在 上建置企業就緒的生成式 AI 平台 AWS

AWS 方案指引

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS 方案指引: 在 上建置企業就緒的生成式 AI 平台 AWS

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商標和商業外觀不得用於任何非 Amazon 的產品或服務，也不能以任何可能造成客戶混淆、任何貶低或使 Amazon 名譽受損的方式使用 Amazon 的商標和商業外觀。所有其他非 Amazon 擁有的商標均為其各自擁有者的財產，這些擁有者可能附屬於 Amazon，或與 Amazon 有合作關係，亦或受到 Amazon 贊助。

Table of Contents

簡介 1
目標對象 2
目標 2
分層方法 3
第 1 層：資料和基礎設施 5
基礎基礎設施 5
向量儲存和擷取 5
運算基礎設施 6
實作建議 6
第 2 層：基礎模型存取 8
實驗和自訂 8
評估 8
目標 8
指標 9
技術 10
實作建議 10
第 3 層：安全與控管 12
核心紀律 12
建議的控制項 12
範圍矩陣 13
實作建議 14
第 4 層：可重複的應用程式模式 15
採用 AI 技術的助理 15
RAG 助理 16
客服人員 17
智慧文件處理 17
產生和摘要 18
程式碼助理 18
多模態 AI 應用程式 19
實作建議 19
最佳實務 21
組織結構和控管 21
AI 卓越中心 21
模型控管委員會 21

標準化和技術卓越性 21
模式程式庫和工具 21
全企業存取架構 22
程序實作 22
服務管理 22
模型選擇和評估 22
效能監控和最佳化 22
安全和合規 23
實作建議 23
後續步驟 24
資源 25
AWS 方案指引 25
AWS 服務文件 25
其他 AWS 資源 26
其他資源 26
文件歷史紀錄 27
..... xxviii

在 上建置企業就緒的生成式 AI 平台 AWS

Ratan Kumar、Jeffrey Zeng 和 Viral Shah，Amazon Web Services

2025 年 6 月 (文件歷史記錄)

通常會聽到原型很容易、示範很酷，但生產很困難。生成式 AI 應用程式尤其如此。此策略文件提供全方位的指引，協助組織實作安全、可擴展且有效的生成式 AI 平台 AWS。它概述了四層架構方法，其中包含可靠的基礎設施、基礎模型選擇、強大的安全性和控管，以及可重複的應用程式模式。

AI 可協助組織提供差異化的產品和服務。在 AI 的不同領域中，生成式 AI 已成為強大的媒介。它重新定義了企業如何模擬案例、最佳化操作，並為他們的客戶提供令人信服的數位體驗。

雖然許多組織正在建立許多生成式 AI 原型，但他們很難將這些原型擴展為生產就緒的解決方案。為了促進廣泛採用生成式 AI，企業必須解決幾個關鍵挑戰：

- 基礎設施準備度和可擴展性
- 安全與合規要求
- 負責任的 AI 實作
- 與現有應用程式和程序整合
- 保護敏感資料和智慧財產權
- 成本最佳化和投資報酬率 (ROI) 測量

此外，您需要考慮採用生成式 AI 的解決方案的 ROI。如何將生成式 AI 整合到現有的應用程式？ 如何確保 AI 產生的建議安全可靠？ 使用現有基礎模型或微調模型時，如何協助保護敏感資料和智慧財產權？

透過開發企業就緒的生成式 AI 功能，您可以克服這些挑戰。此策略文件提供路線圖，協助您了解並加速生成式 AI 計畫，同時維持安全與安全標準的合規性。它還描述了如何使用 AWS 服務，例如 Amazon Bedrock 和 Amazon Q，以實現您的目標業務成果。這些服務可以驅動從客戶服務自動化到開發人員生產力工具的所有功能。他們也可以協助您實作聊天式助理和 AI 客服人員，以轉換您的業務營運。

> **Note**
>
> 在實作本指南之前，我們建議您使用生成式 AI 工作負載評估來評估組織的生成式 AI 準備程度。此策略文件以本指南中記錄的企業生成式 AI 採用已建立的 AWS 最佳實務和模式為基礎。

目標對象

此策略文件適用於執行長 (CEOs)、技術長 (CTOs)、資訊長 (CIOs) 和其他負責推動 AI 創新和數位轉型的資深技術領導者。若要使用此策略文件，您應該熟悉商業和技術管理概念，例如企業架構、雲端運算模型、IT 控管架構、策略技術規劃和數位轉型計劃。

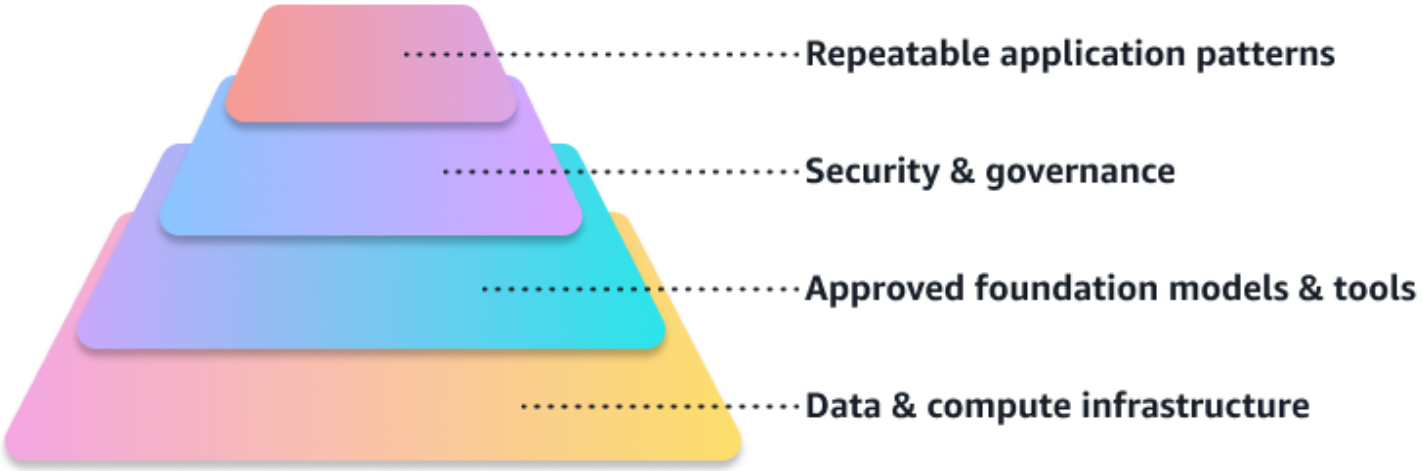
目標

透過實作此策略文件中的建議，組織可以實現以下目標業務成果：

- 將原型轉換為生產就緒的生成式 AI 解決方案，以提供可衡量的商業價值
- 協助分散式團隊更快速地進行實驗，同時維護安全性和管理
- 透過適當的控制和管控，建立對已核准基礎模型的全企業存取權
- 透過可重複使用的應用程式模式和標準化工具來減少開發時間和成本
- 透過安全的自助式存取生成式 AI 功能加速創新
- 實作負責任的 AI 實務，以促進道德使用和合規
- 建立可擴展的架構來評估和採用新的基礎模型
- 透過適當的驗證和控制機制，建立對 AI 產生內容的信任
- 最佳化營運成本，同時維持高效能和安全標準
- 將生成式 AI 功能無縫整合到現有的企業應用程式中

採用分層方法來進行生成式 AI 開發

為了建置強大且可擴展的生成式 AI 平台，組織需要一個全面的環境來實現創新，同時保持控制和安全性。此策略文件建議由四個基本層組成的企業就緒環境，如下圖所示。



下列每個層在企業成功採用和擴展生成式 AI 方面都扮演重要角色：

1. 資料和運算基礎設施 – 可靠的基礎設施為開發、訓練和部署生成式 AI 應用程式提供了基礎。此層可確保組織擁有必要的運算資源和資料管理功能，以支援從小型實驗到大規模生產部署的各種使用案例。對於 Amazon Elastic Compute Cloud (Amazon EC2)，請考慮使用 AWS Trainium、AWS Inferentia、UltraClusters、Elastic Fabric Adapter、適用於 ML、Nitro System 和 Neuron 的容量區塊。 AWS AWS
2. 已核准的基礎模型和工具 – 此層可讓您的組織存取已核准、預先訓練的模型，同時維護安全性和合規性。團隊可以實驗不同的模型、評估其有效性，並針對特定使用案例選擇最適合的模型。您可以使用 Amazon Bedrock 建立護欄、建置代理程式、自訂模型或匯入您自己的自訂模型。
3. 安全性與控管 – 此層可協助您確保生成式 AI 應用程式遵守組織政策、合規要求、法律和隱私權要求，以及道德準則。在此圖層中，您會在所有生成式 AI 計畫中實作一致的安全控制和控管機制。
4. 可重複的應用程式模式 – 此層透過為常見的生成式 AI 使用案例提供標準化範本和最佳實務，加速開發並提升一致性。提供可重複的模式有助於防止重複工作，並快速擴展成功的解決方案。可重複模式的範例包括智慧文件處理、擷取增強生成 (RAG)、聊天式助理和客服人員工作流程。

這種分層方法提供下列主要優點：

- 它可清楚區分生成式 AI 實作的不同層面。
- 它在所有生成式 AI 計畫中促進一致的安全性和管理。

- 它可協助您建置可擴展的基礎設施，以滿足組織需求。
- 它將加速開發的方法標準化。
- 它透過經過驗證的模式和實務降低風險。

透過實作這四個層，組織可以為其生成式 AI 計畫建立堅實的基礎，同時保持適應新需求和機會的靈活性。

第 1 層：為生成式 AI 建置可靠的資料和運算基礎設施

對於開發生成式 AI 應用程式，特別是需要訓練或微調基礎模型時，強大的資料基礎和運算基礎設施至關重要。隨著企業開始生成式 AI 旅程，他們需要可支援整個機器學習生命週期的基礎設施。他們還需要在效能、成本和營運效率之間取得適當的平衡。

可靠的生成式 AI 基礎設施應提供下列三個關鍵元件：基礎基礎設施、向量儲存和擷取基礎設施，以及運算基礎設施。這些元件共同提供彈性，以滿足任何專案的需求，無論其規模、需求或環境。

本節包含下列主題：

- API 型實作的基礎基礎設施
- 向量儲存和擷取基礎設施
- 用於模型訓練和微調的高效能運算基礎設施
- 實作建議

API 型實作的基礎基礎設施

大多數組織透過 APIs 使用預先訓練的基礎模型開始生成式 AI 旅程。Amazon Bedrock 透過統一 API 提供領導基礎模型的無伺服器存取。這可協助您實驗和部署生成式 AI 應用程式，而無需管理複雜的基礎設施。對於這些實作，您需要下列項目：

- 執行應用程式的 Amazon Elastic Compute Cloud (Amazon EC2) 一般用途執行個體。
- Amazon Simple Storage Service (Amazon S3) 用於儲存應用程式資料和輸出。
- Amazon CloudWatch 用於監控和記錄應用程式的效能和用量。
- （選用）AWS Lambda 用於建置事件驅動生成式 AI 應用程式時的無伺服器運算。

向量儲存和擷取基礎設施

隨著生成式 AI 應用程式成熟，您可能需要使用特定領域的知識和內容來增強這些應用程式。使用擷取增強生成 (RAG)，基礎模型會參考訓練資料來源之外的授權資料來源，例如組織的資料或文件，然後再產生回應。如需詳細資訊，請參閱 上的擷取增強產生選項和架構 AWS。

Amazon Bedrock 知識庫提供建置 RAG 應用程式的完整受管解決方案。它可協助您安全地將企業資料與基礎模型連線。如需更多彈性，您可以使用 Amazon OpenSearch Service 搭配向量搜尋功能，或者

您可以使用 Amazon Relational Database Service (Amazon RDS) for PostgreSQL 搭配 pgvector 延伸模組來存放向量內嵌。

為了有效率的向量操作，請考慮使用 Amazon EC2 運算最佳化執行個體進行內嵌產生。另請考慮使用 Amazon ElastiCache 快取經常存取的內嵌，以最佳化效能並降低成本。

如需向量資料庫和快取的詳細資訊，請參閱什麼是向量資料庫？以及為 RAG 使用案例選擇 AWS 向量資料庫。

用於模型訓練和微調的高效能運算基礎設施

對於準備好自訂基礎模型的組織， 為模型訓練和微調 AWS 提供全面的基礎設施。您可以使用 Amazon Simple Storage Service (Amazon S3) 作為低成本、可擴展且高耐用性的物件儲存體，用於建置資料平台，以及儲存訓練資料和訓練模型。此外， AWS Glue 是一種無伺服器資料整合服務，可協助您準備資料以進行模型訓練。

對於訓練基礎設施，Amazon SageMaker AI 提供全受管的機器學習環境，其中包含建置、訓練和部署模型所需的工具和工作流程。使用 SageMaker AI 可以大幅降低您的營運開銷。請考慮使用加速運算執行個體，例如 G 和 P 執行個體系列。這些執行個體系列可讓您從 NVIDIA 存取最新的 GPUs，以進行機器學習訓練和推論。您也可以使用 AWS Trainium，這是專門建置的機器學習加速器，可將訓練時間加速高達 50%，同時降低成本。

對於真正的大型專案，您可以使用 Amazon EC2 UltraClusters。UltraClusters 由數千個位於指定 AWS 可用區域中並相互連接的加速 Amazon EC2 執行個體組成。UltraClusters 可以擴展到數千個 GPUs 或機器學習加速器。它們提供每秒 exa-floating 點操作 (exaflops) 的彙總運算容量，可減少訓練時間，並可將time-to-solution從數週縮短為幾天。

實作建議

請考慮下列建議，為您的生成式 AI 專案設定可擴展且符合成本效益的基礎設施：

- 若要快速實驗和初始部署，請從 Amazon Bedrock 開始，並為您的應用程式使用一般用途運算執行個體。
- 隨著您的需求不斷演進，請使用 Amazon Bedrock 知識庫或 Amazon OpenSearch Service 實作向量儲存解決方案。相應地擴展您的基礎設施。
- 如需進階自訂，請將安全且受管的機器學習環境的佈建標準化並自動化，以支援分散式團隊的需求。如需詳細資訊，請參閱在 上設定安全且受管良好的機器學習環境 AWS (AWS 部落格文章）。

- 採用機器學習操作 (MLOps) 來自動化和標準化整個機器學習生命週期的程序。這些程序包括模型開發、測試、整合、發行和基礎設施管理。如需詳細資訊，請參閱什麼是 MLOps？
- 對於小規模實驗或概念驗證，請從 Amazon SageMaker AI 和一般用途運算執行個體開始。當您擴展至大型生產部署時，請考慮使用 Amazon EC2 加速運算執行個體，以獲得最高效能。
- 與隨需執行個體相比，在 SageMaker AI 中使用受管 Spot 訓練可將訓練模型的成本最佳化高達 90%。SageMaker AI 會代表您管理 Spot 中斷。

第 2 層：核准的一組基礎模型和工具

隨著組織瀏覽生成式 AI 採用的早期階段，他們很快意識到沒有任何單一模型可以有效地解決所有使用案例。不同的模型在各種網域和任務中表現卓越，企業需要平衡每個特定應用程式的功能、成本和效能。這種現實推動了對靈活但受控制的基礎模型存取方法的需求。

本節包含下列主題：

- 模型實驗和自訂
- 模型評估
- 實作建議

模型實驗和自訂

Amazon Bedrock 旨在協助您實驗各種基礎模型，並支援可擴展的生產部署。透過 Amazon Bedrock 知識庫，您擁有全受管解決方案，可建置end-to-end擷取增強生成 (RAG) 工作流程。Amazon Bedrock 也支援受管代理程式，可以執行從預訂行程到管理庫存的複雜任務，無需程式碼。由於它是無伺服器，Amazon Bedrock 可減少基礎設施管理問題，並安全地與其他 整合 AWS 服務。

此外，您可以使用 Amazon Bedrock 來使用您自己的專屬資料私下自訂基礎模型，並將其安全地提供給組織內的使用者。如需詳細資訊，請參閱《Amazon Bedrock 文件》中的自訂模型，以改善其使用案例的效能。對於存取管理， AWS Identity and Access Management (IAM) 會與 Amazon Bedrock 整合，以便您可以設定精細存取控制。控制項決定哪些使用者可以啟用和存取特定模型。如需詳細資訊，請參閱本指南中的 第 3 層： 上生成式 AI 平台的安全和管理 AWS。

模型評估

隨著組織從生成式 AI 原型進展到生產，為基礎模型建立嚴格的評估程序至關重要。雖然開放基準提供模型效能的一般洞見，但在判斷模型是否適合特定企業需求時，通常會很短。量身打造的評估策略可協助使用者選擇最適當的模型，以符合組織的獨特需求。

自訂模型評估的目標

自訂評估方法可協助組織執行下列動作：

- 評估業務相關任務的效能 – 評估模型處理與企業使用案例直接相關任務的能力。

- 識別潛在的偏差和限制 – 偵測模型可能呈現偏差或失敗的區域，以便您可以確保模型適用於實際部署。
- 比較模型與相關指標 – 使用符合您組織優先順序和目標的指標來比較不同的模型。
- 選擇明智的模型 – 對模型選擇、微調和部署到生產環境做出資料驅動型決策。

選擇模型評估指標

有效的模型評估利用各種技術，提供模型效能的整體檢視。透過這些技術，組織不僅可以評估模型的技術準確性，還可以評估其與企業特定需求的一致性。若要評估模型，您可以結合量化和定性指標來判斷效能、識別偏差，然後選擇模型。組織應同時使用基本實境指標（含參考資料）和彈性指標（不含參考資料），以全面了解模型適用性。在 AI 中，地面事實是指在模型訓練期間保留的事實資料，讓您可以將其用於模型評估。

如果參考資料存在，請使用以地面事實為基礎的指標。這些指標提供具體的量化評估。同時，沒有基本事實的技術可以提供靈活性。這些技術可協助您在可讀性和完整性等維度上評估模型，即使沒有預先定義的正確答案也一樣。

以 Ground Truth 資料為基礎的指標

如果有參考資料，則以 Ground Truth 為基礎的指標可以提供量化評估。下列指標提供量化效能評估：

- ROUGE-L 分數 – 針對最長常見子序列 (LCS) 的裝訂評估 (ROUGE) 的召回導向研究，也稱為 ROUGE-L，會測量產生和參考文字之間的最長常見子序列。指標會評估一致性和內容重疊。
- 餘弦相似性 – 此指標會評估文字之間的語意相似性。它提供對模型情境理解的洞察。
- METEOR 分數 – 以明確順序 (METEOR) 評分評估翻譯的指標結合了單字對齊和語意比對，以提供內容準確性和意義的平衡評估。
- 二進位相似性 – 此指標會檢查是否完全相符，因此對於需要精確輸出的任務特別有用，例如命令產生。
- LLM-as-a-judge 分數 – 此指標使用另一個大型語言模型 (LLM) 在定義的規模上評定相似性。它提供精細的品質評估。

Amazon SageMaker Clarify 和 Amazon Bedrock 包含可協助您評估模型的功能。他們可以自動化模型評估任務，讓您可以量化模型風險和回應品質。如需詳細資訊，請參閱評估、解釋和偵測模型中的偏差，以及評估 Amazon Bedrock 資源的效能。

沒有 Ground Truth 資料的指標

當參考資料不可用或作為補充方法時，您可以使用 LLM-as-a-judge 技術。它使用單獨的模型，根據對業務重要的維度來評估生成式 AI 解決方案的輸出。此技術提供評估各種模型品質的彈性。如需詳細資訊，請參閱《Amazon Bedrock 文件》中的使用另一個 LLM 做為判斷器評估模型效能。

您可以使用計算指標來評估檢索增強生成 (RAG) 系統從資料來源擷取相關資訊的有效性，以及產生的回應在回答問題方面的效果。RAG 評估的結果可讓您比較不同的 Amazon Bedrock 知識庫和其他 RAG 來源，然後為應用程式選擇最佳的知識庫或 RAG 系統。如需詳細資訊，請參閱 Amazon Bedrock 文件中的評估 RAG 來源的效能。

實作模型評估技術

為了有效符合具有模型功能的企業需求，請使用一組強大的評估條件來引導模型評估程序。這些條件涵蓋一系列優先順序，從正確性和完整性到事實和敏感資料處理。每個條件都符合特定技術，以提供可行的洞見。例如，當事實準確性至關重要時，ROUGE-L 分數或餘弦相似性等指標可提供具體、可量化的基準。相反地，使用 LLM-as-a-judge 等技術來評估可讀性和新鮮度。這些技術提供針對組織標準量身打造的靈活、定性洞見。評估模型的關鍵維度包括：

- 正確性 – 驗證所提供資訊的準確性
- 完整性 – 驗證回應的深度和涵蓋範圍
- 可讀性 – 評估清晰度和理解程度
- 資訊的新鮮度 – 檢查內容是否相關且為最新版本
- 敏感資料抑制 – 檢查是否適當處理機密資訊
- 準確性 – 測量與事實資訊的一致性
- 一致性 – 檢查邏輯流程和一致性
- 事實 – 驗證內容的真實性
- 全面性 – 評估涵蓋範圍和全面性

透過在這些明確準則中錨定模型評估技術，您可以針對其特定目的徹底審核模型。這種結構化方法使模型符合企業目標、合規要求和使用者期望。它還為在生產環境中大規模部署生成式 AI 應用程式奠定了堅實的基礎。

實作建議

若要建立有效的基礎模型策略，請考慮下列建議：

- 組成具有明確角色、責任和決策程序的模型控管委員會。

- 在基礎模型可用於整個組織之前，開發評估標準和評分機制來評估基礎模型。
- 請記住，最大的基礎模型不一定是您使用案例的最佳模型。使用最上層模型開始proof-of-concept開發，以驗證商業價值，然後系統性地評估較小的模型以進行成本最佳化。
- 開發儀表板來追蹤關鍵指標，例如推論延遲、輸送量、錯誤率和每個推論的成本。
- 為團隊提供有關如何為其使用案例選擇正確模型的明確指導，包括實驗程序和評估標準。

第 3 層： 上生成式 AI 平台的安全和管理 AWS

強大的安全性和管理架構對於擴展整個企業的生成式 AI 採用至關重要。我們建議採用以平台為中心的方法。此方法可確保所有生成式 AI 驅動的應用程式，無論團隊建置哪些應用程式，都能受益於一組預設的安全和負責任的 AI 護欄。

本節包含下列主題：

- 核心安全紀律
- 建議的安全控制
- 安全性範圍矩陣
- 實作建議

核心安全紀律

LLM 應用程式的 OWASP 前 10 名描述了生成式 AI 工作負載的其他安全考量。不過，許多傳統安全紀律仍然至關重要。生成式 AI 工作負載的安全考量包括下列項目：

- 控管與合規 – 建立必要的政策、程序和報告，以強化業務，同時將風險降至最低。對於生成式 AI 應用程式，這包含模型選擇、資料用量和輸出驗證的指導方針。
- 法律和隱私權 – 符合使用或建立生成式 AI 解決方案的特定法規、法律和隱私權要求。在實作生成式 AI 解決方案時，組織必須仔細考慮資料保護法律、智慧財產權和產業特定法規。
- 風險管理 – 識別生成式 AI 解決方案的潛在威脅和建議的緩解措施。這包括解決資料中毒、提示注入攻擊或模型輸出中非預期偏差等風險。
- 控制 – 實作可降低風險的安全控制。對於生成式 AI 應用程式，這包括輸入淨化、輸出篩選和模型使用的存取控制。如需詳細資訊，請參閱本指南中的建議安全控制。
- 彈性 – 架構師生成式 AI 解決方案，以維持可用性並符合商業服務水準協議 (SLAs)。這包括在不同負載條件下對模型備援、備用機制和可擴展性的考量。

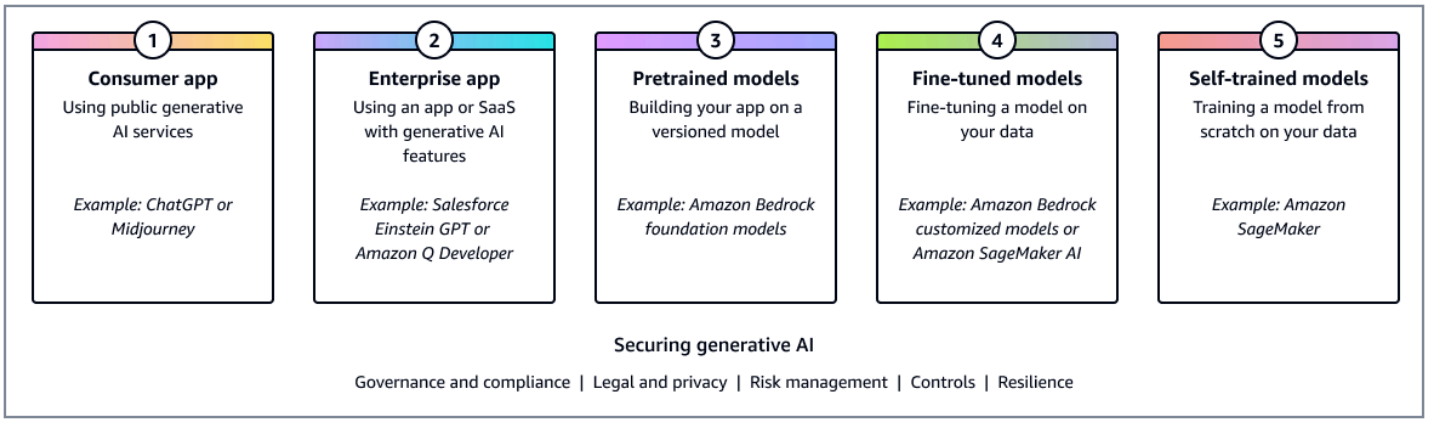
建議的安全控制

Defense-in-depth是一種網路安全方法，使用多層安全控制來保護系統、網路或工作負載。不同的 layer 有助於在某個 layer 失敗時偵測和停止攻擊。我們建議至少採用下列控制項，為生成式 AI 工作負載及其環境採用defense-in-depth方法：

- 身分和存取管理 – AWS 服務，包括 Amazon Bedrock 和 Amazon SageMaker AI，原生整合 AWS Identity and Access Management (IAM)。IAM 可讓您精細控制誰可以對 AWS 帳戶 和 資源執行動作，例如訂閱基礎模型或在模型上執行推論。我們建議您建立具有精細許可的多個 IAM 角色。例如，您可以建立下列角色：
  - 評估角色 – 當新的模型新增至 Amazon Bedrock 時，具有此角色的使用者可以在沙盒環境中評估基礎模型。這些使用者應先與您的法務和採購團隊協調，再啟用允許更廣泛存取的訂閱。
  - 一般存取角色 – 具有此角色的使用者可以存取標準使用的基礎模型。
  - 微調模型存取角色 – 具有此角色的使用者可存取使用專屬資料微調的模型。
  - 專用模型存取角色 – 具有此角色的使用者可以存取特定使用案例的先進、高成本模型。
- 透過 進行私有網路存取 AWS PrivateLink – AWS PrivateLink搭配 Amazon Bedrock 和 Amazon SageMaker AI 使用，從您的 VPC 和內部部署網路調用模型。這有助於將敏感資料保留在您的私有網路中。
- Amazon Bedrock 護欄 – 實作護欄來管理和篩選對基礎模型的請求和回應。這會新增額外的控制層。
- 安全地存放叫用日誌 – 為確保 的使用符合您的合規需求，您可以將所有請求和回應儲存到 Amazon S3 儲存貯體或 Amazon CloudWatch Logs。如需詳細資訊，請參閱《Amazon Bedrock 文件》中的使用 CloudWatch Logs 和 Amazon S3 監控模型調用Amazon S3。
- 稽核和追蹤對基礎模型的存取 – 使用 Amazon CloudWatch，這是一種記錄使用者、角色或 所採取動作的服務 AWS 服務。CloudTrail 會將 Amazon Bedrock 的所有 API 呼叫擷取為事件。如需詳細資訊，請參閱使用 CloudTrail 監控 Amazon Bedrock API 呼叫和使用 記錄 Amazon SageMaker AI API 呼叫 AWS CloudTrail。
- 針對 LLM 安全性，請遵循 OWASP 前 10 項 – 遵守 OWASP 對 LLM 應用程式前 10 項安全風險，並確保您已清楚識別策略和安全控制，以架構分層防禦。如需詳細資訊，請參閱使用適用於 LLMs 的 OWASP 前 10 大生成式 AI 應用程式的 Architect defense-in-depth安全性 (AWS 部落格文章）。

安全性範圍矩陣

您需要的特定安全控制取決於生成式 AI 應用程式的性質。AWS 生成式 AI 安全範圍矩陣 (AWS 部落格文章）提供了一個架構，以根據您的部署範圍了解安全需求。下圖顯示生成式 AI 安全性範圍矩陣，這是一種用於分類使用案例的心理模型。



此矩陣有助於組織根據下列因素識別適當的安全控制：

- 應用程式是面向消費者還是企業內部
- 相較於微調，使用預先訓練的模型
- 正在處理的資料的敏感度
- 應用程式對業務營運的重要性

實作建議

若要為您的生成式 AI 工作負載實作全面的安全和控管架構，請考慮下列建議：

- 為模型存取和使用建立明確的政策和程序。
- 在多層實作安全控制，例如網路、應用程式和資料層。
- 執行定期安全評估和合規稽核。
- 維護安全措施和控制的詳細文件。
- 為使用生成式 AI 的團隊提供安全意識培訓。
- 隨著威脅的演進，定期審查和更新安全控制。

第 4 層：常見生成式 AI 使用案例的可重複應用程式模式

隨著企業採用生成式 AI，組織如何使用此技術來推動商業價值出現了明確的模式。這些常用案例示範了在不同業務職能之間套用生成式 AI 的成熟方法。這些使用案例可協助組織識別自己的實作機會，並排定其優先順序。

透過了解這些已建立的使用案例，組織可以更好地評估生成式 AI 可以提供最立即的影響，同時為更廣泛的採用打下基礎。本節說明企業成功部署生成式 AI 解決方案的關鍵領域。

本節包含下列主題：

- 生成式 AI 輔助
- RAG 聊天式助理
- 生成式 AI 代理器
- 智慧文件處理
- 內容產生和摘要
- 程式碼產生和分析
- 多模態 AI 應用程式
- 實作建議

生成式 AI 輔助

採用 AI 技術的生成助理可跨各種業務職能提供個人化的內容感知支援。這些工具非常適合進階使用者。他們加強了那些學習將他們整合到工作流程的員工的輸出。透過使用語言模型，這些助理可增強組織中的效率、生產力和使用者滿意度。

AWS 提供 Amazon Q Business，這是一種企業級 AI 助理，可隨時使用且高度自訂。實作涉及使用組織特定資料來量身打造其知識庫，以推動準確、相關的回應。與客戶關係管理 (CRM) 和企業資源規劃 (ERP) 系統整合可增強其公用程式，並提升整個企業的順暢資料流程和完整功能。

對於軟體開發團隊， AWS 提供 Amazon Q Developer。此編碼助理旨在透過協助程式碼產生、錯誤偵測、重構建議和文件建立來提高生產力。當無縫整合到整合開發環境 (IDEs) 時，開發人員可以直接在其平常的編碼環境中存取其功能，從而簡化工作流程並減少內容切換。如需詳細資訊，請參閱此策略文件中程式碼產生和分析的 。

生成式 AI 支援助理的常見使用案例包括下列項目：

- 支援排程管理、報告摘要和簡短準備的執行助理。
- 支援員工加入、回答政策相關問題，以及協助管理效能的人力資源助理。
- 銷售啟用工具，可提供即時產品資訊和競爭洞察。

為了最大限度地提高生成式 AI 支援助理的有效性，請考慮從目標部署開始，例如在特定部門內。這有助於開發團隊評估影響並收集使用者意見回饋，然後再在整個組織中更廣泛地推出。這種分階段方法可簡化採用程序，並最佳化助理與企業需求的一致性。

如需建置企業生成式 AI 輔助程式的詳細資訊，請參閱使用 Amazon Bedrock 代理程式和知識庫開發全自動化聊天輔助程式。

RAG 聊天式助理

Out-of-the-box的基礎模型雖然功能強大，但受到預先訓練資料的限制。他們缺乏訓練截止日期後最近事件的知識，也無法了解公司特定的程序或自訂資料。此限制使它們不適合需要組織知識的許多商業使用案例。

透過微調自訂模型是一個複雜且昂貴的程序，因此收集高品質的訓練資料可能是一項挑戰。擷取增強生成 (RAG) 透過結合大型語言模型 (LLMs) 與組織的專屬資料來解決這些限制。RAG 方法會建立由特定領域知識告知的 AI 互動。如需 RAG 和微調的詳細資訊，請參閱查詢自訂文件的生成式 AI 選項。

RAG 聊天式助理的常見使用案例包括下列項目：

- 可存取產品文件和FAQs的客戶支援系統
- 可快速擷取公司特定資訊的內部知識管理工具
- 參考組織政策和法規的法務或合規聊天式助理

AWS 透過 Amazon Bedrock 知識庫簡化 RAG 聊天式助理的實作。此服務可協助您設定向量資料庫，從自訂文件查詢和擷取資料。主要功能包括與 Amazon Simple Storage Service (Amazon S3) 整合以進行資料儲存、與第三方服務（例如 Microsoft SharePoint 和 Atlassian Confluence) 的相容性、相關文件擷取的高效向量搜尋功能，以及與基礎模型（例如 Anthropic Claude) 整合以建置對話式聊天式助理。若要開始，請識別定義良好的有限範圍知識庫，並使用它來建立proof-of-conceptRAG 聊天式助理。隨著實作擴展，請注意資料新鮮度和大規模擷取操作的運算成本。

如需建置 RAG 聊天式助理的詳細資訊，請參閱下列內容：

- 在 上擷取增強產生選項和架構 AWS

- AWS 使用 Terraform 和 Amazon Bedrock 在 上部署 RAG 使用案例
- 使用 RAG 和 ReAct 提示，開發進階生成式 AI 聊天式助理

生成式 AI 代理器

生成式 AI 代理器代表自動化複雜業務流程的重大進展。這些代理程式可以透過分解任務並與各種系統互動，來處理傳統上需要人工介入的複雜工作流程。

生成式 AI 代理程式的典型應用程式包括多步驟業務流程的自動化工作流程協同運作、跨部門任務完成的智慧型虛擬助理，以及可查詢多個資料庫的資料分析和報告產生代理程式。

AWS 提供工具，例如 Amazon Bedrock 代理程式，協助您建立和部署 AI 代理程式。實作涉及定義專屬於您業務需求的自訂動作和 API 整合。您也需要設定強大的錯誤處理和記錄機制。您可以使用 AWS Lambda 執行無伺服器函數、Amazon Amazon API Gateway 管理 APIs，以及AWS Step Functions工作流程協同運作。

組織可以從識別重複的多步驟程序開始，並映射其決策點和系統互動。這可做為初始生成式 AI 代理器的藍圖。重要考量包括驗證適當的錯誤處理，以及維護決策過程中的透明度。

智慧文件處理

智慧文件處理 (IDP) 使用生成式 AI 和自然語言處理 (NLP) 自動擷取、分類和處理各種文件類型的資訊。IDP 可以將非結構化文件處理轉換為自動化、可擴展的工作流程。

現代 IDP 解決方案可以採取兩種方法：傳統的管道型處理，或透過基礎模型直接進行多模式處理。若要在 上建置傳統管道 AWS，您可以結合 Amazon Textract 進行光學字元辨識 (OCR)，並將了解文件與 Amazon Bedrock 基礎模型進行進階分析。或者，對於透過基礎模型的直接多模態處理，您可以使用state-of-the-art基礎模型 （ 例如 Amazon Bedrock 中的 Anthropic Claude)，以單一步驟直接處理影像、PDFs和掃描的文件。

組織可以根據其特定需求選擇其方法。傳統管道在需要精細控制的高容量處理和情境中表現卓越。透過基礎模型進行直接多模式處理，非常適合需要透過複雜情境理解提高準確性的複雜文件。

IDP 的常見使用案例包括下列項目：

- 可自動化保險理賠文件檢閱、驗證和資料擷取的理賠處理。
- 發票處理會擷取應付帳戶自動化的明細項目、總計和廠商資訊。
- 合約分析會檢閱法律文件，以擷取關鍵術語、義務和過期日期。

- 分析患者記錄以擷取相關醫療資訊和編碼的醫療記錄處理。

實作通常會使用使用 Amazon S3 AWS Lambda和 Amazon Simple Queue Service (Amazon SQS) 的事件驅動架構。當您將文件上傳至 Amazon S3 儲存貯體時，它會啟動處理工作流程。結果會存放在資料庫，例如 Amazon DynamoDB，或透過 APIs 擷取至業務系統。為了獲得最佳結果，組織應該實作批次處理策略、強大的錯誤處理和稽核線索。展開範圍之前，請先從特定文件類型開始。

此自動化可大幅減少手動介入、處理時間和人為錯誤。它還有助於員工專注於需要判斷和專業知識的活動。

如需文件處理工作流程的詳細資訊，請參閱使用 Amazon Textract 從 PDF 檔案自動擷取內容。

內容產生和摘要

生成式 AI 模型擅長產生和摘要內容。這使它們成為各種商業應用程式的重要工具。您可以使用大型語言模型的自然語言處理功能來建立或精簡文字型資訊。

內容產生和摘要的常見使用案例包括下列項目：

- 從結構化資料自動產生報告
- 建立行銷資料、產品描述和社交媒體文章的內容
- 研究論文、法律文件或冗長報告的文件摘要

在 上的實作 AWS 通常涉及使用 Amazon Bedrock 產生文字和摘要任務。Amazon Comprehend 可以用於實體辨識和情緒分析，增強產生內容的內容和品質。

內容產生和摘要應用程式的最佳實務包括實作human-in-the-loop程序、使用微調模型進行網域特定內容產生，以及實作內容篩選條件以確保品牌一致性和遵守準則。

例如，您可以使用生成式 AI 來分析患者臨床資料，並產生摘要和臨床備註。如需詳細資訊，請參閱使用擴增患者資料建置醫療智慧應用程式。

程式碼產生和分析

生成式 AI 模型具有卓越的功能，可了解和產生各種程式設計語言的程式碼。您的開發人員可以使用這些功能來協助撰寫、重構和分析程式碼。

常見的使用案例包括下列項目：

- 樣板任務的自動程式碼產生
- 程式碼重構和最佳化建議
- 錯誤偵測和安全漏洞分析

AWS 提供多種 服務來協助支援此模式。Amazon Q Developer 提供 AI 輔助程式碼產生，協助開發人員更快撰寫程式碼，並減少錯誤。Amazon CodeGuru 提供智慧型程式碼檢閱和效能建議。您可以將這些服務與 整合 AWS Lambda ，以進行無伺服器程式碼執行和測試。

程式碼產生和分析應用程式的最佳實務包括將 AI 輔助編碼工具整合到現有的開發環境中。考慮實作嚴格的程式碼檢閱程序，以確保產生的程式碼符合品質標準。將 AI 用於初始程式碼草稿，但依賴人類專業知識進行最終實作決策。

如需建置編碼助理的詳細資訊，請參閱使用 Amazon Q Developer 作為編碼助理以提高您的生產力。

多模態 AI 應用程式

多模態 AI 應用程式結合不同類型的資料輸入和輸出，例如文字、影像、音訊和影片。此模式使用進階生成式 AI 模型的能力，來了解和產生跨多個模態的內容。

多模態 AI 應用程式的常見使用案例包括下列項目：

- 客戶支援或內部文件的視覺化問答系統
- 智慧型影像和影片內容管制
- 商業智慧的多模式資料分析

在 上 AWS，您可以將 Amazon Bedrock 用於多模態基礎模型。其他服務包括用於影像和影片分析的 Amazon Rekognition、用於speech-to-text轉換的 Amazon Transcribe，以及用於text-to-speechAmazon Polly。這些服務可以增強應用程式的多模式功能。

多模態 AI 應用程式的最佳實務包括確保用於訓練和微調多模態模型的高品質、多樣化資料集。此外，針對不同的輸入模式實作強大的錯誤處理，並仔細考慮多模式 AI 系統中的道德影響和潛在偏差。

如需建置多模式應用程式的詳細資訊，請參閱使用 Amazon Bedrock 和 Amazon Transcribe 從語音輸入記錄機構知識。

實作建議

若要在整個組織中有效實作生成式 AI 的可重複應用程式模式，請考慮下列建議：

- 從記錄成功實作的模式目錄開始，並提供加速開發的範本、參考架構和程式碼範例。
- 建立集中式模式控管程序，以評估、核准和維護標準模式，以提升整個企業的一致性和品質。
- 根據業務影響、技術可行性以及符合策略目標，排定使用案例的優先順序。
- 建立跨功能模式開發團隊，其中包含 AI 專家、網域專家和最終使用者，他們可以確保模式滿足真正的業務需求。
- 透過建立可組合成不同應用程式模式的模組化、可編譯元件，實作「建置一次，重複使用多次」方法。
- 為每個模式類型制定明確的評估條件，以衡量有效性。包含商業價值、使用者滿意度和技術效能的指標。
- 建立模式成熟度模型，將模式分類為實驗性、經過驗證或企業標準，以引導團隊進行選擇程序。
- 建立全面的文件，以促進每種模式的成功採用，包括實作指南、限制和最佳實務。
- 執行定期模式審查，以識別隨著技術和業務需求發展而改善、合併或淘汰的機會。

企業生成式 AI 採用和擴展的最佳實務

在整個企業中成功採用和擴展生成式 AI 需要組織結構、標準化流程和技術功能的策略平衡。下列最佳實務取自各種組織的成功實作，為整個企業提供有效採用的架構。

本節包含下列主題：

- 組織結構和控管
- 標準化和技術卓越性
- 程序實作
- 實作建議

組織結構和控管

考慮建立 AI 卓越中心和模型控管委員會。

AI 卓越中心

建立 AI 卓越中心 (AI CoE) (AWS 部落格文章），以引導整個組織的生成式 AI 計畫。AI CoE 應提供建置生成式 AI 應用程式的指引、最佳實務和技術功能。透過定期與業務單位互動，AI CoE 有助於識別採用生成式 AI 的機會，同時維持控管和品質標準。

模型控管委員會

擁有具有明確角色和責任的專用模型控管委員會至關重要。該委員會為基礎模型制定評估標準，審查用量請求，並驗證是否符合道德 AI 原則。委員會與法律和合規團隊緊密合作，監督模型效能和風險評估。這為創新和負責任的 AI 使用建立了平衡的方法。

標準化和技術卓越性

請考慮建立可重複使用的生成式 AI 模式和工具程式庫，並建立全企業存取架構，以普及對 AI 資源的存取。

模式程式庫和工具

為常見的生成式 AI 應用程式開發一套標準化工具和預先定義的模式。建立包含妥善記錄模式、程式碼範本和架構圖表的集中式儲存庫。此標準化可減少進入的障礙、改善跨實作的一致性，並透過提供明確的起點加速開發。

全企業存取架構

透過請求存取、訓練和加入的簡化程序，促進整個組織對進階工具的存取，例如 Amazon Q Business 和 Amazon Q Developer。AI 資源的普及化可讓跨部門的團隊有能力，同時維持適當的安全控制和管理。

程序實作

實作可協助您管理下列項目的程序：

- 服務
- 模型選擇和評估
- 效能監控和評估
- 安全和合規

服務管理

內部服務管理程序對於組織和控制生成式 AI 採用至關重要。這些程序應涵蓋新模型的技術評估、法律審查和存取請求。透過建立明確的工作流程和責任，組織可以維持適當的監督，同時改善生成式 AI 解決方案的部署和操作效率。

模型選擇和評估

系統化的模型選擇方法對於平衡功能、成本和效能至關重要。使用頂層模型開始proof-of-concept開發，以快速驗證商業價值並取得利益相關者的支持。在驗證使用案例之後，系統性地根據已建立的效能基準評估較小的模型，以最佳化生產成本。此方法可加速初始開發，同時提升符合成本效益的擴展。此程序需要明確傳達效能期望，並仔細記錄所需功能。

效能監控和最佳化

有效的監控和最佳化需要全方位方法來追蹤技術和業務指標。開發儀表板來追蹤關鍵指標，例如推論延遲、輸送量、錯誤率和每個推論的成本。設定異常或效能降低的提醒。定期執行效能審查，以確保模型持續符合業務需求。

成本管理應該是主動且策略性的。定期檢閱模型使用模式和運算資源最佳化，有助於維持高效率的操作。實作成本分配標籤和預算監控，以維持不同使用案例和業務單位費用的可見性。

安全和合規

必須在生成式 AI 實作生命週期中嵌入安全和合規考量。開發全面的風險管理架構，以解決資料隱私權、模型安全性和道德 AI 考量。此架構應與現有的企業安全政策保持一致，並解決生成式 AI 應用程式的獨特挑戰。

透過分層方法實作安全控制，從強大的存取管理和身分驗證開始。確保採取適當的網路安全和資料保護措施，包括全面的稽核記錄和監控功能。建立生成式 AI 應用程式特有的明確事件回應程序。

實作建議

若要在整個企業成功採用生成式 AI，請考慮下列重要建議：

- 從定義明確的有限範圍專案開始，這些專案可以展現明確的商業價值。
- 記錄成功指標和學習，以通知未來的專案。
- 逐步擴展成功的實作，並確保有適當的控制和支援機制。
- 透過定期審查流程和程序，持續專注於持續改進。
- 建立涵蓋技術和負責 AI 實務的完整訓練計畫。
- 為知識分享和跨團隊協作建立明確的機制。

透過仔細關注這些最佳實務和建議，組織可以為永續生成式 AI 採用奠定堅實的基礎，同時保持安全性、效率和道德考量。

後續步驟

生成式 AI 代表企業創新、最佳化操作並為客戶創造新價值的轉型機會。不過，從實驗到整個企業採用的過程需要一種策略方法，在創新與安全性、管理和卓越營運之間取得平衡。

此策略中介紹的四層架構 - 可靠的基礎設施、核准的基礎模型、強大的安全性和管理能力，以及可重複的應用程式模式 - 為組織提供建置和擴展其生成式 AI 計畫的完整方法。此架構可協助企業超越隔離的原型，以建立永續的生產就緒解決方案，提供可衡量的商業價值。

企業生成式 AI 採用的成功不僅需要技術實作。組織必須培養正確的人員、程序和技術組合。建立 AI 卓越中心、實作適當的控管結構，以及開發明確的模型評估和部署程序，與技術基礎設施本身一樣重要。

在生成式 AI 實作中，安全和控管不能是在思考之後。透過從頭開始嵌入這些考量，組織可以進行創新，同時優先保護敏感資料並符合法規要求。此策略中概述的安全控制和控管架構為負責任採用 AI 奠定了基礎，可隨著不斷變化的需求和新興威脅而發展。

此策略文件中呈現的應用程式模式示範了組織如何使用經過驗證的方法加速其生成式 AI 計畫。無論是實作生成式 AI 輔助、RAG 聊天式輔助或智慧文件處理解決方案，這些模式都為快速開發提供了起點，同時確保跨實作的一致性和品質。

當組織開始生成式 AI 旅程時，請考慮採取下列步驟：

1. 根據此策略中介紹的架構，評估您目前的 AI 和資料基礎設施準備程度。
2. 建立跨職能團隊來監督生成式 AI 計畫，包括 IT、資料科學、法務和業務單位的代表。
3. 從小型、定義明確的proof-of-concept專案開始，這些專案可以展現明確的商業價值。
4. 制定全面的生成式 AI 策略，以符合組織更廣泛的業務目標。
5. 投資持續教育和訓練，以在整個組織中建置生成式 AI 功能。

透過 AWS 服務 和 的全方位套件，以及 Amazon Bedrock 提供的堅實基礎，組織可以放心地從生成式 AI 實驗轉移到可擴展的企業就緒解決方案，這些解決方案旨在安全、符合道德且有效。當您導覽此轉型旅程時， AWS 會繼續致力於透過釋放企業中生成式 AI 的完整潛力所需的工具、專業知識和基礎設施來支援您的成功。

資源

下列資源提供額外的內容、技術詳細資訊和實作指引，以支援組織的生成式 AI 旅程。隨著 透過生成式 AI AWS 持續創新，您可以定期檢查這些資源是否有更新和新功能。

AWS 方案指引

策略資源

- 生成式 AI 工作負載評估
- 在 上採用生成式 AI 的成熟度模型 AWS
- 在 上擷取增強生成選項和架構 AWS
- 提示工程最佳實務，以避免對現代 LLMs提示注入攻擊

RAG 和聊天式助理

- 選擇 RAG 使用案例的 AWS 向量資料庫
- AWS 使用 Terraform 和 Amazon Bedrock 在 上部署 RAG 使用案例
- 使用 RAG 和 ReAct 提示，開發進階生成式 AI 聊天式助理
- 使用 Amazon Bedrock 代理程式和知識庫開發全自動化聊天式助理

編碼助理

- 使用 Amazon Q Developer 作為編碼助理，以提高您的生產力

文件處理

- 使用 Amazon Bedrock 和 Amazon Transcribe 從語音輸入記錄機構知識
- 使用 Amazon Textract 從 PDF 檔案自動擷取內容

AWS 服務文件

- Amazon Bedrock 文件
- Amazon SageMaker AI 文件

其他 AWS 資源

- 使用生成式 AI 轉換您的業務 (AWS 網站）
- 生成式 AI 創新中心 (AWS網站）
- 保護生成式 AI：生成式 AI 安全性範圍矩陣簡介 (AWS 部落格文章）
- 建立 AI/ML 卓越中心 (AWS 部落格文章）
- 使用適用於 LLMs 的 OWASP 前 10 名 （部落格文章），為生成式 AI 應用程式建構defense-in-depth安全性AWS
- 在 上設定安全且受管良好的機器學習環境 AWS (AWS 部落格文章）
- 機器學習鏡頭 (AWS Well-Architected Framework)

其他資源

- 大型語言模型應用程式的 OWASP 前 10 名 (OWASP 網站）

文件歷史紀錄

下表描述了本指南的重大變更。如果您想收到有關未來更新的通知，可以訂閱 RSS 摘要。

變更	描述	日期
初次出版	—	2025 年 6 月 26 日

本文為英文版的機器翻譯版本，如內容有任何歧義或不一致之處，概以英文版為準。