Amazon Rekognition
- Amazon SageMaker 人工智能
- Amazon SageMaker Clarify
- Amazon SageMaker Data Wrangler
- Amazon SageMaker Ground Truth
- Amazon SageMaker JumpStart
- Amazon SageMaker Model Monitor
- Amazon SageMaker 模型注册表
- Amazon SageMaker Neo
- Amazon SageMaker Processing
- Amazon SageMaker 融通式合作开发工作室
- Amazon Textract
- Amazon Titan
- Amazon Transcribe

管理和监管：

- AWS Auto Scaling
- AWS Chatbot
- AWS CloudTrail
- Amazon CloudWatch
- Amazon CloudWatch Logs
- Amazon CloudWatch Synthetics
- AWS 成本异常检测
- AWS Cost Explorer 成本管理服务
- Amazon Managed Grafana
- AWS Service Catalog
- AWS Systems Manager
- AWS Well-Architected Tool

迁移与传输：

- AWS DataSync
- AWS Transfer Family

联网和内容分发：

- Amazon API Gateway
- AWS AppSync
- Amazon CloudFront
- 弹性负载均衡 (ELB)
- AWS Global Accelerator
- AWS PrivateLink
- Amazon Route 53
- Amazon VPC

安全性、身份与合规性：

- Amazon Cognito
- AWS Encryption SDK
- IAM
- IAM 访问权限分析器
- IAM Identity Center

- AWS KMS
- Amazon Macie
- AWS Secrets Manager
- AWS WAF

存储：

- Amazon EBS
- Amazon EFS
- Amazon S3
- Amazon S3 Intelligent-Tiering
- Amazon S3 生命周期策略
- Amazon S3 Cross-Region Replication

超出考试范围的 AWS 服务和功能

下表列出了超出考试范围的 AWS 服务和功能。此列表并非详尽无遗，并且可能会更改。与考试的目标工作职责完全无关的 AWS 产品/服务被排除在此列表之外：

应用程序集成：

- Amazon MQ

分析：

- AWS Clean Rooms
- AWS Data Exchange
- Amazon DataZone
- Amazon FinSpace

区块链：

- Amazon Managed Blockchain (AMB)

业务应用程序：

- 企业版 Alexa
- Amazon Chime
- AWS Wickr
- Amazon WorkDocs
- Amazon WorkMail

云财务管理：

- AWS Budgets
- AWS 成本和使用情况报告
- 预留实例报告
- AWS Savings Plans

计算：

- AWS Batch
- Amazon EC2 Image Builder
- Amazon ECS Anywhere
- Amazon EKS Anywhere
- AWS Elastic Beanstalk
- Amazon Lightsail
- AWS Local Zones
- AWS Serverless Application Repository

容器：

- AWS App2Container
- AWS Copilot
- AWS 云端 Red Hat OpenShift 服务 (ROSA)

客户参与：

- Amazon SES

数据库：

- Amazon Keyspaces
- Amazon Quantum Ledger Database (Amazon QLDB)
- Amazon Redshift
- Amazon Timestream

开发工具：

- AWS Cloud9
- AWS CloudShell
- Amazon CodeGuru
- AWS CodeStar
- Amazon Corretto

终端用户计算：

- Amazon AppStream 2.0
- Amazon WorkLink
- Amazon WorkSpaces
- Amazon WorkSpaces Web

前端 Web 和移动：

- AWS Device Farm
- Amazon Location Service
- Amazon Pinpoint

游戏开发：

- Amazon GameLift
- Amazon Lumberyard

物联网 (IoT)：

- AWS IoT 1-Click
- AWS IoT Analytics
- AWS IoT Button
- AWS IoT Core
- AWS IoT Device Defender
- AWS IoT Device Management
- AWS IoT Events
- AWS IoT FleetWise
- AWS IoT Greengrass
- AWS IoT SiteWise
- AWS IoT TwinMaker

管理和监管：

- AWS 控制台移动应用程序
- AWS Health Dashboard
- AWS License Manager
- AWS Proton
- AWS Trusted Advisor

机器学习：

- AWS DeepComposer
- AWS DeepRacer
- Amazon DevOps Guru
- Amazon Forecast
- Amazon Fraud Detector
- Amazon HealthLake
- Amazon Lookout for Equipment
- Amazon Lookout for Metrics
- Amazon Lookout for Vision
- Amazon Monitron
- AWS Panorama

媒体服务：

- Amazon Elastic Transcoder
- AWS Elemental MediaConnect
- AWS Elemental MediaConvert
- AWS Elemental MediaLive
- AWS Elemental MediaPackage
- AWS Elemental MediaStore
- AWS Elemental MediaTailor
- Amazon Interactive Video Service
- Amazon Kinesis Video Streams
- Amazon Nimble Studio

迁移与传输：

- AWS Application Discovery Service
- AWS Application Migration Service
- CloudEndure Migration
- AWS Migration Hub
- AWS Snow Family

联网和内容分发：

- AWS App Mesh
- AWS Cloud Map
- AWS Direct Connect
- AWS 私有 5G 服务
- AWS Transit Gateway
- AWS VPN

量子技术：

- Amazon Braket

机器人：

- AWS RoboMaker

卫星：

- AWS Ground Station

安全性、身份与合规性：

- AWS Artifact
- AWS Audit Manager
- AWS Certificate Manager
- AWS CloudHSM
- Amazon Detective
- AWS Directory Service
- AWS Firewall Manager
- Amazon GuardDuty
- AWS Network Firewall

- AWS Private CA
- AWS Resource Access Manager (AWS RAM)
- AWS Security Hub
- AWS Shield
- Amazon Verified Permissions

存储：

- AWS Backup
- Amazon FSx
- 适用于 Lustre 的 Amazon FSx
- 适用于 NetApp ONTAP 的 Amazon FSx
- 适用于 OpenZFS 的 Amazon FSx
- 适用于 Windows File Server 的 Amazon FSx
- Amazon S3 Glacier
- AWS Snow Family
- AWS Storage Gateway

调查问卷

本考试指南对您有多大帮助？ 请参与问卷调查，反馈您的看法。