Implementing Generative AI

APPLICATION LIFECYCLE · FM SELECTION · EVALUATION · MONITORING · BEDROCK AGENTS · SERVICE SELECTION

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

01 **The lifecycle: define a use case → select a foundation model → improve performance → evaluate results → deploy → monitor, audit, and get feedback.** It is iterative — insights from production loop back to any earlier stage, even to redefining the use case.

02 **Service ladder: Amazon Q Business** = no-code (you never select or manage a model) → **Amazon Bedrock** = use/customize pre-trained FMs via API → **Amazon SageMaker AI** = build your own with **full control across the entire ML lifecycle**. Skills, control, complexity, and cost all rise together.

03 **Improve performance** with the course's four levers: **adjust inference parameters**, **prompt engineering**, **RAG**, and **fine-tuning / continued pre-training**. RAG costs less up front; a fine-tuned model in the same domain likely responds faster.

04 **Evaluate with a combination: human evaluation** (relevance, coherence, appropriateness) + **benchmarks** (GLUE, SuperGLUE, SQuAD) + **automated metrics** (perplexity, BLEU). Performance is a function of the model **and** the dataset — and datasets shift over time (**model drift**).

05 **Agents:** Amazon Bedrock Agents **orchestrate** interactions between the FM, data sources, applications, and user conversations — breaking tasks into steps, calling APIs, and querying knowledge bases — with no capacity to provision, no infrastructure to manage, no custom code.

02 KEY TERMS

TERM	DEFINITION
Working backwards	AWS practice: define the customer experience first, then iterate backwards to what to build and its value
Context window	How many tokens an LLM can process — counting **both** the input prompt and the output
Throughput	How many requests the application completes within a given time frame
Model drift	Degrading production performance as real-world data distributions shift away from what was tested
Benchmark	Standardized dataset (GLUE, SuperGLUE, SQuAD) giving consistent comparison points across models
Perplexity	Automated quantitative metric of model performance — fast and scalable, misses nuance
BLEU	Bilingual Evaluation Understudy — automated metric scoring generated text against references
AI agent	Autonomous or semi-autonomous software entity that performs tasks on behalf of users — an LLM core plus components that reach external tools, databases, and APIs
Amazon Bedrock Agents	Managed agents that orchestrate FMs, data sources, apps, and conversations; call APIs and knowledge bases automatically
Batch inference	Submit many prompts, receive responses asynchronously — priced below on-demand (Bedrock)
Provisioned Throughput	Purchased Bedrock model units (max tokens/minute), billed hourly — **required to run customized models**
Amazon A2I	Augmented AI — human-review workflows that auto-select random samples or low-confidence predictions

03 THE GENERATIVE AI APPLICATION LIFECYCLE

iterative — production insights feed earlier stages

01 **Define** a use case → **02** **Select** an FM → **03** **Improve** performance → **04** **Evaluate** results → **05** **Deploy** the app → **06** **Monitor**, audit, feedback

04 CHOOSING THE AWS SERVICE

skills, control, complexity, and cost rise together

SERVICE	CHOOSE IT WHEN	REMEMBER
Amazon Q Business	No-code entry point — RAG over your own documents, fast launch, low technical bar	Powered by Bedrock · you never select, evaluate, or manage models
Amazon Bedrock	Developers add gen AI to an app with pre-trained FMs — knowledge bases, fine-tuning, agents — without building or managing models	Also: SageMaker JumpStart for pre-trained models
Amazon SageMaker AI	Build your own model; fine-grained control over architecture, training, and deployment	Full ML lifecycle control · pay-as-you-go compute/storage — variable costs

05 EVALUATING FM PERFORMANCE

SageMaker Clarify + Bedrock evaluation jobs: automated and human jobs

METHOD	EXAMPLES	CHARACTER
Human evaluation	Reviewers rate outputs	Qualitative: relevance, coherence, appropriateness — invaluable but slow and costly
Benchmarks	GLUE · SuperGLUE · SQuAD	Standardized datasets; consistent comparison across models, broad task coverage
Automated metrics	Perplexity · BLEU	Efficient, scalable, quantitative — quick insight, misses language nuance

06 MONITOR → ACT

MONITOR — KEY METRICS

proactive and reactive

- Compute, network, storage utilization
- System logs — errors, security issues
- Model performance: accuracy, latency (degradation ⇒ drift)
- Bias & fairness across demographic groups
- Compliance adherence

ACT — ON WHAT YOU SEE

automated + human-driven responses

- CloudWatch automated alerts
- SageMaker Model Monitor drift detection · A/B tests
- Investigate root cause in logs, inputs, predictions
- Retrain on drift · adjust resource allocation
- A2I human review of low-confidence outputs
- Canary releases · refine thresholds

07 BEDROCK PRICING FOR A CUSTOMIZED MODEL

customized models cannot run on standard on-demand pricing

COST	HOW IT'S CHARGED
On-demand (standard)	Pay-as-you-go per input token processed + per output token generated — base models
Batch inference	Multiple prompts, asynchronous responses — priced **lower** than on-demand
Custom model training	Total number of tokens processed by the model during training
Custom model storage	Charged per month, per model
Provisioned Throughput	**Required for customized models** — hourly per model unit (max input/output tokens per minute); no-commitment or discounted longer commitments

08 EXAM TRAPS

where the knowledge check gets you

- × “Amazon Q Business makes you pick and manage the underlying model.” — **False.** Q Business is powered by Bedrock; you don't select, evaluate, or manage models at all.
- × “A model is inherently good or bad.” — Performance is a function of the model **and** the dataset it's tested on; the same model can improve on one dataset and degrade on another.
- × “Evaluation is a one-time gate before deployment.” — Evaluate **continuously** across the lifecycle; datasets evolve and deployed models drift.
- × “RAG and fine-tuning are interchangeable.” — RAG has **less upfront cost and effort**; a fine-tuned model in the same domain probably **responds faster**. Choose by trade-off.
- × “Customized Bedrock models run on on-demand pricing.” — You **must purchase Provisioned Throughput** to run a customized model, plus training (per token) and monthly storage.
- × “Agents just generate text.” — Agents break requests into steps, **call APIs into company systems**, and **query data sources** before the LLM ever writes the answer.
- × “Token counts don't matter to cost.” — Some models charge a flat fee per API call, but under token-based pricing, managing input and output sizes **is** cost management.
- × “With managed services you still provision servers and build monitoring.” — AWS handles that; Bedrock Agents need no capacity provisioning, infrastructure management, or custom code.

09 SELF-QUIZ

answers are upside-down below

Q1 List the six stages of the generative AI application lifecycle in order.

Q2 Which AWS service lets you build your own model with full control across the entire ML lifecycle?

Q3 Which AWS generative AI service is the no-code option, where you never select or manage the underlying model?

Q4 What exactly does a model's context window measure?

Q5 Name the three popular benchmarks the module lists for comparing models.

Q6 Name the two automated evaluation metrics the module lists.

Q7 What is the term for a deployed model's performance degrading as real-world data distributions shift?

Q8 Which Bedrock pricing option must you purchase to run a customized model?

Q9 Which service builds human-review workflows that auto-select random samples or low-confidence predictions?

Q10 Sample exam: a media company wants to generate personalized news articles from user preferences and browsing history. Most effective workflow?

ANSWER KEY (rotate the page)

Q1 define use case → select FM → improve performance → evaluate results → deploy → monitor, audit, feedback **Q2** Amazon SageMaker AI
Q3 Amazon Q Business **Q4** tokens the LLM can process — input prompt plus output **Q5** GLUE, SuperGLUE, SQuAD **Q6** perplexity and BLEU
Q7 model drift **Q8** Provisioned Throughput **Q9** Amazon Augmented AI (A2I) **Q10** Amazon Bedrock with a fine-tuned FM