Session 2 Handout: Generative AI Fundamentals (Domain 2, 24%)

FAST-12 AI Practitioner Certification Prep · 7formations · Instructor: Wadood Chaudhary · All examples fictional · Exam guide v1.2

Why this session matters

Domain 2 tests vocabulary and trade-offs: what foundation models are, what they're good and bad at, and what drives cost. Know the terms precisely, because distractors use near-synonyms.

Key terms

Term	Meaning
Foundation model (FM)	Large model pre-trained on broad data, adaptable to many tasks
LLM	Foundation model for language
Token	Unit of text a model reads and writes; drives cost and limits
Context window	Maximum tokens per request (input plus output)
Embedding	Numeric vector capturing meaning; used for semantic search
Transformer	Architecture using attention; basis of most LLMs
Diffusion model	Generates images by removing noise step by step
Multimodal	Handles more than one data type (text, image, audio)
Hallucination	Fluent but false or unsupported output
Temperature	Randomness of sampling; low = more consistent

Advantages and limits

Strengths: adaptability, speed, natural-language interfaces, content creation. Limits: hallucination, non-determinism, knowledge cutoff, bias inherited from data, lack of explainability, cost at scale. Treat output as a draft that needs checking.

Cost drivers

Price per input and output token × volume. Bigger models cost more and are slower. Options: smaller models, shorter prompts, batch inference, provisioned throughput for steady high volume, caching.

Worked example (fictional sandbox rates)

120 claim summaries a day, 3,000 input + 300 output tokens each, 21 workdays.

- **Model A** ($0.80 / $3.20 per M tokens): 7.56 M input = $6.05; 0.756 M output = $2.42; **≈ $8.47 a month**.

- **Model B** ($3 / $15): $22.68 + $11.34 = **≈ $34.02 a month**.

Pick by evaluation score first, then cost.

AWS services in this domain

Amazon Bedrock (managed access to FMs including Amazon Nova), SageMaker AI and JumpStart (build, train, deploy, including open models), Amazon Quick (AI assistant and analytics for business users), Kiro (AI-assisted software development).

Checklist

- ☐ I can define each key term in one sentence.
- ☐ I can list three limits of generative AI and a mitigation for each.
- ☐ I can compute monthly token cost.
- ☐ I know what temperature does.