🧠 AWS Certified AI Practitioner (AIF-C01)

Domain 2: Fundamentals of Generative AI

Task Statement 2.2: Understand the Capabilities and Limitations of Generative AI for Solving Business Problems

---

📌 OBJECTIVES

1. Describe the **advantages** of generative AI
2. Identify **disadvantages** and limitations
3. Understand **factors in model selection**
4. Determine **business value and performance metrics**

✅ 1. ADVANTAGES OF GENERATIVE AI

Generative AI offers a range of benefits to businesses, particularly in automating, personalizing, and scaling processes.

Advantage	Description
Adaptability	FMs can be quickly adapted to multiple domains with minimal fine-tuning.
Responsiveness	Real-time responses in chatbots and virtual assistants enhance user engagement.
Simplicity	Prompt-based interfaces reduce the need for complex data pipelines or coding.
Creativity	Can generate novel text, designs, code, or media assets (e.g., marketing content).
Scalability	Easily scaled to serve millions of users with consistent performance.
Efficiency	Automates repetitive tasks (e.g., summarizing reports, drafting emails).
Multimodal capabilities	Models like GPT-4 or Claude 3 can handle text, images, and more.
Personalization	Supports tailored user experiences based on input context or history.

⚠️ 2. DISADVANTAGES AND LIMITATIONS

Understanding the risks and boundaries of GenAI ensures responsible and reliable deployment.

Disadvantage	Description
Hallucinations	Models can generate false or misleading information confidently.

Disadvantage	Description
Inaccuracy	May produce outputs with factual or logical errors.
Interpretability	Hard to understand how or why a model makes a specific prediction.
Nondeterminism	Same prompt can produce different results each time due to probabilistic nature.
Bias	Models may reproduce societal, cultural, or data-based biases.
Data Sensitivity	FMs may leak information seen during training or inference if not properly governed.
Cost	Large models can be expensive to run at scale (especially image or video generation).
Context Limits	Token limits can restrict long document comprehension.
Latency	Complex models may have high response times without optimized infrastructure.

Mitigations:

- Use retrieval-augmented generation (RAG) to ground responses.
- Apply guardrails (e.g., Amazon Bedrock Guardrails).
- Apply human-in-the-loop (HITL) workflows with Amazon A2I.

🧠 3. MODEL SELECTION FACTORS

Choosing the right model is critical to success in solving specific business problems.

Factor	Explanation
Model Type	Text (Claude), image (Stable Diffusion), code (CodeWhisperer), multi-modal
Capabilities	Does the model support summarization, question answering, etc.?
Performance Requirements	Does it need low latency, high throughput, or support for large context windows?
Cost Constraints	Smaller models may suffice for simpler tasks; avoid overpaying.
Compliance	Does the model/provider support required data governance or region restrictions?
Tuning Options	Can you fine-tune the model or use parameter-efficient techniques (e.g., LoRA)?
Inference Needs	Does the solution require real-time, batch, or edge inference?
Deployment Method	SaaS (Amazon Bedrock) vs fully managed (SageMaker) vs containerized (ECS/EKS)
Security & Privacy	Consider encryption, VPC support, and data isolation.

Tip: Use **Amazon Bedrock** to experiment with multiple models behind a common API (e.g., Claude, Cohere, Titan).

💼 4. BUSINESS VALUE & METRICS

You must align generative AI solutions with measurable outcomes.

🎯 Common Business Metrics

Metric	What It Measures
Cross-Domain Performance	Ability to generalize across tasks (e.g., summarization + question answering)
Efficiency	Time saved (e.g., automated email generation, faster onboarding)
Conversion Rate	Improvement in leads → customers (e.g., via GenAI-driven chat or content)
Average Revenue per User (ARPU)	Additional revenue generated from personalized recommendations or content
Accuracy	How often outputs are factually correct or pass validation checks
Customer Lifetime Value (CLV)	Long-term value of customers due to improved experiences and personalization
Resolution Time	Time to answer customer queries (chatbots with GenAI can reduce this significantly)
Engagement Rate	Clicks, responses, or time spent on content generated by AI
Return on Investment (ROI)	Overall benefit of deploying a generative AI system vs. cost

---

🔍 EXAMPLES OF BUSINESS VALUE

Use Case	Business Value
Chatbots with GenAI	24/7 support, reduced human staffing needs, increased customer satisfaction
Document Summarization	Analysts save hours reviewing reports, increasing operational efficiency
Marketing Content Creation	Quicker campaign turnaround, A/B tested AI-generated ad copy
Product Recommendations	Higher conversion rates, increased cart size
Internal Knowledge Search (RAG)	Reduces employee search time, improves decision-making speed
Code Generation	Faster software delivery, fewer bugs, less developer fatigue

---

🛠️ AWS Services Related to GenAI Business Solutions

Capability	Service(s)

Capability	Service(s)
Model Access	Amazon Bedrock, Amazon SageMaker
Prompt Engineering	Amazon Bedrock, Amazon Q
Human Feedback	Amazon A2I (Augmented AI)
Knowledge Retrieval	Amazon Kendra, Amazon OpenSearch + RAG
Compliance & Governance	AWS Artifact, AWS Audit Manager, IAM, CloudTrail
Observability & Metrics	Amazon CloudWatch, SageMaker Model Monitor

---

🧠 Study Tips

- Pair each **advantage** with a **limiting factor** to understand trade-offs.
- Memorize key **metrics** that tie GenAI to **business outcomes**.
- Understand when **to use Bedrock** vs **SageMaker** vs **prebuilt tools** (like Q or Lex).
- Review **real-world case studies** on AWS or partner sites for context.