Working with Foundation Models

MODEL SELECTION · EVALUATION & MONITORING · INFERENCE PARAMETERS · RAG · CUSTOMIZATION · BEDROCK & SAGEMAKER AI

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

01 **Three approaches: reuse** an existing FM (prompts + inference parameters), **adapt** it with your data (reference a knowledge base or customize), or **build** your own from scratch. Complexity and cost rise in that order.

02 **Inference knobs: temperature**, **top P**, **top K** — higher values ⇒ more diverse and creative output; lower values ⇒ more deterministic and focused. **Max length** and **stop sequences** cap how much the model generates.

03 **RAG** = **retrieve** relevant data from an authoritative knowledge base *outside* the LLM, **augment** the prompt with it, then **generate**. The model itself never changes — cheaper than retraining, current without retraining, but each request pays a small **latency** cost.

04 **Customization = more training: fine-tuning** (small *labeled* dataset → better at a target task) · **instruction-based fine-tuning** (instruction + expected-output pairs → responds the way you want) · **continued pre-training** (large dataset, same techniques as original training → deeper knowledge, highest cost).

05 **Size trade-off:** more **parameters** = more capable, but costlier and slower to respond. Large models for complex reasoning, multimodal work, and long-form content; small models for narrow tasks, mobile deployment, and real-time speed.

02 KEY TERMS

TERM	DEFINITION
Parameters	Internal variables learned and adjusted during training to capture patterns; the count (billions–trillions) measures model size and complexity
Context window	How many tokens the LLM can process — counting both the input prompt and the output it produces
Latency	The time between the request to and the response from an FM
Inference speed	Encompasses latency plus throughput — how many requests are processed in a given time frame
Temperature	Controls randomness of output; higher = more diverse/creative, lower = more deterministic/focused (range 0–1)
Top P (nucleus sampling)	Cumulative-probability cutoff for candidate next tokens — P is a percentage of tokens (range 0–1)
Top K	Restricts sampling to the K most probable next tokens — K is an actual number of tokens (range 1–K)
Stop sequence	A token or string that signals the model to stop generating — a word, phrase, punctuation mark, or special character
RAG	Referencing an authoritative knowledge base outside the LLM before generating a response for a prompt
Vector database	Stores embeddings created from the data source you want a RAG solution to reference
Orchestrator	Software that manages the flow between user input, the vector database, and the LLM in a RAG solution
Transfer learning	Fine-tuning a model for a new *related* task (a dog-image classifier learns cats from a smaller dataset)

03 THREE APPROACHES TO USING FMS

complexity and cost rise top to bottom

APPROACH	WHAT YOU DO	AWS SERVICES
Reuse (off the shelf)	Prompt engineering + inference parameters — least complex, least costly path to tailored outputs	Amazon Q · Bedrock · SageMaker AI

APPROACH	WHAT YOU DO	AWS SERVICES
Adapt with data	Reference a knowledge base (RAG) or customize via fine-tuning / continued pre-training	RAG: Q, Bedrock, SageMaker AI · customization: Bedrock, SageMaker AI
Build from scratch	Develop your own FM — complete control; demands deep ML skills, time, and resources	Amazon SageMaker AI

04 CHOOSING A PRE-TRAINED FM — SEVEN FACTORS

weigh relative priority per use case

FACTOR	THE QUESTION TO ASK
Capabilities	What does the model need to do? Task specialization, multimodal support, languages, context window, customization flexibility
Size & complexity	How complex are the tasks and how critical is accuracy? More parameters ⇒ more capability and more cost
Inference speed	How quickly must the model respond? Smaller = faster; hardware matters (GPU, Trainium, Inferentia)
Training data	What can you find out about it? Quality · volume · diversity · relevance · recency · bias
Ethical practices	Risk of bias or harm? Privacy, transparency, regulatory compliance (GDPR), organizational values
Costs	Licensing, flat vs. token-based pricing, compute, customization fees, scaling, operations
Implementation	What is the effort to integrate, deploy, run, and scale?

05 INFERENCE PARAMETERS

lower = deterministic · higher = creative

PARAMETER	WHAT IT CONTROLS	RANGE	REMEMBER
Temperature	Randomness of the output	0–1	0.2 factual/technical · 0.8 brainstorming
Top P	Cumulative **probability** of candidate next tokens (nucleus sampling)	0–1	P is a *percentage* of tokens · 0.1 focused Q&A, 0.9 ideas
Top K	Sampling restricted to the K most probable tokens	1–K	K is an actual *number* of tokens · 10 precise, 50 creative
Max length	Maximum tokens generated in a single response	—	≈50 for chat answers · ≈200 for summaries
Stop sequences	Strings that halt generation at a boundary or format	—	Words, phrases, punctuation, special characters

06 THE RAG PIPELINE

01 User submits input → **02** Orchestrator searches the **vector database** (embeddings of your data) → **03** Retrieved data **augments** the prompt → **04** Augmented prompt → LLM → **05** Grounded, more relevant response

07 ADAPTING A MODEL — THE COST/EFFORT SPECTRUM

know when each is the right answer

METHOD	DATA NEEDED	GOAL	COST / EFFORT
Prompt engineering	None — better prompts	Tailor outputs of the model as-is	Least
RAG	Your docs → embeddings in a vector DB; model unchanged	Ground answers in current, relevant data; reduce hallucination	Low — no retraining; adds latency
Fine-tuning	Relatively small **labeled** dataset	Perform better on the target task	Moderate — quality dataset can be hard to get
Instruction-based fine-tuning	Sample instructions + expected outputs	Follow instructions; respond in a desired way	Moderate — careful curation required

METHOD	DATA NEEDED	GOAL	COST / EFFORT
Continued pre-training	Large volume of new high-quality data; same techniques as original training	Expand knowledge without changing capabilities	Highest — compute-expensive, time-consuming

08 EXAM TRAPS

where the knowledge check gets you

- × "Raising temperature makes output more predictable." — Backwards. **Higher** temperature, top P, or top K ⇒ more diverse and creative; **lower** ⇒ more deterministic and focused.
- × "Top P and top K do the same thing." — Top P is a cumulative **probability threshold** (0–1); top K is a fixed **count** of candidate tokens (starts at 1).
- × "RAG retrains or modifies the model." — RAG never touches the model; it augments the **prompt** with retrieved context. That is why the data stays current without retraining.
- × "RAG has no downside." — The retrieve-and-augment step adds **latency** to every request.
- × "Fine-tuning and continued pre-training are interchangeable." — They differ in goal, data, and cost: fine-tuning targets a **task** with a small labeled set; continued pre-training expands **knowledge** with a large dataset using the original training techniques.
- × "A fine-tuned model gets better at everything." — It might perform **worse** on tasks outside the target.
- × "Always pick the biggest model." — Larger models cost more and respond more slowly; smaller models win for narrow tasks, mobile deployment, and real-time applications.
- × "You must build a vector store to use Amazon Bedrock Knowledge Bases." — It is a **fully managed** RAG workflow, and its "chat with your document" feature answers from a sample document with no vector store setup at all.

09 SELF-QUIZ

answers are upside-down below

Q1 Which inference parameter controls output randomness on a 0–1 range, where lower values give more focused results?

Q2 One of top P and top K is a probability threshold; the other is a token count. Which is which?

Q3 Which two inference parameters limit how much or how far a model generates?

Q4 In a RAG solution, which component manages the flow between user input, the vector database, and the LLM?

Q5 Which Amazon Bedrock feature provides a fully managed RAG workflow with citations?

Q6 Which customization technique uses a relatively small labeled dataset to improve performance on a particular task or domain?

Q7 Which customization technique uses the same learning techniques as the model's initial training to deepen knowledge without changing capabilities?

Q8 Which AWS service builds an interactive chat application over your enterprise data without coding?

Q9 Name the three general approaches to powering a generative AI application with an FM.

Q10 Sample exam: for a creative writing task needing diverse, unexpected outputs — decrease top P, reduce top K, increase temperature, or expand the context window?

ANSWER KEY (rotate the page)

Q1 temperature **Q2** top P = probability threshold; top K = number of tokens **Q3** max length and stop sequences **Q4** the orchestrator **Q5** Amazon Bedrock Knowledge Bases **Q6** fine-tuning **Q7** continued pre-training **Q8** Amazon Q Business **Q9** reuse an existing model, adapt it with data, build from scratch **Q10** increase temperature