Capstone Lab: Creating a Project Plan with Generative AI

ILT-TF-100-MLGAIE-1 - Version 1.0.0

Lab overview

In this lab, you produce a comprehensive "Generative AI Implementation Playbook" containing five distinct documents:

- A Generative AI Use Case Selection document
- A Model Selection Rubric and Recommendation document
- A Performance Improvement Plan document
- A Foundation Model Evaluation Framework document
- A Deployment Checklist and Deployment Plan document

This playbook serves as a practical guide for implementing future generative AI projects in your organization, demonstrating your ability to utilize Generative AI to navigate the key stages of the implementation lifecycle.

This capstone lab guides you through the process of implementing a generative AI project following the Generative AI Implementation lifecycle. You go through each major stage of the lifecycle, using AI prompts to generate key documents and deliverables for a hypothetical generative AI application.

Objectives

By the end of this lab, you should be able to do the following:

- Define a clear use case and requirements for a generative AI implementation
- Select an appropriate foundation model based on key criteria
- Develop a plan to improve model performance through techniques like fine-tuning
- Create an evaluation framework to assess model results
- Outline a deployment strategy for putting the model into production

Icon key

Various icons are used throughout this lab to call attention to different types of instructions and notes. The following list explains the purpose for each icon:

- **Caution:** Information of special interest or importance (not so important to cause problems with the equipment or data if you miss it, but it could result in the need to repeat certain steps).
- **Expected output:** A sample output that you can use to verify the output of a command or edited file.
- **Note:** A hint, tip, or important guidance.
- **Task complete:** A conclusion or summary point in the lab.
- **Warning:** An action that is irreversible and could potentially impact the failure of a command or process (including warnings about configurations that cannot be changed after they are made).

AWS services used in this lab

Amazon Bedrock

Amazon Bedrock provides you with access to high-performing foundation models from leading AI companies through a single, easy-to-use API. As a fully-managed service, Amazon Bedrock simplifies the process of building generative AI applications.

Task 1: Create a Generative AI Use Case Selection document

In this task, you define a clear use case for implementing generative AI in your organization. You'll use Amazon Bedrock to generate key components of a Generative AI Use Case Selection document.

- **Note:** Thoroughout this lab, if you get an error like: **Too many tokens, please wait before trying again.** This indicates that the model is momentarily busy. You can wait a few seconds up to one minute and re-submit the prompt.
- **Caution:** To ensure a smooth experience, it's important to keep the Bedrock Chat open throughout the lab. Navigating away or refreshing the browser clears the chat memory, requiring you to restart the current task.

Task 1.1: Download the Lab Supporting Files

1. Choose (right-click) the link below to download a zip file containing the Lab Supporting files.

   Lab 3 Supporting files

- **Note:** This zip contains a text file with a the use-case for the ficticious company AnyOrganization. This text file is used for context to get the lab started and to create subsequent prompts and documents. In a real-life scenario, to create a ***Generative AI Implementation Playbook*** for your Organization, you need a similar file with the specifc details of your organizaton to get the process started.
- **Note:** At the end of this lab, the **Supporting Files** section gives you a full list of all the files included in this zip.

Task 1.2: Access Amazon Bedrock

1. On the AWS Management Console, in the search box, search for and choose `Amazon Bedrock`.
2. In the Amazon Bedrock console, expand the left side panel if it is not already visible, then scroll down to *Playgrounds* and choose **Chat/Text**.

- **Note:** Since we are using prompts iteratively, for this lab, you use the Chat mode to maintain memory for LLM recall between prompts. This interface provides a conversational aspect. You can issue a prompt and follow up with a clarifying request.

1. For *Mode*, select Chat.
2. Under the *Mode*, choose Select model.
3. Select **Anthropic** and then choose the **Claude 3 Sonnet**.

4. Choose Apply.

Task 1.3: Define the Problem Statement

1. In the Bedrock console, click the *Upload a file* button that looks like a paper clip under the prompt text area. Attach the `AnyOrg_UseCase.txt` that you downloaded from step 1.1.

- **Note:** AnyOrg_UseCase.txt contains challenges and opportunities that AnyOrganization is facing in regards to customer engagement and employee productivity we've discussed so far.

1. In the prompt text-area, enter the following prompt:

```
Act as a business analyst. Write a clear problem statement for
implementing generative AI in the company described in the attached
[Use Case]. The statement should identify current challenges and
potential opportunities (200 word maximum).
```

   - **Note:** The LLM remembers the AnyOrganization use case throughout the steps of this task. You can refer to the use case as `[Use Case]` in any chat prompt throughout this task.

2. Review the generated problem statement and refine it if necessary by prompting the LLM to make specific changes, for example:

```
In the previous problem statement, please change "potential biases,
and security of sensitive information." to "potential biases, and
security of our company's sensitive information." Please restate the
whole problem statement.
```

Task 1.4: Gather Relevant Requirements

1. Use the following prompt to generate a list of requirements:

```
Based on the problem statement for implementing generative AI in the
company described in the [Use Case], list 3 key requirements for the
solution. Include both functional and non-functional requirements.
```

2. Review the generated requirements and add any additional ones specific to your organization.

3. "In the Configuration pane, to the left of the response, change the temperature for the prompt to 0 for more deterministic and conservative responses. Next, re submit the prompt by copying it again to regenerate the list of functional and non-functional requirements.

   - **Note:** The temperature default is 1, which leads to more non-deterministic and creative responses. Observe how the outputs differ in terms of creativity, diversity, and potential novelty. By increasing the temperature parameter, the generative AI model produces outputs

that are more diverse and creative, as it explores a wider range of possibilities beyond the most likely or "safe" options.

Task 1.5: Align Stakeholder Expectations

1. Enter this prompt to generate stakeholder alignment strategies:

```
Provide 3 strategies for aligning stakeholder expectations when
implementing a generative AI solution in the company described in the
[Use Case]. Consider different stakeholder groups like executives, IT
teams, and end-users.
```

2. Review and adapt the strategies based on your organization's structure. For example if the model output stated one stakeholder for "IT Team" but you have several IT Teams, then you can ask the model to regenerate this response based on the specific number of teams in the organization.

Task 1.6: Identify Key Metrics

1. Use the following prompt to generate relevant metrics:

```
List and briefly explain 5 key metrics that should be considered when
evaluating the success of a generative AI implementation in the
company described in the [Use Case]. Include both technical and
business metrics.
```

2. Review the metrics and ensure they align with your organization's goals. Refine as needed.

Task 1.7: Compile the final Generative AI Use Case Selection document

1. Enter the following prompt to generate a structured Generative AI Use Case Selection document:

```
Using all the information we've generated about the generative AI
project for the company described in the [Use Case], create a
structured "Generative AI Use Case Selection" document. Include
sections for Problem Statement, Requirements, Stakeholder Alignment,
Key Metrics, Selection Criteria, and Organizational Readiness.
Summarize and organize the information we've discussed into a single
cohesive document (300 words maximum). Add text formatting markdown.
```

2. Review the generated document carefully.

3. Make any necessary edits or additions to ensure the document accurately reflects your organization's needs and context.

4. Save the final document for use in future exercises and as a template for future generative AI projects.

- **Note:** When you are copying the final `Generative AI Use Case Selection` document do not use the copy icon at the top-right of the output window, since it copies the whole conversation from the beginning of the chat. We only want to save the final document, so only copy the last response.

Task complete: You successfully created a Generative AI Use Case Selection document.

Task 2: Create a Model Selection Rubric and Recommendation document

In this task, you create a Model Selection Rubric and Recommendation document, using Amazon Bedrock to generate key components of the document.

In this task, you create a Model Selection Rubric and Recommendation document by using Amazon Bedrock to generate key components of the Model Selection Rubric and Recommendation document.

Task 2.1: Identify Key Selection Factors

1. Refresh your browser tab to delete the conversation history from the previous task.
2. Re-upload the `AnyOrg_UseCase.txt` file as you did in Task 1.3.
3. In the prompt text area, enter the following prompt:

```
List and briefly explain 5 key factors to consider when selecting a
pre-trained foundation model for a generative AI project. Focus on
general criteria that would apply across different use cases, but
highlight any factor that you think would apply given the attached
[Use case].
```

4. Review the generated list of factors and refine it if necessary by prompting the LLM to make specific changes and to restate the whole problem statement with your changes.

Task 2.2: Evaluate Trade-offs

1. Use the following prompt to explore trade-offs:

```
Based on the key factors for selecting a foundation model that we just
discussed, explain how you would evaluate the trade-offs between
different model options. Provide 5 examples of potential trade-offs
and how they might impact the decision-making process.
```

2. Review the generated analysis and consider how these trade-offs apply to your specific use case.

Task 2.3: Assess Model Capabilities

1. Enter this prompt to generate a capability assessment process:

```
Describe a step-by-step process for assessing a foundation model's
capabilities against specific use case requirements. Include methods
for testing and evaluating model performance on relevant tasks.
```

2. Review the generated process and adapt it as needed for your use case.

Task 2.4: Compile the final Model Selection Rubric

1. Use the following prompt to generate a comprehensive rubric:

```
Based on all the information we've discussed about selecting
foundation models, create a detailed model selection rubric. The
rubric should include categories for evaluation, specific criteria
within each category, and a scoring system. Ensure the rubric is
comprehensive and can be used to compare different foundation model
options objectively. Add text formatting markdown.
```

2. Review the generated rubric carefully.
3. Make any necessary edits or additions to ensure the rubric aligns with your specific use case and organizational needs.
4. Save the final rubric for use in evaluating model options.

- **Note:** Copy only the last response to save the Rubric and Recommendation document.

Task complete: You successfully created a Model Selection Rubric and Recommendation document.

Task 3: Create a Performance Improvement Plan document

In this task, you create a Performance Improvement Plan document, using Amazon Bedrock to generate key components of the document.

Task 3.1: Explore Performance Improvement Techniques

1. Refresh your browser tab to delete the conversation history from the previous task.
2. Re-upload the `AnyOrg_UseCase.txt` file as you did in Task 1.3.
3. In the Bedrock console, enter the following prompt:

```
List and briefly explain 5 key techniques that can be used to improve
the performance of a pre-trained foundation model given the attached
[Use case].
```

4. Review the generated techniques. In a real-world scenario, you would make notes on the ones most relevant to your use organization.

Task 3.2: Plan for Fine-tuning

1. Use this prompt to generate a fine-tuning plan:

```
Provide a concise, step-by-step plan for fine-tuning the selected
foundation model on our domain-specific data to improve performance
for [Use Case]. Include details on data preparation, training process,
and potential challenges.
```

2. Review and refine the generated plan, adding any specific details or considerations for your use case. Since this is a ficticious example for AnyOrganization, you can leave it as-is. Or you can ask the model to regenerate the output with changes. In a real-world scenario, you can ask the model to update the output with specific changes that apply to your organization.

Task 3.3: Implement Retrieval Augmented Generation (RAG)

1. Enter the following prompt to learn about RAG:

```
Explain the concept of Retrieval Augmented Generation (RAG) and how it
can be applied to enhance the performance of a generative AI model for
[Use Case]. Provide specific steps for implementing RAG with the
selected foundation model. (200 words maximum)
```

2. Review the generated information to assess the posibility of implementing RAG for your use case.

Task 3.4: Compile the final Performance Improvement Plan document

1. Enter the following prompt to generate a structured plan:

```
Using all the information we've generated about improving the
performance of the selected foundation model for [Use Case], create a
concise "Performance Improvement Plan" document. Include sections for
Fine-tuning, Retrieval Augmented Generation, Prompt Engineering, and
any other relevant techniques. Summarize the key steps,
considerations, and potential challenges for each technique. Add text
formatting markdown.
```

2. Review the generated text carefully, making any necessary edits.

3. Save the text for the final Performance Improvement Plan document for use in subsequent exercises and as a reference for future generative AI projects.

Task complete: You successfully created a Performance Improvement Plan document.

Task 4: Create a Foundation Model Evaluation Framework document

In this task, you create a Foundation Model Evaluation Framework document, using Amazon Bedrock to generate key components of the document.

Task 4.1: Define Evaluation Metrics

1. Refresh your browser tab to delete the conversation history from the previous task.
2. Re-upload the `AnyOrg_UseCase.txt` file as you did in Task 1.3.
3. In the Bedrock console, enter the following prompt:

```
Suggest 5 key metrics that should be used to evaluate the performance
of a generative AI model for the attached [Use Case]. Briefly explain
the importance of each metric.
```

4. Review the generated metrics and refine the list based on your requirements.

Task 4.2: Design a Test Dataset

1. Use this prompt to generate guidelines for test dataset design:

```
Provide recommendations for designing a representative test dataset to
evaluate a generative AI model for [Use Case]. Consider factors like
dataset size, diversity, edge cases, and real-world scenarios.
```

2. Based on the recommendations, outline a plan for creating or acquiring a suitable test dataset.

Task 4.3: Plan Automated Evaluation

1. Enter the following prompt to generate an approach for automated evaluation:

```
Describe a process for conducting automated evaluation of a generative
AI model's outputs for [Use Case]. Include details on metrics to
track, tools or frameworks that can be used, and how to interpret the
results.
```

2. Review and adapt the generated process based on your technical requirements.

Task 4.4: Plan for Continuous Monitoring

1. Enter the following prompt to generate a continuous monitoring strategy:

```
Provide a strategy for continuously monitoring and evaluating a
generative AI model's performance in a production environment for [Use
Case]. Consider aspects like real-time monitoring, user feedback
collection, and periodic reassessment.
```

2. Review and refine the generated strategy based on your organization's needs.

Task 4.5: Compile the final Foundation Model Evaluation Framework document

1. Use this prompt to generate a structured evaluation plan document:

```
Using the information we've generated about evaluating a generative AI
model for [Use Case], create a comprehensive "Generative AI Model
Evaluation Plan" document. Include sections for Evaluation Metrics,
Test Dataset Design, Automated Evaluation Process, Human Evaluation
Methodology, and Continuous Monitoring Strategy. Summarize and
organize the information into a cohesive document. Add text formatting
markdown.
```

2. Review the generated document carefully.
3. Make any necessary edits or additions to ensure the plan accurately reflects your evaluation needs and requirements.
4. Save the final Foundation Model Evaluation Framework document for future reference and implementation.

Task complete: You successfully created a Foundation Model Evaluation Framework document.

Task 5: Create a Deployment Checklist and Deployment Plan document

In this task, you create a Deployment Checklist and Deployment Plan document, using Amazon Bedrock to generate key components of the document.

Task 5.1: Identify Key Deployment Considerations

1. Refresh your browser tab to delete the conversation history from the previous task.
2. Re-upload the `AnyOrg_UseCase.txt` file as you did in Task 1.3.
3. In the Bedrock console, enter the following prompt:

```
Act as a DevOps engineer. List and briefly explain 5 key
considerations for deploying a generative AI model in production.
Include aspects related to infrastructure, scalability, security, and
monitoring. Highlight special considerations that you think would
apply given the attached [Use case].
```

4. Review the generated list and refine it if necessary.

Task 5.2: Address Scalability and Performance Requirements

1. Use the following prompt to generate scalability and performance strategies:

```
Based on the key considerations for deploying a generative AI model,
provide 5 specific strategies for addressing scalability and
performance requirements. Include recommendations for infrastructure
choices, load balancing, and caching mechanisms.
```

2. Review the generated strategies and adapt them to your organization's specific needs. As stated before, since this is a ficticious example for AnyOrganization, you can leave it as-is, or you can ask the model to regenerate the output with any changes you like. In a real-world scenario, you can ask the model to update the output with specific changes that apply to your organization.

Task 5.3: Outline Security and Privacy Measures

1. Enter this prompt to generate security and privacy recommendations:

```
Considering the deployment of a generative AI model, outline essential
security and privacy measures. Include recommendations for data
protection, access control, encryption, and compliance with relevant
regulations.
```

2. Review the security and privacy measures, ensuring they align with your organization's policies and regulatory requirements.

Task 5.4: Plan for Version Control and Model Updates

1. Enter this prompt to generate version control and update strategies:

```
Outline a strategy for managing version control and updates for a
deployed generative AI model. Include considerations for testing new
versions, rolling out updates, and maintaining backward compatibility.
```

2. Review and refine the version control and update strategies to fit your development and operations processes.

Task 5.5: Compile the final Deployment Checklist and Deployment Plan document

1. Enter the following prompt to generate a structured Deployment Checklist and Deployment Plan document:

```
Using all the information we've generated about deploying a generative AI model, create a structured Deployment Checklist and Deployment Plan document. Include sections for Pre-Deployment Preparation, Deployment Process, Post-Deployment Tasks, and Ongoing Maintenance. Summarize and organize the information we've discussed into a cohesive plan that could be followed by a DevOps team. Add text formatting markdown.
```

2. Review the generated final Deployment Checklist and Deployment Plan document carefully.
3. Make any necessary edits or additions to ensure the plan accurately reflects your organization's deployment process and requirements.
4. Save the final final Deployment Checklist and Deployment Plan document for use in your actual deployment and as a template for future generative AI projects.

Task complete: You successfully created a Deployment Checklist and Deployment Plan document.

Task 6: Final Deliverable

Task 6.1: Create a Generative AI Implementation Playbook

As a final deliverable, assemble the five document created in each task 1 through 5 into a single "Generative AI Implementation Playbook" document that can serve as a guide for future projects in your organization.

Task complete: You successfully created a Generative AI Implementation Playbook document for **AnyOrganization**. Similar steps can be applied to other projects in your organization to create a Generative AI Implementation Playbook.

Supporting Files

The zip containing Supporting files that you downloaded in Task 1 has the supporting files:

- The initial use-case for the ficticious company **AnyOrganization**. This was used as the original file to get the lab started.
- A Generative AI Use Case Selection document. You should have created a similar document at the end of Task 1. However, we are adding one here for reference.
- A Model Selection Rubric and Recommendation document. You created a similar document in Task 2. Here for reference.
- A Performance Improvement Plan document. You created a similar document in Task 3. Here for reference.
- A Foundation Model Evaluation Framework document. You created a similar document in Task 4. Here for reference.
- A Deployment Checklist and Deployment Plan document. You created a similar document in Task 5. Here for reference.
- A PDF document containing the steps of the lab, which include prompts that you can use in later projects.

- A PDF document called **Amazon Q Business Reference Guide** with additional information regarding how to implement an Amazon Q Business project in your own organization.
- Supporting sample files **dog-breeds.txt** and **anycompany.txt** that can be used as sample input files when setting up and testing Amazon Q Business.

Conclusion

You successfully completed the following:

- Defined a clear use case and requirements to implement a generative AI project.
- Defined the steps to select an appropriate foundation model based on key criteria.
- Developed a plan to improve model performance through techniques like fine-tuning.
- Created an evaluation framework to assess model results.
- Outlined a deployment strategy for putting the model into production.
