AI Engineering 1st edition Huyen Test Bank

SKU: 9781098166304-TEST-BANK

AI Engineering

ISBN 9781098166304 | Chapter 1: Introduction to Building AI Applications with Foundation Models

Total Questions: 150

Multiple Choice

Q1.
According to the chapter, the one word that best describes AI after 2020 is:

A) alignment
B) scale ✓ Correct
C) interpretability
D) supervision

Rationale: The chapter explicitly characterizes post-2020 AI primarily by its scale.

Q2.
What is the primary process referred to as AI engineering in the chapter?

A) Building applications on top of foundation models ✓ Correct
B) Designing semiconductor hardware for model training
C) Creating only task-specific supervised models from scratch
D) Replacing software engineering with autonomous systems

Rationale: AI engineering is defined as building applications on top of readily available foundation models.

Q3.
A language model encodes statistical information about language that is used to estimate:

A) the legal ownership of text corpora
B) the emotional intent of all speakers
C) how likely a word or token is in a given context ✓ Correct
D) the grammatical rules of only one language

Rationale: The chapter defines a language model as encoding statistical information about how likely tokens are in context.

Q4.
In the chapter, the basic unit of a language model is called a:

A) parameter
B) token ✓ Correct
C) embedding
D) gradient

Rationale: The text states that the basic unit of a language model is the token.

AI Engineering 1st edition Huyen Test Bank

Q5.

The process of breaking original text into smaller units a model can process is called:

A) serialization
B) vectorization
C) tokenization ✓ Correct
D) normalization

Rationale: Tokenization is the term used for breaking text into tokens.

Q6.

The set of all tokens a model can work with is known as its:

A) context window
B) training corpus
C) vocabulary ✓ Correct
D) parameter space

Rationale: A model's vocabulary is the set of all tokens it can use.

Q7.

Which type of language model predicts missing tokens using context from both before and after the missing position?

A) Autoregressive language model
B) Masked language model ✓ Correct
C) Contrastive language model
D) Retrieval language model

Rationale: Masked language models use bidirectional context to fill in missing tokens.

Q8.

Which type of language model predicts the next token using only the preceding tokens?

A) Masked language model
B) Bidirectional encoder model
C) Autoregressive language model ✓ Correct
D) Contrastive embedding model

Rationale: Autoregressive models generate the next token based only on prior tokens.

Q9.

As presented in the chapter, BERT is best classified as a:

A) masked language model ✓ Correct
B) speech synthesis model
C) retrieval-augmented model
D) reinforcement learning agent

Rationale: BERT is given as the canonical example of a masked language model.

AI Engineering 1st edition Huyen Test Bank

Q10.

In this book, unless explicitly stated otherwise, the term "language model" refers to:

A) a multilingual translation system
B) a masked language model
C) an autoregressive model ✓ Correct
D) an embedding-only model

Rationale: The chapter notes that language model will refer to an autoregressive model unless stated otherwise.

Q11.

A model that can generate open-ended outputs is described in the chapter as:

A) discriminative
B) deterministic
C) generative ✓ Correct
D) symbolic

Rationale: Open-ended output generation is the defining characteristic of a generative model.

Q12.

The chapter suggests thinking of a language model primarily as a:

A) compression engine
B) completion machine ✓ Correct
C) search index
D) rule-based parser

Rationale: The text explicitly describes a language model as a completion machine.

Q13.

Why were language models able to grow to today's scale, according to the chapter?

A) They rely only on manually curated labels
B) They avoid using parameters
C) They can be trained with self-supervision ✓ Correct
D) They operate only on images

Rationale: The chapter attributes modern scaling of language models to self-supervision.

Q14.

What is supervision in machine learning as described in the chapter?

A) Training models using unlabeled data only
B) Training models using labeled data ✓ Correct
C) Deploying models without evaluation
D) Updating prompts without changing data

Rationale: Supervision refers to training ML algorithms using labeled examples.

AI Engineering 1st edition Huyen Test Bank

Q15.

In self-supervised learning, labels are:

A) inferred from the input data ✓ Correct
B) provided only by domain experts
C) unnecessary because there are no outputs
D) generated after deployment by users only

Rationale: Self-supervision infers labels directly from the input data.

Q16.

According to the chapter, self-supervised learning differs from unsupervised learning because self-supervised learning:

A) requires no data preprocessing
B) uses labels inferred from the input data ✓ Correct
C) can only be used for images
D) never involves prediction tasks

Rationale: The text distinguishes self-supervision by the presence of inferred labels, unlike unsupervised learning.

Q17.

In the training examples derived from "I love street food.", the marker <BOS> denotes:

A) bag of samples
B) beginning of sequence ✓ Correct
C) binary output state
D) base optimization step

Rationale: <BOS> is defined as the beginning-of-sequence marker.

Q18.

Which marker is especially important because it helps language models know when to end their responses?

A) <PAD>
B) <CLS>
C) <EOS> ✓ Correct
D) <MASK>

Rationale: The chapter highlights the end-of-sequence marker <EOS> as crucial for ending responses.

Q19.

A model's size is typically measured by its number of:

A) tokens
B) parameters ✓ Correct
C) layers of supervision
D) output classes

Rationale: The chapter states that model size is typically measured by parameter count.

AI Engineering 1st edition Huyen Test Bank

Q20.

A parameter in an ML model is best defined as:

A) a fixed prompt template used during inference
B) a variable updated through the training process ✓ Correct
C) a token reserved for sequence boundaries
D) a benchmark used to compare models

Rationale: Parameters are variables within a model that are updated during training.

Q21.

Why do larger models generally need more data, according to the chapter?

A) Because larger models have more capacity to learn and need more data to realize it ✓ Correct
B) Because larger models cannot process small datasets at all
C) Because larger models always use supervised labels
D) Because larger models reduce tokenization efficiency

Rationale: The chapter explains that larger models have greater learning capacity and need more data to maximize performance.

Q22.

Foundation models are described as an extension beyond text to include additional data:

A) licenses
B) benchmarks
C) modalities ✓ Correct
D) hyperparameters

Rationale: The chapter emphasizes that foundation models incorporate multiple data modalities beyond text.

Q23.

A model that can work with more than one data modality is called a:

A) multimodal model ✓ Correct
B) self-supervised model
C) task-specific model
D) sequence-only model

Rationale: The text defines models operating across multiple modalities as multimodal models.

Q24.

A generative multimodal model is also called a:

A) large masked model
B) large multimodal model ✓ Correct
C) multi-agent model
D) latent memory model

Rationale: The chapter states that a generative multimodal model is also called a large multimodal model (LMM).

AI Engineering 1st edition Huyen Test Bank

Q25.

OpenAI's CLIP is characterized in the chapter primarily as a:

A) generative chatbot model
B) multimodal embedding model ✓ Correct
C) purely autoregressive coding model
D) speech recognition model

Rationale: CLIP is described as an embedding model that produces joint text-image embeddings.

Q26.

Which training approach did OpenAI use in a variant form for CLIP by leveraging co-occurring image-text pairs from the internet?

A) Human reinforcement supervision
B) Natural language supervision ✓ Correct
C) Manual expert annotation
D) Federated supervised learning

Rationale: The chapter identifies CLIP's approach as a variant of self-supervision called natural language supervision.

Q27.

According to the chapter, foundation models mark a transition from task-specific models to:

A) hardware-optimized models
B) rule-based systems
C) general-purpose models ✓ Correct
D) fully deterministic models

Rationale: The text contrasts older task-specific models with newer general-purpose foundation models.

Q28.

Which of the following is identified as a common AI engineering technique for adapting a model without changing model weights?

A) Prompt engineering ✓ Correct
B) Pre-training
C) Quantization-aware retraining
D) Gradient checkpointing

Rationale: Prompt engineering adapts model behavior through instructions and context rather than weight updates.

Q29.

Using a database of customer reviews to supplement instructions for better generation is called:

A) curriculum learning
B) retrieval-augmented generation ✓ Correct
C) contrastive decoding
D) active supervision

Rationale: The chapter defines this approach as retrieval-augmented generation, or RAG.

AI Engineering 1st edition Huyen Test Bank

Q30.
Further training a model on a dataset of high-quality examples to adapt it to a task is called:

A) tokenization
B) finetuning ✓ Correct
C) embedding
D) distillation-only inference

Rationale: Finetuning is the process of continuing training on task-relevant data.

Q31.
Which of the following is NOT one of the three common AI engineering adaptation techniques emphasized in the chapter?

A) Prompt engineering
B) RAG
C) Finetuning
D) K-means clustering ✓ Correct

Rationale: The chapter highlights prompt engineering, RAG, and finetuning, not k-means clustering.

Q32.
Which factor is identified as contributing to the rapid growth of AI engineering by expanding the usefulness of AI across many tasks?

A) General-purpose AI capabilities ✓ Correct
B) Declining internet connectivity
C) Reduced need for evaluation
D) Exclusive reliance on task-specific models

Rationale: General-purpose capabilities increase the number of tasks AI can perform and therefore demand for applications.

Q33.
Which factor lowered the barrier to building AI applications by exposing models through APIs?

A) Natural language supervision
B) Model as a service ✓ Correct
C) Unsupervised benchmarking
D) Static feature deployment

Rationale: Model as a service makes powerful models accessible via APIs without requiring teams to build them from scratch.

Q34.
According to the chapter, why did the author prefer the term "AI engineering" over terms ending in "Ops"?

A) Because operations no longer matter in AI systems
B) Because the focus is more on tweaking foundation models than on operations alone ✓ Correct
C) Because only infrastructure teams use AI
D) Because AI engineering excludes evaluation and monitoring

Rationale: The chapter says the emphasis is on engineering and adapting models, not only on operations.

AI Engineering 1st edition Huyen Test Bank

Q35.

Which use case category is described as the most popular across multiple generative AI surveys?

A) Education
B) Workflow automation
C) Coding ✓ Correct
D) Data organization

Rationale: The chapter explicitly states that coding is the most popular use case in multiple surveys.

Q36.

A consumer or enterprise application that lets users ask questions about their own documents is referred to as:

A) few-shot prompting
B) talk-to-your-docs ✓ Correct
C) closed-book inference
D) zero-retention search

Rationale: The chapter names this information-aggregation use case talk-to-your-docs.

Q37.

Applications that can plan and use external tools to accomplish tasks are called:

A) embedders
B) agents ✓ Correct
C) encoders
D) benchmarks

Rationale: The chapter defines AIs that can plan and use tools as agents.

Q38.

If an application can still function without its AI component, the AI is considered:

A) critical
B) complementary ✓ Correct
C) dynamic
D) proactive

Rationale: The chapter distinguishes complementary AI as nonessential to the app's basic functioning.

Q39.

A feature that responds only after a user request or specific action is best described as:

A) reactive ✓ Correct
B) proactive
C) static
D) foundational

Rationale: Reactive features produce outputs in response to user actions or requests.

AI Engineering 1st edition Huyen Test Bank

Q40.

A feature that is continually updated with user feedback is described as:

A) static
B) reactive
C) dynamic ✓ Correct
D) complementary

Rationale: Dynamic features are updated continuously, often using user feedback or personalization.

Q41.

Involving humans in AI's decision-making process is called:

A) human-in-the-loop ✓ Correct
B) zero-shot learning
C) model distillation
D) cross-validation

Rationale: The chapter uses the term human-in-the-loop for human participation in AI decisions.

Q42.

In Microsoft's Crawl-Walk-Run framework, which stage means human involvement is mandatory?

A) Run
B) Walk
C) Crawl ✓ Correct
D) Scale

Rationale: Crawl is the stage where human involvement remains required.

Q43.

Which metric abbreviation in the chapter refers to time to first token?

A) TPOT
B) TTFT ✓ Correct
C) MMLU
D) RAG

Rationale: TTFT stands for time to first token.

Q44.

Which metric abbreviation refers to time per output token?

A) TPOT ✓ Correct
B) TTFT
C) BOS
D) EOS

Rationale: TPOT is defined as time per output token.

AI Engineering 1st edition Huyen Test Bank

Q45.
The chapter describes the "last mile challenge" of AI products as the idea that:

A) evaluation is unnecessary once a demo works
B) early demos can be easy, but turning them into robust products is much harder ✓ Correct
C) models should always be built from scratch
D) inference costs never change after deployment

Rationale: The text emphasizes that a fun initial demo does not guarantee a production-ready product.

Q46.
What are the three layers of the AI stack described in the chapter?

A) Research, governance, and compliance
B) Application development, model development, and infrastructure ✓ Correct
C) Prompting, coding, and deployment
D) Data labeling, model hosting, and customer support

Rationale: The chapter organizes the AI stack into application development, model development, and infrastructure.

Q47.
Which layer of the AI stack involves providing a model with good prompts, necessary context, and strong interfaces?

A) Infrastructure
B) Application development ✓ Correct
C) Model pre-training
D) Regulatory compliance

Rationale: Application development focuses on prompts, context, interfaces, and evaluation of the user-facing system.

Q48.
Compared with traditional ML engineering, AI engineering focuses less on training new models and more on:

A) manual labeling at scale
B) model adaptation ✓ Correct
C) removing evaluation from workflows
D) single-modality pipelines only

Rationale: The chapter contrasts AI engineering with ML engineering by emphasizing adaptation over original model development.

Q49.
Which statement best distinguishes prompt-based techniques from finetuning?

A) Prompt-based techniques update model weights, whereas finetuning does not
B) Prompt-based techniques adapt behavior without updating weights, whereas finetuning updates weights ✓ Correct
C) Both always require the same amount of data and compute
D) Neither can improve task performance

Rationale: Prompt-based methods use instructions and context, while finetuning changes model weights.

AI Engineering 1st edition Huyen Test Bank

Q50.

Which training phase refers to training a model from scratch with randomly initialized weights?

A) Inference optimization
B) Post-training
C) Pre-training ✓ Correct
D) Prompt engineering

Rationale: Pre-training is the phase in which model weights start random and the model is trained from scratch.

Q51.

A hospital innovation team wants to pilot an AI solution for drafting patient education handouts without investing in model training infrastructure. Which approach from the chapter best lowers the barrier to entry?

A) Pretraining a new language model on hospital data
B) Using a model as a service through an API ✓ Correct
C) Building a custom GPU cluster for inference
D) Labeling millions of examples for supervised training

Rationale: Model as a service allows teams to use powerful existing models via APIs without building models from scratch.

Q52.

A compliance analyst needs a model to identify a missing term in a contract clause by using both the words before and after the blank. Which model type is most appropriate?

A) Autoregressive language model
B) Masked language model ✓ Correct
C) Recommender system
D) Speech synthesis model

Rationale: Masked language models use context on both sides of a missing token to fill in blanks.

Q53.

A telehealth startup is building a chatbot that must generate a reply one token at a time based only on prior text in the conversation. Which model class fits this requirement?

A) Autoregressive language model ✓ Correct
B) Masked language model
C) Image classification model
D) Topic model

Rationale: Autoregressive models predict the next token using only preceding tokens, enabling sequential generation.

Q54.

A health IT student claims that self-supervised learning and unsupervised learning are identical. Which response is most accurate?

A) They are identical because neither uses tokens
B) They are identical because both require manual labels
C) They differ because self-supervision infers labels from input data ✓ Correct
D) They differ because unsupervised learning always uses more compute

Rationale: In self-supervised learning, labels are derived from the data itself, unlike unsupervised learning where labels are not used at all.

AI Engineering 1st edition Huyen Test Bank

Q55.
A team is estimating the size of a foundation model before procurement. According to the chapter, model size is typically measured by the number of what?

A) Tokens
B) Parameters ✓ Correct
C) Prompts
D) Benchmarks

Rationale: The chapter states that model size is typically measured by its number of parameters.

Q56.
An imaging vendor wants one model to interpret both radiology text reports and associated images. Which term best describes this type of model?

A) Task-specific model
B) Multimodal model ✓ Correct
C) Unsupervised clusterer
D) Rule-based expert system

Rationale: A multimodal model works with more than one data modality, such as text and images.

Q57.
A researcher uses paired pathology images and their web captions to train a model without manually assigning class labels. Which training approach from the chapter does this resemble?

A) Natural language supervision ✓ Correct
B) Reinforcement learning from human feedback
C) Manual annotation with adjudication
D) K-means clustering

Rationale: Natural language supervision uses naturally co-occurring image-text pairs instead of manual labels.

Q58.
A team is choosing among adaptation strategies for a retail description generator. Which option requires updating model weights?

A) Prompt engineering
B) Retrieval-augmented generation
C) Finetuning ✓ Correct
D) Template prompting

Rationale: Finetuning adapts a model by changing its weights through additional training.

Q59.
A health insurer wants AI to answer member questions only after a user submits a request in chat. In the chapter's product framing, this feature is best classified as:

A) Proactive
B) Reactive ✓ Correct
C) Static
D) Critical-only

Rationale: Reactive features generate responses in reaction to user requests or specific actions.

AI Engineering 1st edition Huyen Test Bank

Q60.
A pharmacy app pushes refill reminders automatically when it detects an upcoming medication gap. In the chapter's terminology, this AI behavior is:

A) Reactive
B) Static
C) Proactive ✓ Correct
D) Complementary

Rationale: Proactive features surface outputs when there is an opportunity rather than waiting for a request.

Q61.
A health system is deciding whether an AI symptom-checking feature is core to the app or merely helpful. Which distinction from the chapter addresses whether the app can still function without AI?

A) Reactive versus proactive
B) Dynamic versus static
C) Critical versus complementary ✓ Correct
D) Prompting versus finetuning

Rationale: Critical versus complementary distinguishes whether the application still works without AI.

Q62.
A rehabilitation app personalizes coaching over time using each patient's prior interactions and preferences. This is best described as which type of AI feature?

A) Static
B) Dynamic ✓ Correct
C) Masked
D) Task-specific

Rationale: Dynamic features are continually updated using user feedback or personalization signals.

Q63.
A revenue-cycle department wants to start AI cautiously by having staff review every suggested denial-response draft before anything is sent. Which stage of Microsoft's Crawl-Walk-Run framework does this represent?

A) Run
B) Walk
C) Crawl ✓ Correct
D) Scale

Rationale: Crawl means human involvement is mandatory before AI output is used.

AI Engineering 1st edition Huyen Test Bank

Q64.
A payer has evidence that AI-generated responses to simple inquiries are accepted verbatim by agents 95% of the time, so it plans limited direct automation for those inquiries. Which framework transition does this most closely support?

A) From run to crawl
B) From crawl to walk or run ✓ Correct
C) From static to dynamic only
D) From prompt engineering to tokenization

Rationale: High acceptance of AI suggestions can justify moving from mandatory human review toward greater automation.

Q65.
A company is comparing a weekend prototype built with prompts to the effort of training a custom model from scratch for the same use case. Which chapter idea does this illustrate?

A) Foundation models increase data labeling costs
B) Adapting existing models usually reduces time to market ✓ Correct
C) Task-specific models always outperform general-purpose models
D) Open-ended outputs are easier to evaluate than closed-ended outputs

Rationale: The chapter emphasizes that adapting existing foundation models is generally much easier and faster than building from scratch.

Q66.
A clinical documentation team wants a model to produce concise notes in the organization's voice while grounding output in a repository of prior approved examples. Which technique best matches this need if they want to supplement prompts with external content?

A) Tokenization
B) Retrieval-augmented generation ✓ Correct
C) Masked pretraining
D) Speech recognition

Rationale: RAG supplements model generation with information retrieved from an external database or corpus.

Q67.
A medical coding manager asks why larger models generally need more data. Which answer best reflects the chapter?

A) Larger models need more data because tokenization increases label cost
B) Larger models need more data because they have greater capacity to learn ✓ Correct
C) Larger models need more data because APIs require it for each request
D) Larger models need more data because masked models cannot scale

Rationale: The chapter explains that larger models have more learning capacity and therefore need more data to fully realize performance.

AI Engineering 1st edition Huyen Test Bank

Q68.

A startup wants to build an AI feature that extracts information from uploaded insurance cards and receipts into structured fields. Which use-case category best fits this project?

A) Conversational bots
B) Data organization ✓ Correct
C) Education
D) Image and video production

Rationale: Extracting structured information from unstructured documents is a data organization use case.

Q69.

A public health team wants a system that summarizes policy memos, compares related documents, and returns answers conversationally from those files. Which use case from the chapter best describes this?

A) Talk-to-your-docs information aggregation ✓ Correct
B) Image generation
C) Speech synthesis
D) Lead scoring only

Rationale: Conversational retrieval and summarization over documents is described as talk-to-your-docs within information aggregation.

Q70.

A hospital CIO prefers to begin with internal knowledge-management copilots rather than patient-facing chatbots. According to the chapter, what is the main reason organizations often start this way?

A) Internal tools always require no evaluation
B) Internal applications usually reduce privacy, compliance, and catastrophic-failure risks ✓ Correct
C) Internal tools eliminate latency concerns entirely
D) Internal tools do not need model adaptation

Rationale: The chapter notes that companies often deploy internal-facing applications first because they carry lower operational and regulatory risk.

Q71.

A nurse educator wants to use AI to generate personalized practice scenarios for language learners and adapt content to learner preferences. Which chapter theme does this most directly reflect?

A) Education use cases emphasize personalization and tutoring ✓ Correct
B) Coding is the dominant use case in all sectors
C) Image generation is necessary for all learning tasks
D) AI cannot support debate or quiz generation

Rationale: The chapter highlights education applications such as personalized lesson plans, tutoring, and language-practice scenarios.

AI Engineering 1st edition Huyen Test Bank

Q72.
A software team is building an AI assistant that converts natural-language requests into SQL for quality reporting dashboards. Which use-case category best fits this tool?

A) Writing
B) Coding ✓ Correct
C) Conversational therapy
D) Image production

Rationale: Converting English into code such as SQL is presented as a coding use case.

Q73.
A clinic wants AI to draft responses for schedulers, but humans will choose, edit, or reject the suggestions before sending. What role for humans does this represent?

A) Human-out-of-the-loop
B) Human-in-the-loop ✓ Correct
C) Pretraining supervision
D) Static inference

Rationale: When humans review or use AI suggestions in the decision process, the system is human-in-the-loop.

Q74.
A health startup's entire product is a thin wrapper around a popular general chatbot with little proprietary data or distribution. Which business concern from the chapter is most relevant?

A) Vocabulary size inflation
B) Product defensibility is weak ✓ Correct
C) Tokenization failure
D) Self-supervision bottleneck

Rationale: Applications that are easy to replicate and lack unique technology, data, or distribution have weak defensibility.

Q75.
A hospital plans to evaluate a patient-messaging assistant by measuring time to first token, time per output token, and end-to-end response time. These are examples of which metric group?

A) Fairness metrics
B) Latency metrics ✓ Correct
C) Revenue metrics
D) Embedding metrics

Rationale: TTFT, TPOT, and total latency are listed as latency metrics.

AI Engineering 1st edition Huyen Test Bank

Q76.

A support center currently has a median human response time of one hour. When assessing an AI assistant, which chapter principle is most relevant to defining acceptable latency?

A) Latency should be judged relative to the use case and current workflow ✓ Correct

B) Latency never matters for reactive systems

C) Only total token count matters

D) All AI systems require subsecond responses

Rationale: The chapter notes that acceptable latency depends on the use case and can be compared with current human processing times.

Q77.

A health app team celebrates a polished demo built in two days and assumes production deployment will be easy. Which warning from the chapter most directly challenges this assumption?

A) General-purpose models cannot be adapted

B) The last mile from demo to product is often much harder than the initial prototype ✓ Correct

C) Model APIs prevent systematic experimentation

D) Infrastructure is no longer needed for AI products

Rationale: The chapter emphasizes that impressive demos can mask the difficulty of reaching production quality.

Q78.

A biomedical startup is deciding whether to continue with an in-house model after API providers sharply cut prices. Which maintenance principle from the chapter applies most directly?

A) Once chosen, model strategies rarely need reevaluation

B) Teams must continually run cost-benefit analyses because the AI landscape changes quickly ✓ Correct

C) Provider APIs never converge

D) Only regulatory changes affect maintenance decisions

Rationale: Rapid changes in pricing and capabilities require ongoing reassessment of build-versus-buy choices.

Q79.

A genomics company worries that new AI-related laws could abruptly limit access to needed GPUs in its country. Which maintenance risk from the chapter does this illustrate?

A) Prompt drift only

B) Regulatory changes affecting compute availability ✓ Correct

C) Vocabulary collapse

D) Data labeling inflation for self-supervision

Rationale: The chapter notes that regulation can quickly affect compute access and create major operational risk.

AI Engineering 1st edition Huyen Test Bank

Q80.

A game studio avoids certain generative tools because it is uncertain whether outputs built on models trained on others' data could create future ownership disputes. Which concern is this?

A) Latency overrun
B) Interpretability gap
C) Intellectual property risk ✓ Correct
D) Token budget risk

Rationale: The chapter identifies evolving IP rules as a potentially serious risk for AI product builders.

Q81.

A hospital is mapping its AI work into application development, model development, and infrastructure. Which responsibility belongs primarily to the infrastructure layer?

A) Crafting prompts and interfaces
B) Finetuning on domain examples
C) Managing serving, compute, data, and monitoring ✓ Correct
D) Generating benchmark tasks

Rationale: Infrastructure includes model serving, resource management, data management, and monitoring.

Q82.

A product team is mainly focused on prompt design, adding relevant context, user interface design, and rigorous testing of outputs. Which layer of the AI stack is it operating in most directly?

A) Application development ✓ Correct
B) Infrastructure
C) Hardware fabrication
D) Regulatory affairs

Rationale: Application development centers on prompts, context, interfaces, and evaluation for end-user use.

Q83.

A machine learning group is working on training frameworks, finetuning pipelines, dataset engineering, and inference optimization. Which layer does this describe?

A) Application development
B) Model development ✓ Correct
C) Pure infrastructure only
D) Customer success

Rationale: These responsibilities are core functions of the model development layer.

AI Engineering 1st edition Huyen Test Bank

Q84.

An ML engineer asks how AI engineering differs from traditional ML engineering. Which statement best reflects the chapter?

A) AI engineering eliminates the need for evaluation

B) AI engineering focuses more on adapting and evaluating existing models than training new ones from scratch ✓ Correct

C) AI engineering uses only closed-ended models

D) AI engineering no longer needs infrastructure

Rationale: The chapter contrasts AI engineering with ML engineering by emphasizing model adaptation and evaluation over model creation.

Q85.

A health chatbot generates free-form explanations, summaries, and draft replies that can vary widely across prompts. What challenge does this create compared with many traditional ML systems?

A) Open-ended outputs are harder to evaluate systematically ✓ Correct

B) Closed-ended outputs require more GPUs

C) Open-ended outputs remove the need for business metrics

D) Open-ended outputs make latency irrelevant

Rationale: Because foundation models produce open-ended outputs, evaluation becomes a much bigger challenge.

Q86.

A team wants the fastest way to test many vendor models for a utilization-review assistant before collecting a large task-specific dataset. Which adaptation approach is the best fit initially?

A) Prompt-based adaptation ✓ Correct

B) Full pretraining

C) Manual labeling of one million examples

D) Quantization-only retraining

Rationale: Prompt-based techniques are easier to start with, require less data, and support quick experimentation across models.

Q87.

A payer needs to adapt a model to a specialized task it was not exposed to during prior training and also wants potential gains in quality, latency, and cost. Which approach is most appropriate?

A) Prompt engineering alone

B) Finetuning ✓ Correct

C) Tokenization redesign only

D) Removing the EOS token

Rationale: The chapter notes that finetuning can enable new-task adaptation and significant improvements in quality, latency, and cost.

AI Engineering 1st edition Huyen Test Bank

Q88.

A data scientist says, “We trained the chatbot by pasting examples into the prompt.” Based on the chapter, what is the most accurate correction?

A) Correct, because any change to outputs is training

B) Incorrect, because prompt engineering adapts behavior without updating weights and is not training ✓ Correct

C) Correct, because post-training always occurs in prompts

D) Incorrect, because prompts are only used with masked models

Rationale: Prompt engineering changes model behavior through instructions and context, not through weight updates.

Q89.

A biomedical NLP team is discussing training phases. Which statement is correct?

A) Pretraining starts from previously learned weights

B) Finetuning starts from randomly initialized weights

C) Pretraining from scratch is usually the most resource-intensive phase ✓ Correct

D) Post-training never changes model weights

Rationale: The chapter explains that pretraining begins from random initialization and is often by far the most resource-intensive phase.

Q90.

A hospital wants to classify complaint emails as likely spam by framing the task as: “Question: Is this email likely spam?... Answer:” What broader chapter concept does this demonstrate?

A) Many tasks can be reframed as completion tasks ✓ Correct

B) Only translation can be done through completion

C) Completion guarantees factual correctness

D) Classification requires masked models only

Rationale: The chapter shows that tasks such as translation and classification can be framed as text completion.

Q91.

A documentation assistant sometimes responds to a user’s question by continuing the text with another question instead of answering. Which chapter point explains this behavior?

A) Completion is not the same as conversation ✓ Correct

B) Masked models cannot read punctuation

C) Self-supervision prevents dialogue alignment

D) Vocabulary size is too small

Rationale: A completion machine predicts plausible continuations, which may not align with conversational expectations unless post-trained or otherwise adapted.

AI Engineering 1st edition Huyen Test Bank

Q92.

A multilingual model developer is choosing between words, characters, and subword units. According to the chapter, why are tokens commonly used instead of whole words?

A) Tokens always correspond to full medical terms

B) Tokens balance efficiency with meaningful subword structure and help with unknown words ✓ Correct

C) Tokens eliminate the need for vocabulary design

D) Tokens guarantee better factual accuracy

Rationale: Tokens reduce vocabulary size while preserving more meaning than characters and helping models process novel words.

Q93.

A research fellow asks why BOS and EOS markers are included in training examples for language models. Which answer is best?

A) They are used only for image classification

B) They help models work with multiple sequences and indicate when to start and stop ✓ Correct

C) They reduce the number of parameters

D) They are equivalent to labels from human annotators

Rationale: Beginning- and end-of-sequence markers define sequence boundaries and help the model know when to end output.

Q94.

A startup is considering a model that can jointly understand text and images but does not generate open-ended outputs; instead, it creates shared vector representations. Which description best fits it?

A) A multimodal embedding model ✓ Correct

B) An autoregressive decoder-only model

C) A rule-based parser

D) A supervised-only classifier with manual labels

Rationale: The chapter describes models like CLIP as multimodal embedding models that produce joint text-image embeddings rather than open-ended generations.

Q95.

A hospital executive asks why AI engineering has grown so quickly. Which combination best matches the chapter's three main drivers?

A) Task-specific models, reduced regulation, and lower storage costs

B) General-purpose capabilities, increased investment, and low barriers to building applications ✓ Correct

C) More labeled datasets, fewer APIs, and slower hardware cycles

D) Smaller models, less demand, and less competition

Rationale: The chapter attributes rapid growth to broad model capabilities, surging investment, and accessible model-as-a-service tooling.

AI Engineering 1st edition Huyen Test Bank

Q96.

A healthcare startup wants to prioritize one initial AI project. Which option most closely reflects the chapter's guidance on use-case evaluation?

A) Choose the most technically impressive demo regardless of business need

B) Start with a use case tied to clear risks or opportunities for the business ✓ Correct

C) Always build externally facing tools before internal ones

D) Avoid buying existing tools even when they perform better

Rationale: The chapter advises grounding AI projects in concrete business risks and opportunities rather than novelty alone.

Q97.

A utilization management team is considering whether to buy an existing AI summarization tool or build its own. Which chapter framing is most directly relevant?

A) Masked versus autoregressive modeling

B) The classic buy-or-build decision ✓ Correct

C) Vocabulary compression ratio

D) Inference token entropy

Rationale: The chapter explicitly frames this choice as a buy-or-build question teams must assess for themselves.

Q98.

A medical claims startup has strong proprietary workflow data from early customers and uses that usage information to improve prompts, collect better examples, and refine product decisions. Which competitive advantage from the chapter is this leveraging most directly?

A) Distribution only

B) Data moat ✓ Correct

C) Compute monopoly

D) Vocabulary expansion

Rationale: The chapter highlights data and usage insights as a nuanced but powerful moat for AI products.

Q99.

A clinical operations leader wants to know which measurement best reflects business impact for a customer support bot rather than technical performance alone. Which is the best example?

A) Number of parameters in the base model

B) Percentage of customer messages the bot can automate ✓ Correct

C) Vocabulary size of the tokenizer

D) Average embedding dimension

Rationale: Automated message percentage is a business-facing success metric tied to operational impact.

AI Engineering 1st edition Huyen Test Bank

Q100.

A health system is selecting between a smaller task-specific model and a larger general-purpose foundation model for prior-authorization support. Which conclusion is most consistent with the chapter?

A) General-purpose models always dominate task-specific models in speed and cost

B) Task-specific models may still be preferable because they can be smaller, faster, and cheaper ✓ Correct

C) Task-specific models cannot be adapted with retrieval

D) General-purpose models remove the need for evaluation

Rationale: The chapter notes that despite the flexibility of foundation models, task-specific models can offer advantages in size, speed, and cost.

Q101.

A startup can launch an image-captioning product in days by calling a multimodal model API rather than training its own model. Which chapter concept best explains why this is now feasible for many teams?

A) The rise of model as a service lowered the barrier to application development ✓ Correct

B) Masked language models replaced autoregressive models for generation

C) Unsupervised learning eliminated the need for prompts

D) Vocabulary expansion removed the need for compute infrastructure

Rationale: Model as a service lets teams access powerful foundation models through APIs without building and serving models themselves.

Q102.

A team is deciding whether to use a masked language model or an autoregressive model for a code assistant that must generate new code line by line. Which choice is most appropriate?

A) Masked language model, because it uses future tokens to produce open-ended generation efficiently

B) Autoregressive model, because it predicts the next token using preceding context and supports sequential generation ✓ Correct

C) Masked language model, because it is the standard architecture for text generation in current production systems

D) Either model, because both are equally suited for long-form token-by-token generation

Rationale: Autoregressive models are preferred for text generation because they generate one next token at a time from prior context.

Q103.

An executive claims that self-supervision and unsupervised learning are interchangeable terms. Which response best evaluates this claim?

A) The claim is correct because both avoid all labels entirely

B) The claim is partially correct because self-supervision uses manually curated labels only during evaluation

C) The claim is incorrect because self-supervision infers labels from the input data, whereas unsupervised learning does not use labels at all ✓ Correct

D) The claim is incorrect because unsupervised learning requires more labeled data than self-supervision

Rationale: Self-supervised learning derives training targets from the data itself, unlike unsupervised learning, which uses no labels.

AI Engineering 1st edition Huyen Test Bank

Q104.

Why did self-supervision become a key enabler of scaling language models into LLMs?

A) It made tokenization unnecessary, reducing preprocessing costs

B) It allowed models to learn from abundant text without manual labeling bottlenecks ✓ Correct

C) It guaranteed factual correctness in generated outputs

D) It eliminated the need for large parameter counts

Rationale: Self-supervision leverages naturally occurring text to create training signals without expensive human labeling.

Q105.

A researcher argues that adding image understanding to a text model does not change its category because it still generates text. Which critique is most aligned with the chapter?

A) The argument is valid because output modality alone defines the model type

B) The argument is weak because models that process multiple modalities are better characterized as foundation or multimodal models ✓ Correct

C) The argument is valid because all multimodal systems are just embedding models

D) The argument is weak only if the model also performs speech recognition

Rationale: The chapter distinguishes multimodal foundation models from text-only LLMs based on supported input modalities, not only output form.

Q106.

A company wants a model that can classify images using internet-scale image-text pairs without manually labeling every image category. Which example from the chapter most directly supports this strategy?

A) BERT using masked token prediction

B) GPT-2 using next-token prediction

C) CLIP using natural language supervision on image-text pairs ✓ Correct

D) AlexNet using ImageNet category labels

Rationale: CLIP used naturally co-occurring image-text pairs as supervision rather than manually labeled image classes.

Q107.

Which scenario best illustrates a transition from task-specific models to general-purpose foundation models?

A) A translation model that only translates between English and French

B) A sentiment classifier retrained separately for every domain

C) A single model used out of the box for translation, summarization, and classification ✓ Correct

D) An OCR pipeline that requires different handcrafted rules for each document type

Rationale: Foundation models are general-purpose because one model can perform many tasks reasonably well without separate task-specific training.

AI Engineering 1st edition Huyen Test Bank

Q108.

A retail team needs product descriptions that match brand voice and customer language. Which adaptation strategy most strongly synthesizes chapter guidance for improving an out-of-the-box model?

A) Use only a larger vocabulary tokenizer

B) Combine prompt engineering, retrieval from customer reviews, and possibly finetuning ✓ Correct

C) Replace the model with a masked language model

D) Reduce the number of parameters to avoid clichés

Rationale: The chapter identifies prompt engineering, RAG, and finetuning as common techniques for adapting general-purpose models to specific needs.

Q109.

An instructor asks why completion is considered powerful if it is not the same as conversation. Which answer best captures the chapter's reasoning?

A) Completion is powerful because many tasks can be reframed as predicting what text comes next ✓ Correct

B) Completion is powerful because it always returns the correct answer if the prompt is clear

C) Completion is powerful only for translation and not for classification

D) Completion is powerful because it removes the probabilistic nature of language generation

Rationale: Many tasks such as translation, summarization, coding, and classification can be cast as completion problems.

Q110.

A product manager says, 'If our language model answers a question with another question, the model is broken.' Which evaluation is most accurate?

A) Correct, because completion models are designed exclusively for dialogue

B) Incorrect, because a completion machine predicts likely continuations and may not naturally follow conversational norms without post-training ✓ Correct

C) Correct, because autoregressive models cannot generate interrogative text

D) Incorrect, because only masked language models can answer questions directly

Rationale: The chapter notes that completion alone does not ensure appropriate conversational behavior; post-training helps align responses to user requests.

Q111.

A team plans to train a 100-billion-parameter model on a very small corpus due to budget constraints. What is the strongest chapter-based critique?

A) Large models cannot technically be trained on small datasets

B) Small datasets always improve generalization in larger models

C) Training a large model on too little data wastes compute because a smaller model could achieve similar or better results on that dataset ✓ Correct

D) Parameter count is unrelated to data requirements

Rationale: The chapter states that larger models have more capacity and need more data to realize it; otherwise compute is wasted.

AI Engineering 1st edition Huyen Test Bank

Q112.

Which statement best explains why tokens, rather than full words or characters alone, are commonly used as the basic unit in language models?

A) Tokens always correspond one-to-one with dictionary words

B) Tokens balance semantic usefulness and vocabulary efficiency while helping with unknown words ✓ Correct

C) Tokens eliminate the need for special sequence markers

D) Tokens are used only in multilingual models

Rationale: Tokens preserve more meaning than characters, require fewer unique units than words, and help decompose unseen words.

Q113.

An analyst wants to estimate text length for API budgeting. Based on the chapter, what is the closest approximation?

A) 100 tokens is approximately 25 words

B) 100 tokens is approximately 50 words

C) 100 tokens is approximately 75 words ✓ Correct

D) 100 tokens is approximately 100 words

Rationale: The chapter states that for GPT-4 an average token is about three-quarters of a word, so 100 tokens is roughly 75 words.

Q114.

A healthcare documentation vendor is deciding whether to build a proprietary model from scratch or adapt an existing foundation model. Which recommendation best reflects the chapter's buy-or-build framing?

A) Always build from scratch because task-specific models are universally superior

B) Always buy because foundation models make custom development obsolete

C) Evaluate trade-offs such as time to market, adaptation effort, cost, and whether smaller task-specific models offer operational advantages ✓ Correct

D) Delay the decision until regulations fully stabilize worldwide

Rationale: The chapter frames this as a classic buy-or-build decision shaped by performance, speed, cost, and deployment considerations.

Q115.

Which combination of factors best explains the rapid rise of AI engineering as a discipline?

A) Task-specific architectures, lower evaluation burden, and declining internet usage

B) General-purpose capabilities, increased AI investment, and lower barriers to building applications ✓ Correct

C) Reduced need for models, increased labeling costs, and slower deployment cycles

D) Smaller models, less demand for applications, and fewer APIs

Rationale: The chapter identifies those three factors as creating ideal conditions for AI engineering's growth.

AI Engineering 1st edition Huyen Test Bank

Q116.

A board member cites increased stock prices after companies mention AI on earnings calls as proof that AI causes improved performance. What is the best analytical response?

A) The conclusion is valid because market reactions establish causality

B) The conclusion is invalid because earnings calls never affect stock prices

C) The conclusion is uncertain because the relationship may reflect correlation rather than causation ✓ Correct

D) The conclusion is valid only for S&P 500 technology firms

Rationale: The chapter explicitly notes uncertainty about whether such associations reflect causation or correlation.

Q117.

Why does the chapter prefer the term 'AI engineering' over labels ending in 'Ops' for this context?

A) Because operational concerns no longer matter in production AI

B) Because the emphasis is on tweaking and building with foundation models, not only operations ✓ Correct

C) Because AI engineering excludes evaluation and monitoring

D) Because foundation models cannot be deployed through operational pipelines

Rationale: The chapter argues that the focus is more on engineering and adaptation than on purely operational activities.

Q118.

A hospital innovation team wants a low-risk first generative AI deployment. Which use case is most consistent with the chapter's observed enterprise adoption pattern?

A) An external autonomous diagnostic chatbot for patients

B) An internal knowledge-management summarization assistant for staff ✓ Correct

C) A fully automated public relations bot posting to social media

D) A consumer-facing AI companion app for behavioral health

Rationale: The chapter notes that organizations tend to deploy lower-risk internal-facing applications before external-facing ones.

Q119.

An exam author wants to use AI for a task that is easy to evaluate and therefore lower risk than open-ended generation. Which application best fits that logic?

A) Generating unrestricted poetry for publication

B) Creating an always-on public chatbot with no review

C) Classifying emails as likely spam or not spam ✓ Correct

D) Writing interactive fiction with branching plots

Rationale: The chapter notes that close-ended tasks like classification are easier to evaluate and have more estimable risks.

AI Engineering 1st edition Huyen Test Bank

Q120.

A software leader is assessing where coding assistants are likely to yield the greatest productivity gains first. Which conclusion aligns best with the chapter?

A) Highly complex tasks show the largest gains, while documentation gains are minimal

B) Productivity gains are strongest for simpler tasks such as documentation, with less improvement on highly complex tasks ✓ Correct

C) Backend development uniformly benefits more than frontend development

D) Coding is a poor use case because AI cannot write or refactor code

Rationale: The chapter reports larger gains for simpler tasks like documentation and smaller gains for highly complex work.

Q121.

A marketing team wants to personalize seasonal ad creatives across locations. Which capability of foundation-model-based image tools most directly supports this use case?

A) Deterministic rule-based rendering from fixed templates only

B) Probabilistic generation of multiple visual variations for testing and localization ✓ Correct

C) Exclusive reliance on human-labeled segmentation masks

D) Inability to modify existing images once generated

Rationale: The chapter highlights AI's probabilistic creativity and ability to generate many ad variations, including seasonal and regional changes.

Q122.

Why is writing considered an especially suitable domain for AI assistance according to the chapter?

A) Because writing tasks require zero factual accuracy

B) Because users write frequently, the work can be tedious, and people often tolerate imperfect suggestions they can ignore ✓ Correct

C) Because all writing tasks are low stakes

D) Because language models are deterministic when editing prose

Rationale: The chapter emphasizes writing frequency, tedium, and relatively high tolerance for imperfect suggestions.

Q123.

A dean is deciding whether to prohibit or integrate generative AI in coursework. Which chapter-based argument most strongly supports integration over blanket prohibition?

A) AI eliminates the need for teachers to evaluate student work

B) AI can personalize materials, roleplay practice scenarios, generate quizzes, and support tutoring ✓ Correct

C) AI guarantees academic integrity if students use it

D) AI is useful only for grading, not learning

Rationale: The chapter presents AI as a tool for personalization, tutoring, practice, and assessment support in education.

Q124.

A company is designing a customer support system and wants to minimize risk while still learning from AI performance. Which deployment path best matches the chapter's Crawl-Walk-Run framework?

A) Immediately let AI answer all external customer requests autonomously

B) Start with AI-generated suggestions for human agents, then expand automation as acceptance rates justify it ✓ Correct

C) Avoid any human involvement because it slows learning

D) Use AI only for infrastructure monitoring, not support workflows

Rationale: Crawl-Walk-Run begins with mandatory human involvement and increases automation as system quality is validated.

Q125.

A founder proposes a startup whose only function is a simple layer over a popular foundation model feature likely to be added natively by the provider. What is the most important strategic concern raised in the chapter?

A) The startup may face weak defensibility because its layer can be subsumed by the underlying model or platform ✓ Correct

B) The startup will necessarily have stronger distribution than incumbents

C) The startup's data moat is guaranteed from day one

D) The startup avoids all maintenance risk by depending on a major provider

Rationale: The chapter warns that thin layers over foundation models can be absorbed as the base models or platforms expand capabilities.

Q126.

Which set of moats does the chapter identify as the main forms of competitive advantage in AI products?

A) Branding, office location, and patent volume

B) Technology, data, and distribution ✓ Correct

C) Compute, tokenization, and entropy

D) Latency, sequence markers, and star counts

Rationale: The chapter identifies technology, data, and distribution as the three main competitive advantages.

Q127.

A startup lacks proprietary preexisting datasets but is first to market and can collect rich usage signals. According to the chapter, how should this be interpreted strategically?

A) It has no path to defensibility because only historical data matters

B) Usage data can become a moat by informing product improvement and guiding future data collection and training ✓ Correct

C) Usage data is irrelevant unless directly used for pre-training

D) The startup should avoid collecting feedback because it slows deployment

Rationale: The chapter notes that first movers can build a data advantage through usage information even when direct training use is limited.

AI Engineering 1st edition Huyen Test Bank

Q128.

A product owner says success for a support chatbot should be judged only by how many messages it answers. Which evaluation is best?

A) Correct, because volume is the sole meaningful business metric

B) Incomplete, because business impact should also consider speed, labor savings, and customer satisfaction ✓ Correct

C) Incorrect, because only model-level perplexity matters

D) Correct, because satisfaction cannot be measured for AI systems

Rationale: The chapter stresses aligning AI metrics with business metrics such as automation rate, response time, labor savings, and user satisfaction.

Q129.

Which metric grouping is most directly associated with response speed in foundation model applications?

A) Interpretability, fairness, and calibration

B) TTFT, TPOT, and total latency ✓ Correct

C) Precision, recall, and F1 score

D) MMLU, BLEU, and ROUGE

Rationale: The chapter specifically lists time to first token, time per output token, and total latency as latency metrics.

Q130.

A team built a compelling weekend demo and is now forecasting a near-term polished product launch. Which chapter concept most strongly challenges this assumption?

A) Natural language supervision

B) The last mile challenge from demo quality to production usefulness ✓ Correct

C) Vocabulary compression

D) Bidirectional encoding

Rationale: The chapter warns that impressive early demos can mask the long and difficult path to a reliable production product.

Q131.

A manager is surprised that moving a model-powered feature from 80% to 95% quality took far longer than reaching 80%. What explanation is most consistent with the chapter?

A) Late-stage gains are often disproportionately difficult because product kinks and hallucinations dominate the final improvement stages ✓ Correct

B) Foundation models improve linearly with every engineering hour

C) The issue proves the model should have been unsupervised rather than self-supervised

D) This only occurs in image generation, not language applications

Rationale: The chapter emphasizes that progress from good to highly reliable performance is much harder than initial gains.

AI Engineering 1st edition Huyen Test Bank

Q132.

A CIO is choosing between building in-house infrastructure and relying on a model provider. Three months later, provider prices drop dramatically. Which lesson from the chapter is most relevant?

A) AI infrastructure decisions are static once made

B) Rapid changes in model cost and capability require ongoing cost-benefit reassessment ✓ Correct

C) Price declines eliminate all vendor risk

D) In-house systems are always cheaper over time

Rationale: The chapter highlights the need for continual reassessment because the best option today may become the worst tomorrow.

Q133.

A multinational firm is preparing an AI rollout and underestimates the impact of emerging regulation on compute, data, and compliance. Which chapter-based evaluation is strongest?

A) Regulatory change is mostly cosmetic for AI teams

B) Regulation can materially affect compliance costs, compute access, and even product viability ✓ Correct

C) Only open source models are affected by regulation

D) Regulation matters only after a product reaches profitability

Rationale: The chapter notes that regulations around privacy, compute, and national security can create major operational and strategic disruptions.

Q134.

A game studio hesitates to use generative AI in art pipelines despite technical benefits. Which concern from the chapter most directly explains this caution?

A) Foundation models cannot generate images at production quality

B) IP ownership and legal uncertainty around training data and generated outputs remain unresolved ✓ Correct

C) AI tools require no maintenance planning

D) Only text applications face legal risk

Rationale: The chapter notes that IP-heavy companies worry about evolving legal rules around ownership and training data.

Q135.

Which statement best characterizes the three layers of the AI stack described in the chapter?

A) Application development, model development, and infrastructure ✓ Correct

B) Prompting, tokenization, and entropy

C) Pre-training, post-training, and deployment only

D) Data labeling, annotation, and benchmarking

Rationale: The chapter organizes the AI stack into application development, model development, and infrastructure.

AI Engineering 1st edition Huyen Test Bank

Q136.

A consultant claims that foundation models completely changed AI infrastructure requirements, making prior ML operational knowledge mostly obsolete. Which response is best supported by the chapter?

A) Correct, because serving and monitoring are no longer needed

B) Incorrect, because many core infrastructure needs such as resource management, serving, and monitoring remain the same ✓ Correct

C) Correct, because application development replaces all infrastructure concerns

D) Incorrect, but only for image models

Rationale: The chapter states that while applications and models changed, core infrastructural needs largely remained constant.

Q137.

Which difference most clearly distinguishes AI engineering from traditional ML engineering at a high level?

A) AI engineering avoids evaluation because outputs are open-ended

B) AI engineering focuses more on adapting and evaluating existing large models than on training task-specific models from scratch ✓ Correct

C) AI engineering uses fewer compute resources than traditional ML systems

D) AI engineering cannot use feedback loops in production

Rationale: The chapter frames AI engineering as less about model creation and more about adaptation and evaluation of existing foundation models.

Q138.

Why does evaluation become a larger problem in AI engineering than in many traditional ML settings?

A) Foundation models are smaller and easier to benchmark

B) Open-ended outputs increase flexibility but make correctness and quality harder to assess systematically ✓ Correct

C) Evaluation is unnecessary when using API-based models

D) Prompt engineering removes output variability

Rationale: The chapter emphasizes that open-ended outputs are much harder to evaluate than close-ended predictions.

Q139.

A nontechnical founder wants to build an AI application without learning gradient descent or neural network internals. Which chapter-based conclusion is most appropriate?

A) This is unrealistic because ML knowledge is mandatory for any AI application

B) This is feasible because foundation models reduce the need for deep ML expertise, though such knowledge remains valuable ✓ Correct

C) This is feasible only if the founder builds a model from scratch

D) This is impossible unless the application uses masked language models

Rationale: The chapter notes that ML expertise is no longer required to start building AI applications, though it still helps with tool choice and troubleshooting.

AI Engineering 1st edition Huyen Test Bank

Q140.

Which statement most accurately differentiates pre-training from finetuning?

A) Pre-training updates an already trained model, while finetuning starts from random weights

B) Pre-training is prompt-based adaptation, while finetuning changes only tokenization

C) Pre-training trains from scratch with randomly initialized weights, while finetuning continues training from an existing model ✓ Correct

D) Pre-training and finetuning are unrelated processes with different toolchains

Rationale: The chapter defines pre-training as training from scratch and finetuning as continued training from a previously trained model.

Q141.

A team says quantizing model weights counts as training because the weight values change. Which answer best reflects the chapter's distinction?

A) Correct, because any weight change is training

B) Incorrect, because training changes weights through learning, whereas quantization changes precision and is not considered training ✓ Correct

C) Correct, but only during post-training

D) Incorrect, because quantization applies only to embeddings

Rationale: The chapter explicitly notes that not all weight changes are training and gives quantization as an example.

Q142.

A model provider improves an LLM's instruction-following behavior before public release. An enterprise customer later adapts that released model to internal legal documents. Which pairing best classifies these two steps?

A) Provider finetuning, then customer pre-training

B) Provider post-training, then customer finetuning ✓ Correct

C) Provider quantization, then customer supervision

D) Provider prompting, then customer tokenization

Rationale: The chapter uses post-training for provider-side training after pre-training and finetuning for customer-side adaptation.

Q143.

A team must adapt a model quickly with very little labeled data and wants to compare many candidate models before deeper investment. Which adaptation approach best fits?

A) Prompt-based adaptation, because it avoids changing weights and is easy to start with ✓ Correct

B) Pre-training, because it requires the least data and compute

C) Full finetuning, because it is always simpler than prompting

D) Masked language modeling, because it guarantees domain alignment

Rationale: Prompt-based techniques are easier to start, require less data, and make it easier to experiment across models.

AI Engineering 1st edition Huyen Test Bank

Q144.

When is finetuning more likely to be justified over prompt engineering alone?

A) When the task is simple and performance requirements are loose

B) When strict performance, quality, latency, or novel task adaptation requires changing model behavior more deeply ✓ Correct

C) When the team wants to avoid updating model weights

D) When no training data is available at all

Rationale: The chapter notes that finetuning is more complex but can significantly improve quality, latency, cost, and adaptation to new tasks.

Q145.

A CTO asks why AI engineering puts more pressure on efficient inference optimization than many earlier ML systems. Which answer is best?

A) Because foundation models are typically larger, more compute-intensive, and higher-latency than traditional models ✓ Correct

B) Because foundation models no longer need serving infrastructure

C) Because API-based access makes latency irrelevant

D) Because open-ended outputs reduce computational cost

Rationale: The chapter highlights model scale, compute consumption, and latency as key reasons inference optimization matters more.

Q146.

A hiring manager wants to know which engineering skill demand is likely to grow as organizations adopt compute-intensive foundation models. Which chapter-based answer is strongest?

A) Less need for GPU and cluster expertise because providers abstract everything away

B) More need for engineers who know how to work with GPUs and large compute clusters ✓ Correct

C) Less need for evaluation expertise because benchmarks are standardized

D) More need for manual data labeling specialists than infrastructure engineers

Rationale: The chapter notes that larger compute demands increase the need for engineers skilled with GPUs and big clusters.

Q147.

A product team is mapping business goals to technical success criteria for an AI support assistant. Which approach best reflects the chapter's planning guidance?

A) Track only model-centric metrics and ignore business outcomes until launch

B) Define business metrics and usefulness thresholds, then evaluate off-the-shelf models against them before planning adaptation work ✓ Correct

C) Choose the most popular model and assume it will meet the business need

D) Delay evaluation until after full-scale deployment

Rationale: The chapter recommends setting measurable business and technical goals, evaluating existing models, and planning from the baseline.

AI Engineering 1st edition Huyen Test Bank

Q148.

A feature automatically surfaces AI-generated recommendations when the user has not explicitly requested them. According to the chapter, what implication follows?

A) Latency always matters more than for reactive systems

B) The feature is proactive and therefore usually faces a higher quality bar because users may find low-quality outputs intrusive ✓ Correct

C) The feature is static and does not require updates

D) The feature is complementary and therefore can tolerate any error rate

Rationale: The chapter states that proactive features can feel intrusive if low quality, so they generally require a higher quality threshold.

Q149.

A biometric login system would fail entirely if its AI component stopped working, whereas an email client could still function without text suggestions. How does the chapter classify this difference?

A) Reactive versus proactive

B) Static versus dynamic

C) Critical versus complementary ✓ Correct

D) Prompted versus finetuned

Rationale: The chapter distinguishes AI that is essential to app function as critical, versus AI that merely augments an app as complementary.

Q150.

A personalized study assistant continually adjusts to each learner's preferences and history, potentially with memory or user-specific adaptation. Which chapter concept best describes this design?

A) Static feature using one shared model with periodic updates only

B) Dynamic feature with ongoing personalization over time ✓ Correct

C) Complementary feature with no need for evaluation

D) Task-specific model that cannot use foundation models

Rationale: The chapter describes dynamic features as those continually updated with user feedback or personalization mechanisms over time.

AI Engineering 1st edition Huyen Test Bank