Generative AI Competency Technical Controls Calibration Guide

Table of Contents

Introduction

This calibration guide is intended for AWS services path partners who are interested in the Amazon Web Services (AWS) Generative AI Competency program. This guide only covers controls under the section of "AWS Generative AI Specific Requirements" and the "Common Customer Example Requirements" are addressed in a separate guide.

The calibration guide has case study qualification, case study examples, and deep-dive for each technical control. It provides clarity on the expected level of details for requested evidence. It helps partner improve application quality and reduce cycle time during the technical validation process. Additionally, partners can use the best practices in this technical guide to improve their Generative AI service offerings.

Each control has the following FAQs:

Why is this important? This section explains the architectural importance of implementing the control for efficient data analytics operations, security, migration, and other relevant aspects.

How can you implement this? This section discusses how to implement the specific control using AWS services. Partners can also implement controls using third-party services, but they must justify how the service meets AWS standards and control requirements.

What are good example responses? This section provides examples of responses that meet the control requirements and demonstrate the depth and expertise expected in the assessment.

What are unacceptable/insufficient information responses? This section includes examples of responses that do not meet the control requirements.

Generative AI Case Study Qualification

Partners need to provide 4 qualified GenAI case studies as part of the Competency

application. All customer examples must use at least one or more AWS native generative AI services: Amazon Bedrock, Amazon SageMaker, Amazon Q, Amazon EC2, Amazon EKS, Amazon ECS, or any other related services.

The four (4) customer examples should showcase different primary AWS native Generative AI services. Use case should demonstrate partners' capability of building a wide variety of Generative AI solutions covering LLMs, data creation, generation of synthetic data, training and deploying GenAI models, monitoring GenAI models, and any other solutions types.

Customer examples migrating from other platforms to AWS with one or more AWS native GEN-AI services mentioned above is also accepted.

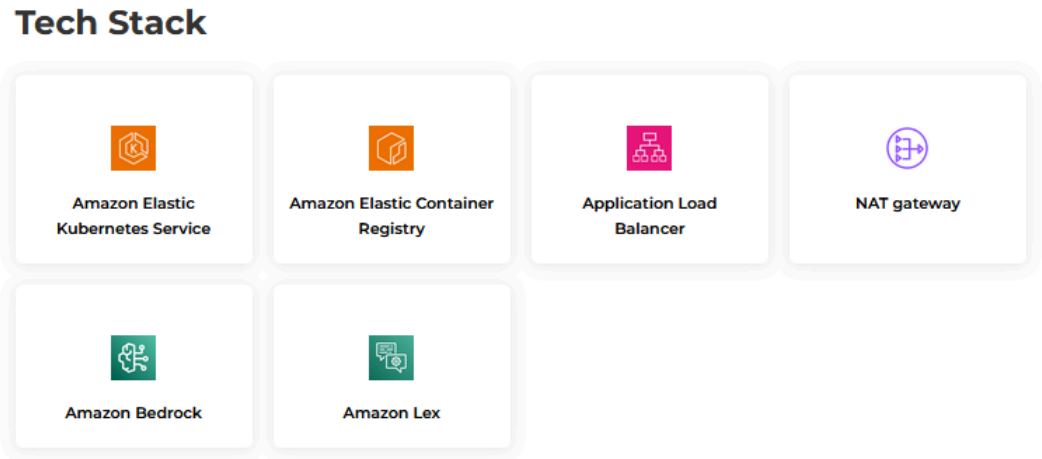
Case Study Example from Zeb

Challenge: Delivering real-time, consistent support for growing wearable user base.

As our client's user base grew, so did the volume of inquiries relate to their wearable fitness devices and associated services. Users increasingly demanded immediate assistance when encountering issues or seeking guidance on optimizing their fitness watches' usage. Maintaining uniform and accurate support across diverse channels was imperative to uphold user trust and satisfaction.

Solution: Enhanced Fitness Support: Integrated Chatbot Solutions

To address these challenges, our team proposed the implementation of a customized chatbot solution integrated seamlessly into the client's web and mobile applications. Leveraging AWS services such as EKS and Bedrock for building scalable applications, the solution included



Results Achieved

Enhanced User Experience: Integration of the chatbot into our client's applications significantly enriched the user experience by providing prompt access to support and information.

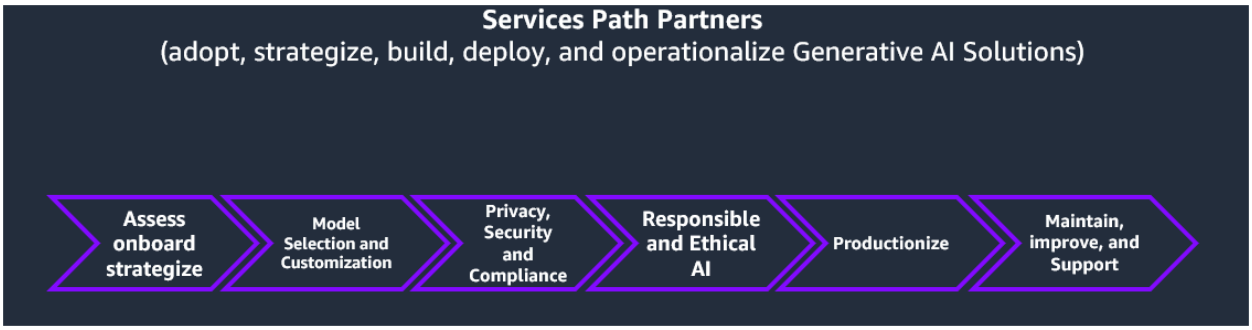
Improved Efficiency: Leveraging AWS EKS for containerized applications, our client witnessed enhanced efficiency in addressing customer inquiries, leading to reduced response times and heightened productivity.

Scalability and Reliability: AWS EKS empowered our client to seamlessly scale their chatbot infrastructure in response to fluctuating user demand while ensuring high availability and reliability.

Handling Edge Cases with Amazon Bedrock: The use of Claude allowed the chatbots to effectively handle edge cases and unique queries, providing intelligent and contextual responses even for uncommon or complex situations, enhancing the overall robustness of the system.

Generative AI Practice Requirements

The following requirements apply to AWS Services Partners' Generative AI Practice.



GENAIPR-001 - Customer Onboarding, Adoption Strategy, and Implementation Plan of Generative AI

Requirement

The AWS partner evaluates the customer's maturity for generative AI to craft a strategy that aligns with the customer's processes, skills, organizational structure, data available, industry, use cases, budget, and tolerance to risk. This includes evaluating readiness, identifying potential use cases, and defining a generative AI project roadmap.

Criteria for Passing

The partner must provide a written description of their generative AI methodology. This should include evaluations of the client's readiness and customer journey for generative AI (data infrastructure, technical skills, organizational readiness), the process for evaluating outcomes of existing generative AI projects (success metrics, lessons

learned), and details on identifying potential use cases within the client's business operations.

Why is this Important?

Effective onboarding and strategic planning are crucial in ensuring the successful adoption and implementation of generative AI. By evaluating a customer's readiness and aligning the strategy with their specific needs and capabilities, partners can maximize the impact of generative AI technologies, mitigate risks, and drive significant business value.

How Can You Implement This?

1. **Assess Customer Readiness:** Conduct thorough assessments of the customer's current data infrastructure, technical capabilities, and organizational structure to determine their readiness for adopting generative AI technologies.
2. **Develop a Tailored Strategy:** Based on the assessment, create a customized onboarding and adoption strategy that addresses specific business needs, budget constraints, and risk tolerance.
3. **Define Roadmap and Use Cases:** Identify and prioritize generative AI use cases that align with the customer's business objectives. Outline a clear roadmap for project implementation and scaling.
4. **Measure and Learn:** Implement metrics to evaluate the success of generative AI projects and use these insights to refine and improve the strategy continuously.

Additional Resources:

- AWS Generative AI Adoption Framework: [Link to framework]
- Data Governance in the Age of AI: [Link to blog]

Good Example

Partner ABC developed a comprehensive generative AI methodology for Client XYZ, which included a detailed assessment of the client's data handling capabilities and a skills audit to gauge readiness. For example, we evaluate client's existing data systems including data storage, management and processing capabilities. The methodology detailed a phased implementation plan starting with a data evaluation stage, low-risk use cases, piloting a POC before production, with success metrics defined for each phase, such as accuracy, efficiency and user engagement, and reduction in operational costs. We use these performance analysis and lessons learned to facilitate continuous improvement and inform future strategies.

Unacceptable/Insufficient Information Example

A simple, generic response such as "We assessed the client's readiness for generative AI and developed a strategy accordingly" without including its deliverables, readiness assessment, success metrics, and identified use cases, outcome and next steps is not acceptable.

GENAIPR-002 – Foundation Model Selection and Customization

Requirement

The AWS partner must be capable of selecting and customizing an appropriate Foundation Model using AWS technologies such as Amazon Bedrock, Amazon Q, SageMaker Jumpstart, or Containers. The customization should consider multiple factors including licensing, customer skills, output quality, context windows, latency, budget, and customizations supported. The partner should also facilitate the adaptation of the selected model for various downstream tasks to ensure optimal performance and strategic alignment with business needs.

Criteria for Passing

Partners provide written procedures and examples of past model selection and customization processes, demonstrating a methodical approach that considers factors like cost, latency, customization, model size, inference options, and context windows. The partner must also demonstrate knowledge of relevant metrics they use, such as language fluency, coherence, contextual understanding, factual accuracy, and ability to generate meaningful responses. The AWS Partner must also demonstrate knowledge and application of different inference options to optimize performance (could be part of OPE-005) and cost or, when needed, meet regulatory requirements.

Why is this Important?

Selecting and effectively customizing a Foundation Model are crucial for tailoring AI solutions that align with specific business needs, budget and capabilities. Customization enhances model performance, ensures relevance to the task, and optimizes resource use, thereby maximizing ROI for clients.

How Can You Implement This?

1. **Identify Foundational Model Selection Factors:** Evaluate various Foundation Models considering client-specific factors like data type, expected latency, budget, and required customizations.

2. **Customization and Optimization:** Customize models using tools like SageMaker for fine-tuning or Amazon Bedrock for model adaptation, focusing on enhancing performance metrics such as fluency, coherence, and factual accuracy.
3. **Integration and Implementation:** Ensure the customized model integrates seamlessly with the client's existing systems, utilizing AWS technologies for deployment and monitoring.
4. **Continuous Improvement:** Establish a process for ongoing optimization of data, prompts and model parameters to keep up with evolving business needs and data.

Additional Resources:

- Prompt Engineering for Foundation Models, Fine Tune a Foundation Model
- Select the right foundation model for your startup: [Link to AWS Blog]

Good Example

Partner XYZ selects foundation model based on (1) output quality performance need, (2) context window limitation (3) latency, (4) cost and customer budget, (5) customizations needed and other licensing agreement.

For large document summarization, we use Claude XX with ## context length and Llama YY models. We have modified the Claude XX prompts by incorporating domain specific context for downstream tasks to meet the customer business requirements (Partner provide example prompts). After adding relevant metadata to each chunk, we directly pass the document text to the model and we iterate this process until the final summarization. We use AWS Bedrock Llama-2 model for Q&A with the document content with customized RAG pipeline. We do parameter tuning like temperature, token length etc. for optimized response. Full fine tuning is performed when there is availability of large corpus of data, self-contained in vocabulary and patterns that the LM must learn for the use case.

For the project of enhancing contact center call interactions, we use a fine-tuned Mistral 7B model based on its performance in language comprehension. It was customized and deployed using AWS Sagemaker using a dataset of anonymized call center transcripts. Optimizations were made to complete sentiment analysis, intent recognition and keywords spotting. We finalized these customizations and optimizations based on customer business and data requirement, budget consideration. (Partner attached additional documents as example).

Unacceptable/Insufficient Information Example

A simple, generic response such as "We customized a Foundation Model to better fit the client's needs" without detailed documentation of the model type, customization

processes, specific AWS technologies used, and the impact on performance metrics is not acceptable.

GENAIPR-003 – Privacy, Security, and Compliance

Requirement

The AWS partner provides the confidentiality, integrity, and availability practices and frameworks to safeguard customer data and the generative AI applications they build, including their certifications for relevant data privacy laws and regulations.

Criteria for Passing

Partners provide privacy-preserving mechanisms, written documentation, risk assessment reports, data anonymization procedures, generative application security practices, and proof of compliance with relevant data privacy laws and regulations.

Why is this important?

Gen-AI security and data protection is critical to customer adoption. To earn customer trust, partners need to establish a framework on how customer data is stored, processed, shared, and used by the provider of the environment that the model runs in.

How can you implement this?

1. **Leverage industry resources:** OWASP AI Security and Privacy Guide and the Artificial Intelligence Risk Management Framework (AI RMF 1.0)
2. **Identify compliance needs relevant to customer data**
3. **Establish data protection and retention framework based on data confidentiality level**
4. **Use existing AWS services (such as AWS PrivateLink) to set up private access to Amazon Bedrock**

Additional Resources:

- Securing generative AI: data, compliance, and privacy considerations: [Link to AWS Blog]
- Use AWS PrivateLink to set up private access to Amazon Bedrock [link to AWS Blog]

Good Example

For all X company's generative AI practices, we ensure the confidentiality, integrity, and availability of customer data through robust frameworks and security measures. We maintain compliance with relevant data privacy laws such as GDPR and ISO standards. For customer data, we have detailed policy on access and credential management by leveraging AWS features like IAM, VPC, encryption, authentication etc. We have established data anonymization procedures which includes redaction of PII data. We use threat monitoring tools to identify potential vulnerabilities and act quickly to mitigate any threats. We conduct regular Well Architected review of the GEN-AI workload including security and networking pillar. See documentation for details of GEN-AI data privacy policy we communicate to customers (Partner provides policy details).

Unacceptable/Insufficient Information Example

A simple response like "we have data protection and retention policy in place to ensure customer data are secure" without information specific to customer examples or policy/mechanism details is not acceptable.

GENAIPR-004 – Responsible and Ethical Generative AI Best Practices

Requirement

Partner includes in their practice the responsible and ethical use of Generative AI by prioritizing accuracy, safety, transparency, user empowerment, intellectual property/copyright infringement, and sustainability while addressing potential risks and biases in AI-generated content.

Criteria for Passing

Partners are required to provide the following related to the use of Generative AI:

- Written documentation of AI ethics policies.
- Proof of bias mitigation strategies implemented in generative AI systems.
- User consent protocols that comply with applicable legal and ethical standards.
- Transparency reports detailing the operations and decisions of generative AI systems.
- Safety measures implemented to prevent misuse of AI technologies.
- Sustainability initiatives demonstrating the eco-friendly use of generative AI.

Why is this Important?

Incorporating ethical considerations in the deployment of generative AI is crucial to building trust with users and ensuring the long-term viability of AI technologies. Responsible practices help mitigate risks related to bias, privacy breaches, and ethical misuse, thereby enhancing the reliability and credibility of AI solutions.

How Can You Implement This?

1. **Develop AI Ethics Policies:** Create clear guidelines that outline ethical considerations specific to generative AI, including how to handle data responsibly, manage biases, and ensure transparency.
2. **Implement Bias Mitigation Techniques:** Use techniques like diversified data sets, regular bias audits, and algorithmic fairness assessments to reduce bias in AI-generated outputs.
3. **Establish Transparent Reporting Mechanisms:** Provide stakeholders with transparency reports that explain decision-making processes, AI behaviors, and the impact of AI solutions.
4. **Enhance User Empowerment:** Develop protocols to obtain explicit user consent and provide users with control over their data and interactions with AI systems.
5. **Prioritize Safety and Security:** Implement safety protocols to detect and prevent potential misuse of AI technologies and ensure that AI systems are secure against cyber threats.
6. **Promote Sustainability:** Optimize resource usage and adopt energy-efficient algorithms to minimize the environmental impact of running AI systems.

Additional Resources:

- Build responsible AI applications with Guardrails for Amazon Bedrock

Good Example

Partner XYZ is committed to the responsible and ethical use of Generative AI practices. We provide customers our AI ethics policies, transparency reports on AI decisions, bias mitigation report leveraging Amazon SageMaker Clarify and model monitoring for safety measures through real-time monitoring systems. We integrate application observability with customer existing platform. Additionally, our sustainability initiatives included using optimized machine learning models to reduce computational resource usage. Our frameworks are designed to prioritize user consent, ethical data usage, and continuous evaluation of our AI systems to enhance safety and reliability. We implement safeguards tailored to customers responsible AI policies using Guardrails in Amazon Bedrock. (partner attaches a responsible and ethical Gen AI policy).

See an example from Quantiphi - Quantiphi Responsible Ai ethics and Policies -

Unacceptable/Insufficient Information Example

A simple, generic response such as "We follow best practices for ethical AI" without detailed documentation of AI ethics policies, specific bias mitigation strategies, transparency efforts, user consent protocols, safety measures, and sustainability initiatives is not acceptable.

GENAIPR-005 – Generative AI Project Production Launch

Requirement

The AWS partner must demonstrate a comprehensive approach to operationalizing Generative AI projects delivered to customers. This includes integrating various components such as application frontend, backend, foundation models, infrastructure, prompt engineering, fine-tuning, data sources, data stores, along with CI/CD automation pipelines, security, compliance, human intervention, troubleshooting, data store refreshing, model and data drift, continuous monitoring and analytics, performance optimization, and model re-training.

Criteria for Passing

Partners are required to provide technical documentation that covers:

- Best practices for maintaining freshness in data stores.
- Automated processes used to customize, serve, monitor, protect, and re-train the foundation models.
- Detailed components involved in the solution such as data management, data privacy and security, platform infrastructure, monitoring, security and compliance, orchestration and automation, model versioning, model performance evaluation, and prompt engineering management.

Why is this Important?

Operationalizing generative AI projects efficiently ensures that the solutions are scalable, reliable, and maintain the highest standards of performance and security. A well-structured production launch strategy helps in reducing the time to market, optimizing resource use, ensuring compliance, and maintaining data integrity and security.

How Can You Implement This?

1. **Finalize Framework for Gen-AI project Components:** Identify development best practices for application front end, backend, foundational model and infrastructure.
2. **Manage Prompt Engineering:** Develop practices for managing and optimizing AI prompts, ensuring they are effective and evolve with the project needs.
3. **Establish Robust CI/CD Pipelines:** Implement continuous integration and deployment pipelines to automate the deployment, updates, and scaling of generative AI applications.
4. **Incorporate Comprehensive Security Measures:** Adopt stringent security and compliance measures to protect data and AI models, including encryption, access controls, and compliance audits.
5. **Implement Continuous Monitoring, Analytics and Improvement:** Utilize tools like AWS CloudWatch to continuously monitor the performance and health of AI applications and infrastructure. Incorporate performance evaluation metrics and human validation mechanisms to improve GEN-AI model outcome.
6. **Optimize Data Management Practices:** Establish protocols for regular data store refreshing, backup, and disaster recovery to ensure data integrity and availability.

Additional Resources:

- AWS Solution - Generative AI Application Builder on AWS

Good Example

Partner XYZ provides a comprehensive technical documentation for all generative AI projects with Client ABC, detailing the CI/CD pipelines established for continuous deployment and updates. The documentation included data management strategies that ensure data freshness and integrity, extensive security protocols aligned with GDPR, and continuous performance monitoring setups using AWS CloudWatch. We also outlined their prompt engineering management processes that ensure prompt relevance and effectiveness over time. (Partner attaches Gen-AI Projects Production Launch Documentation).

Unacceptable/Insufficient Information Example

A generic response such as "We have implemented automation and monitoring for our AI projects" without specific details on the CI/CD processes, data management best practices, security measures, or how the foundation models are maintained, monitored, and optimized is not acceptable.

GENAIPR-006 – Maintenance and Support Services for GenAI-Launched Projects

Requirement

The AWS partner must provide comprehensive post-launch support plans that facilitate the effective use, maintenance, improvement and support of generative AI applications. These plans should address workload-associated risks and continuously improve model performance based on customer feedback.

Criteria for Passing

Partners are required to submit written documentation that details the maintenance and support mechanisms. This documentation must include:

- The duration of the support plan (e.g., stabilization period, production support on a monthly, quarterly, annual basis, etc.).
- Clear contact and escalation procedures for customers to report operational issues.
- Defined response times and Service Level Agreements (SLAs) based on issue severity.

Why is this Important?

Providing robust post-launch support is crucial to ensure the long-term success and sustainability of generative AI applications. Effective support plans help maintain system performance, enhance user satisfaction, and ensure that any issues are swiftly resolved, thereby minimizing disruption to business operations.

How Can You Implement This?

1. **Define Support Periods:** Establish clear timelines for different phases of support, such as immediate post-launch stabilization and ongoing maintenance, to ensure customers understand the level of support at each stage.
2. **Establish Communication Channels:** Set up dedicated channels for support, such as help desks or customer service hotlines, and define escalation paths for swift resolution of issues.
3. **Implement SLAs:** Develop and communicate clear service level agreements that outline expected response times based on the severity of reported issues, ensuring that critical problems are prioritized.
4. **Continuous Improvement:** Use customer feedback and performance data to continually refine and improve the AI models and the support services offered.
5. **Risk Management:** Implement procedures to proactively identify, assess, and mitigate risks associated with the workload of generative AI applications.

Good Example

Partner XYZ has developed a detailed support and maintenance plan for their generative AI project with Client ABC, which includes a X-month stabilization period followed by quarterly reviews. The documentation specifies a tiered support system with escalation procedures clearly outlined for each tier. SLAs are defined with response times ranging from 1 hour for critical issues to 24 hours for low-severity issues. They also provided a feedback mechanism through which they gather and analyze user feedback to continuously improve the model performance.

Unacceptable/Insufficient Information Example

A generic response such as "We provide ongoing support for our AI applications" without specific details on the length of support, contact and escalation procedures, and SLAs tailored to issue severity is not acceptable.

Resources

Visit AWS Specialization Program Guide to get overview of the competency program.
Explore AWS Specialization Program Benefits to understand partner benefits.
Visit How to build a microsite to understand on building a Practice/solution page
Check out How to build an architecture diagram to build an architecture diagrams.
Learn about Well Architected Framework on Well Architected Website

Notices

Partners are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers and partners are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers/partners.