

ENGLISH FOR WORK / ENGINEERING TEAMS / RESEARCH TO PRODUCTION

AI Development English

Precise language for models, evidence, and delivery.



Phil Stilwell

© 2026 Phil Stilwell. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior written permission from the author, except as permitted by law.

Eight practical lessons. Eleven extended conversations.

Vocabulary, phrase banks, guided speaking, and explained answers.

Revised 2026 edition / Upper-intermediate to advanced

Your route through the book

01 Speaking the AI Development Stack, page 4

Distinguish components and ask a diagnostic question without assigning blame.

02 LLMs, Transformers, Tokens, and Context, page 14

Correct a technical misconception tactfully and negotiate a bounded alternative.

03 Data, Datasets, Labels, and Leakage, page 27

Disclose compromised evidence and resist pressure to overstate readiness.

04 Retrieval, Embeddings, Vector Search, and RAG, page 38

Negotiate a trade-off using matched comparisons and explicitly limited claims.

05 Fine-Tuning, Alignment, and Adaptation, page 51

Separate proposed remedies by mechanism and agree on a controlled experiment.

06 Evaluation, Benchmarks, and Regression, page 62

Present disaggregated results and challenge an oversimplified release argument.

07 Inference, Latency, Cost, and Deployment, page 75

Negotiate a staged deployment with explicit conditions and rollback ownership.

08 Safety, Security, Privacy, and Governance, page 85

Give a precise incident account, protect evidence, and distinguish findings from scope assumptions.

[Professional communication notes, page 3](#)

[Three additional workplace cases, page 99](#)

[Answers and explanations, page 110](#)

[Phrase finder, page 168](#)

[Vocabulary index, page 170](#)

[References and publication notes, page 175](#)

Suggested study range: B2-C1. This is an editorial study range, not a certified proficiency assessment.

Name the claim. Name the evidence.

AI development discussions move between research, software, product, and risk. Shared words do not always carry shared meanings. Effective English makes the component, comparison, and uncertainty explicit.

Separate the system from the model

An application can fail in retrieval, prompting, serving, permissions, or presentation even when its model is unchanged.

"The retrieved passage is correct; the generated answer omits the exception."

Keep the denominator visible

A percentage needs a population and criterion. An average can conceal weak performance on an important slice.

"It passed 88 of these 100 cases; that is not an estimate for all customer traffic."

Disagree with a testable claim

Ablations, baselines, and matched comparisons make technical disagreement useful. A request for evidence is not necessarily a personal challenge.

"What changes if we hold the retrieval set constant and swap only the reranker?"

Do not turn a hypothesis into a finding

Use may, appears, and is consistent with when evidence is limited. State what would confirm or weaken the explanation.

"The timing suggests a queueing issue, but we have not isolated the cause."

Professional boundaries

Fictional communication practice, not deployment, security, or compliance instructions. Use authorized review, current documentation, and approved procedures for real systems. Never place real secrets or personal data in practice exercises.

A Read the situation

LESSON 01 / SPEAKING THE AI DEVELOPMENT STACK

Distinguish components and ask a diagnostic question without assigning blame.

The answer is wrong, but which layer failed?

A document assistant displays the wrong refund deadline. Retrieval returned the current policy, which states 30 days, but the generated answer says 60. The interface displays the generated text unchanged. Priya owns the model integration; Alex manages the product. A demonstration is scheduled for Friday. No cause has been confirmed, and Alex cannot approve a production release alone.

People: Priya, AI integration engineer; Alex, Product manager.

Read the record

Choose one answer per question using the facts above.

1. Which observation is confirmed?

- A.** Production release is approved.
- B.** The generated answer says 60 days.
- C.** Retrieval found an obsolete policy.
- D.** The interface altered the number.

2. Who owns the model integration?

- A.** The policy author
- B.** Priya
- C.** Alex alone
- D.** The customer

3. What remains unknown?

- A.** Whether a demonstration is scheduled
- B.** Whether the interface displays the text
- C.** Why the model produced the wrong deadline
- D.** The correct policy deadline

[Answers and explanations, page 110](#)

B Vocabulary in use

CORE TERMS

application stack

The components that together deliver an application.

map the application stack

language model

A model that processes and generates language sequences.

query a language model

orchestration

Coordination of model, retrieval, and tool steps.

trace the orchestration flow

endpoint

An address through which a service accepts requests.

call an endpoint

trace

A linked record of events in a request's path.

examine a request trace

upstream

Earlier in the relevant processing path.

inspect the upstream response

retrieval

Finding source material relevant to a request.

inspect retrieval results

user interface

The part of a system a user sees and operates.

inspect the user interface

inference

Running a trained model to produce an output.

run inference

payload

The data carried in a request or response.

inspect the request payload

dependency

A component or condition another part relies on.

identify a service dependency

downstream

Later in the relevant processing path.

assess downstream impact

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

application stack APPLICATION STACK

map the application stack

Keep the meaning precise: The components that together deliver an application.

inference INFERENCE

run inference

Keep the meaning precise: Running a trained model to produce an output.

B Precision and professional shorthand

EXTEND YOUR RANGE

symptom

An observed sign of a problem, not its cause.

describe the symptom

reproduction

Recreating a reported result under stated conditions.

obtain a reliable reproduction

schema

A defined structure for data and its fields.

validate a response schema

API

Application programming interface; a defined software interaction boundary.

document an API contract

integration

Connecting components into a working system.

test the integration

triage

Initial assessment that determines priority and ownership.

triage the defect

root cause

The underlying cause established by investigation.

verify the root cause

request ID

An identifier used to locate a particular request.

attach the request ID

tool call

A model-requested invocation of an available tool.

inspect a tool call

contract

Agreed expectations between software components.

preserve the service contract

observability

The ability to infer system behavior from available signals.

improve observability

handoff

Transfer of work and its relevant context.

prepare a clear handoff

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

downstream DOWNSTREAM

assess downstream impact

Keep the meaning precise: Later in the relevant processing path.

schema SCHEMA

validate a response schema

Keep the meaning precise: A defined structure for data and its fields.

Phrases that move the work forward

Locate the layer

Which component first shows the wrong value?

Limit the conclusion

That locates the mismatch, but it does not establish the cause.

Separate responsibilities

Product approval and technical diagnosis are different decisions.

Clarify a term

By "the AI," do you mean the model or the whole application?

Check the interface

Does the interface display the response without changing it?

Narrow the test

Change one component while keeping the others fixed.

Confirm a dependency

Does this test depend on the policy index being current?

Set the update

I can commit to an update, not to a fix by that time.

State the observation

The retrieved passage says 30 days; the generated answer says 60.

Request a trace

Can you attach the request ID and the redacted trace?

Ask for reproduction

Can we reproduce it with the same payload and configuration?

Agree on a handoff

I will send the evidence and name the unresolved question.

Avoid blame

Let us inspect the boundary before assigning ownership.

Name the uncertainty

We have an observation, not a verified root cause.

Preserve the record

Please retain the configuration used for that request.

Close precisely

The demonstration remains scheduled; release approval is separate.

Precision in real exchanges

Purpose: Request a trace

Useful wording: Can you attach the request ID and the redacted trace?

Purpose: Separate responsibilities

Useful wording: Product approval and technical diagnosis are different decisions.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which question locates the first mismatch?

- A.** Should the interface team own the cause because it displays the answer?
- B.** Can a single corrected reply establish that the issue is fixed?
- C.** At which boundary does 30 become 60?
- D.** Can we assign the fault to the model because retrieval found the right passage?

2. Complete: "Could you confirm whether ___?"

- A.** the payload does includes the policy
- B.** the payload includes the policy
- C.** does the payload include the policy
- D.** includes the payload the policy

3. A colleague needs the term for this meaning: The part of a system a user sees and operates. Which term fits?

- A.** user interface
- B.** reproduction
- C.** integration
- D.** trace

D Choose the response that fits / continued

4. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "So the model got it wrong? I need to tell the team who is picking this up." Priya: [missing response] Alex: "Wait, could the interface have changed the number? We had a display issue last time." Which response belongs in the gap?

- A.** Tell them the answer is wrong, but leave the cause open. The payload could contain another instruction or an older example.
- B.** That gives us a specific mismatch. Retrieval found the current passage. I still need to inspect exactly what reached the model.
- C.** The recorded response already says 60, and the interface displays it unchanged. That puts the first observed difference upstream of presentation.
- D.** The configuration and a redacted trace. Then I can repeat the request and compare the inputs without changing several variables.

[Answers and explanations, page 110](#)

One useful distinction

A trace records events; a diagnosis explains them. A wrong answer identifies a symptom, not automatically a model defect. Keep retrieval output, model output, and displayed text separate.

E Conversation

The answer is wrong, but which layer failed?

The answer is wrong, but which layer failed?

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

request ID / trace / handoff / reproduction / configuration / mismatch / triage /
upstream / payload / root cause

- 01 **Alex:** The refund answer is wrong again. The policy says 30 days, but our assistant tells the customer 60. Which team should investigate first?
- 02 **Priya:** Do you have the ¹ _____? I want to see the retrieved passage and the answer side by side.
- 03 **Alex:** Yes, the record is D-184. I checked the passage in that record: it says 30 days. The response underneath it says 60.
- 04 **Priya:** That gives us a specific ² _____. Retrieval found the current passage. I still need to inspect exactly what reached the model.
- 05 **Alex:** So the model got it wrong? I need to tell the team who is picking this up.
- 06 **Priya:** Tell them the answer is wrong, but leave the cause open. The ³ _____ could contain another instruction or an older example.
- 07 **Alex:** Wait, could the interface have changed the number? We had a display issue last time.
- 08 **Priya:** The recorded response already says 60, and the interface displays it unchanged. That puts the first observed difference ⁴ _____ of presentation.
- 09 **Alex:** What would you need to distinguish a bad instruction from an unstable model answer? We should keep the Friday demonstration team informed.
- 10 **Priya:** The configuration and a redacted ⁵ _____. Then I can repeat the request and compare the inputs without changing several variables.

Conversation continues on the next page.

The answer is wrong, but which layer failed?

CONVERSATION / CONTINUED

- 11 **Alex:** Is repeating one request enough to say we have fixed the problem? The same customer asked two slightly different versions of the question.
- 12 **Priya:** No. First we need a reliable ⁶ _____. Then we test the change on both variants and related policy cases.
- 13 **Alex:** Please make the ownership clear in the ticket. I can coordinate the demonstration, but I cannot approve production use on my own.
- 14 **Priya:** I will own technical ⁷ _____. I will also name the policy and interface owners if the evidence requires their input.
- 15 **Alex:** Can I tell the sales team we will have a fix by Thursday afternoon? They need to decide whether to keep the live demonstration.
- 16 **Priya:** I can promise findings by then, not a fix. Until we verify a ⁸ _____, that deadline would be an unsupported commitment.
- 17 **Alex:** All right. I will describe the customer-visible problem and keep the demonstration decision separate. What should your update include for that decision?
- 18 **Priya:** The reproduction result, affected cases, remaining uncertainty, and a proposed next step. I will preserve the ⁹ _____ so the comparison is meaningful.
- 19 **Alex:** That works. Please send it by three on Thursday, even if the investigation is incomplete. I will confirm who can decide about Friday.
- 20 **Priya:** Agreed. The ¹⁰ _____ will say what we know, what we do not know, and which decision still needs an owner.

CHECK THE GAPS

1. request ID / 2. mismatch / 3. payload / 4. upstream / 5. trace / 6. reproduction / 7. triage / 8. root cause / 9. configuration / 10. handoff

[Why these answers fit, page 111](#)

F Check accuracy and scope

The answer is wrong, but which layer failed?

Return to the case

A document assistant displays the wrong refund deadline. Retrieval returned the current policy, which states 30 days, but the generated answer says 60. The interface displays the generated text unchanged. Priya owns the model integration; Alex manages the product. A demonstration is scheduled for Friday. No cause has been confirmed, and Alex cannot approve a production release alone.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which statement preserves uncertainty?

- A. The trace suggests a generation issue; the cause is unconfirmed.
- B. The trace proves the engineer caused the defect.
- C. A correct passage guarantees a correct answer.
- D. The demonstration authorizes production use.

2. Which expression means transferring work with context?

- A. prepare a clear handoff
- B. increase the inference
- C. approve a symptom
- D. retrieve the ownership

[Answers and explanations, page 112](#)

Listen for the distinction

Switch roles. Ask for D-184, the redacted trace and the request configuration; retain the Thursday findings update without promising a fix.

Locate the failure before naming an owner

"The AI is broken" does not identify a component or establish responsibility. Separate the observed output from a theory about its cause. A colleague asking for the exact trace is asking for evidence, not necessarily dismissing the report.

F Say it / transfer to a new case

A missing citation in a different assistant

A support assistant generates a correct answer with a citation ID. The interface omits the citation link. The display team needs the exact recorded request, not an assumption that retrieval failed.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

symptom / user interface / request ID / handoff

- 01 **Engineer:** The model response contains the citation; the ¹ _____ does not display it.
- 02 **Lead:** Attach the ² _____ so the display team can locate that exchange.
- 03 **Engineer:** That is the observed ³ _____; we have not proved why it occurred.
- 04 **Lead:** Then send a clear ⁴ _____ to the display owner.

CHECK THE GAPS

1. user interface / 2. request ID / 3. symptom / 4. handoff

[Why these answers fit, page 113](#)

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "Please make the ownership clear in the ticket. I can coordinate the demonstration, but I cannot approve production use on my own." Priya: [missing response] Alex: "Can I tell the sales team we will have a fix by Thursday afternoon? They need to decide whether to keep the live demonstration." Which response belongs in the gap?

- A.** Agreed. The handoff will say what we know, what we do not know, and which decision still needs an owner.
- B.** Do you have the request ID? I want to see the retrieved passage and the answer side by side.
- C.** The configuration and a redacted trace. Then I can repeat the request and compare the inputs without changing several variables.
- D.** I will own technical triage. I will also name the policy and interface owners if the evidence requires their input.

[Answers and explanations, page 114](#)

A Read the situation

LESSON 02 / LLMS, TRANSFORMERS, TOKENS, AND CONTEXT

Correct a technical misconception tactfully and negotiate a bounded alternative.

A larger answer will not make the input fit

A document-summary request reserves 4,000 output tokens. Its instructions and documents contain 29,000 input tokens. The configured total context limit is 32,000 tokens, and the request is rejected before generation. Lina proposes increasing the output allowance. Omar owns the request configuration. The team must retain the exception clauses; no larger-context model has been evaluated for this task.

People: Lina, Application engineer; Omar, Model platform engineer.

Read the record

Choose one answer per question using the facts above.

1. What is the total requested reservation?

- A. 29,000 tokens
- B. 32,000 tokens
- C. 25,000 tokens
- D. 33,000 tokens

2. Why was the request rejected?

- A. The exception clauses were false.
- B. A larger model had already failed evaluation.
- C. It exceeds the configured total context limit.
- D. The generated answer was inaccurate.

3. What must the team preserve?

- A. A 4,000-token answer in every case
- B. The untested larger model
- C. The exception clauses
- D. Every duplicate header

[Answers and explanations, page 115](#)

B Vocabulary in use

CORE TERMS

LLM

Large language model; a model that processes and generates language.

evaluate an LLM

token

A unit processed by a model, not necessarily a whole word.

count input tokens

context window

The amount of context a model can handle in a request.

stay within the context window

output tokens

Tokens generated by the model.

reserve output tokens

system instruction

Higher-priority guidance supplied by an application's configuration.

inspect the system instruction

attention

A mechanism that weights relationships between representations.

compute attention weights

transformer

A neural architecture built around attention mechanisms.

use a transformer architecture

tokenizer

The component that converts text into tokens.

use the model's tokenizer

input tokens

Tokens supplied to the model before generation.

measure input tokens

token budget

An allocated number of tokens for a task or component.

allocate a token budget

prompt

Instructions and other context supplied for a model response.

revise the prompt

parameter

A learned numerical value within a model.

update model parameters

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

LLM LLM

evaluate an LLM

Keep the meaning precise: Large language model; a model that processes and generates language.

input tokens INPUT TOKENS

measure input tokens

Keep the meaning precise: Tokens supplied to the model before generation.

B Precision and professional shorthand

EXTEND YOUR RANGE

weight

A learned numerical coefficient in a model.

load model weights

truncation

Removal of content to fit a size limit.

detect silent truncation

reserved capacity

Capacity set aside rather than already consumed.

reduce reserved capacity

instruction hierarchy

Priority relationships among different instruction sources.

respect the instruction hierarchy

position

A token's place within the input sequence.

vary evidence position

deduplication

Removal of repeated copies of content.

apply exact deduplication

checkpoint

Saved model state from a particular stage.

select a checkpoint

compression

Reduction of information size, potentially losing detail.

assess compression losses

delimiter

A marker separating sections of supplied text.

preserve clear delimiters

long-context

Able to handle relatively large input context.

evaluate long-context performance

exception clause

Wording that limits or changes a general rule.

retain the exception clause

headroom

Unused capacity left below a limit.

leave sufficient headroom

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

parameter **PARAMETER**

update model parameters

Keep the meaning precise: A learned numerical value within a model.

reserved capacity **RESERVED CAPACITY**

reduce reserved capacity

Keep the meaning precise: Capacity set aside rather than already consumed.

Phrases that move the work forward

Restate a proposal

You are proposing a larger output allowance, correct?

Show the arithmetic

Twenty-nine thousand plus four thousand is thirty-three thousand.

Distinguish quantities

The allowance is a maximum, not the length already generated.

Offer an alternative

Could we remove exact duplicates before shortening substantive text?

Check counting

Are those counts from the tokenizer used by this model?

Check a transformation

Which information would compression remove?

Bound a comparison

Let us compare the same documents under both configurations.

Separate evaluation

The larger-context model needs its own quality check.

Locate the limit

The request fails before generation because the combined reservation is too large.

Correct tactfully

I see the aim, but that change would increase the reservation.

Preserve a requirement

We must retain the exception clauses in the source context.

State a condition

A smaller output allowance is acceptable only if the summary still meets the task.

Avoid a guarantee

Fitting within the limit does not prove the answer is complete.

Reject silent loss

Please report truncation instead of silently dropping the end.

Keep a fallback

If the exceptions cannot fit, route the request for a different workflow.

Read back the decision

We will remove duplicates, recount, and test coverage before changing models.

Precision in real exchanges

Purpose: Correct tactfully

Useful wording: I see the aim, but that change would increase the reservation.

Purpose: Distinguish quantities

Useful wording: The allowance is a maximum, not the length already generated.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which response explains the rejection?

- A.** The combined reservation is 1,000 tokens over the limit.
- B.** The output has already used 33,000 tokens.
- C.** The model generated an incorrect exception.
- D.** The tokenizer guarantees the summary is complete.

2. Which proposal preserves the stated requirement most directly?

- A.** Remove exact duplicate headers and recount.
- B.** Delete exception clauses first.
- C.** Increase the output reservation without recounting.
- D.** Treat words and tokens as equal.

3. A colleague needs the term for this meaning: The component that converts text into tokens. Which term fits?

- A.** system instruction
- B.** position
- C.** tokenizer
- D.** truncation

D Choose the response that fits / continued

4. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "All right, I had treated the limit as input only. Could we simply trim the final document until the request is accepted?" Omar: [missing response] Lina: "We also repeat the same header on every page. Removing those exact copies would preserve the policy wording and reduce the input somewhat." Which response belongs in the gap?

- A.** Exactly. Also record whether the client performs truncation. A request can fit after important text has been removed without anyone noticing.
- B.** Not without checking what it contains. The final pages include an exception clause that changes the refund rule for renewals.
- C.** Can we check the token budget first? Your input is 29,000 tokens, and you have reserved another 4,000 for output.
- D.** For this configuration, yes. The combined reservation is 33,000. Increasing the output allowance would move us further over the limit.

[Answers and explanations, page 115](#)

One useful distinction

A token is not a word. Input size, reserved output, and total context are different quantities. The numerical limits in this case are fictional configuration facts, not specifications for a named model.

E Conversation

A larger answer will not make the input fit

A larger answer will not make the input fit

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

long-context / token budget / exception clause / truncation / headroom / deduplication
/ position / delimiters / reservation / capacity constraint

- 01 **Lina:** The summary request is rejected again. I was going to increase the output allowance so the model has enough room to explain the policy.
- 02 **Omar:** Can we check the ¹ _____ first? Your input is 29,000 tokens, and you have reserved another 4,000 for output.
- 03 **Lina:** That sounds close to the 32,000 limit. Are you saying the unused output allowance counts even though the model has not answered yet?
- 04 **Omar:** For this configuration, yes. The combined ² _____ is 33,000. Increasing the output allowance would move us further over the limit.
- 05 **Lina:** All right, I had treated the limit as input only. Could we simply trim the final document until the request is accepted?
- 06 **Omar:** Not without checking what it contains. The final pages include an ³ _____ that changes the refund rule for renewals.
- 07 **Lina:** We also repeat the same header on every page. Removing those exact copies would preserve the policy wording and reduce the input somewhat.
- 08 **Omar:** That is a reasonable first test. Use exact ⁴ _____, then count again with the tokenizer used by this model rather than a word count.
- 09 **Lina:** Suppose that saves only 600 tokens. We would still be 400 over the limit. Could we reserve 3,000 output tokens instead?
- 10 **Omar:** We could test it. That would leave 600 tokens of ⁵ _____, but we need to verify that the shorter answer still covers the required points.

Conversation continues on the next page.

A larger answer will not make the input fit

CONVERSATION / CONTINUED

- 11 **Lina:** I'll check renewals and exceptions against the same checklist. Getting the request accepted won't help if the summary drops the important part.
- 12 **Omar:** Exactly. Also record whether the client performs ⁶ _____. A request can fit after important text has been removed without anyone noticing.
- 13 **Lina:** Someone suggested using a larger-context model instead. Would that let us keep the original request and stop spending time on these smaller adjustments?
- 14 **Omar:** It is another option, not an automatic solution. Its ⁷ _____ performance on our policy task has not been evaluated, and its cost may differ.
- 15 **Lina:** Then I will keep that as a separate comparison. Should the test vary where the exception appears, or is the current document order sufficient?
- 16 **Omar:** Vary its ⁸ _____ while keeping the wording unchanged. We want to see whether the model preserves the exception across the tested arrangements.
- 17 **Lina:** I'll remove the repeated headers first. If we're still over, I'll test the smaller output allowance and bring both the counts and the coverage results.
- 18 **Omar:** Correct. Keep the ⁹ _____ between documents so the revised input does not merge separate policies into one undifferentiated block.
- 19 **Lina:** I will report the before-and-after counts with the coverage results. If the necessary clauses still cannot fit, I will stop rather than silently remove them.
- 20 **Omar:** Good. Describe the limit as a ¹⁰ _____, not a model-quality finding. We can then choose the next workflow using the actual test results.

CHECK THE GAPS

1. token budget / 2. reservation / 3. exception clause / 4. deduplication / 5. headroom / 6. truncation / 7. long-context / 8. position / 9. delimiters / 10. capacity constraint

[Why these answers fit, page 116](#)

F Check accuracy and scope

A larger answer will not make the input fit

Return to the case

A document-summary request reserves 4,000 output tokens. Its instructions and documents contain 29,000 input tokens. The configured total context limit is 32,000 tokens, and the request is rejected before generation. Lina proposes increasing the output allowance. Omar owns the request configuration. The team must retain the exception clauses; no larger-context model has been evaluated for this task.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Complete: "We can reduce the allowance, provided that ____."

- A. the summary covering the required points
- B. cover the summary the required points
- C. the summary still covers the required points
- D. does the summary cover the required points

2. Which claim is appropriately limited?

- A. Removing duplicates proves the policy is current.
- B. The request fits; we still need to test summary coverage.
- C. A fitting request must produce a complete answer.
- D. More context eliminates all reasoning errors.

[Answers and explanations, page 118](#)

Listen for the distinction

Switch roles. Read the proposal to remove exact duplicates while retaining the exception clauses; do not treat a fitting request as proof of coverage.

Correct the model of the problem

Restate the proposal before rejecting it. Explain which limit applies and show the arithmetic. "That would increase the reservation" is more useful than "You do not understand tokens." A correction can be direct without judging the colleague.

F Say it / transfer to a new case

A contract summary with repeated appendices

A fictional request has a 20,000-token total limit, 18,000 input tokens, and a 3,000-token output reservation. Repeated appendix covers may be removed; unique obligations must remain.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

reservation / obligation / coverage / deduplication

- 01 **Engineer:** The combined ¹ _____ is 21,000 tokens, so the request is too large.
- 02 **Reviewer:** Start with exact ² _____ of the repeated covers.
- 03 **Engineer:** I will preserve every unique ³ _____ in the supplied contract.
- 04 **Reviewer:** After recounting, test ⁴ _____ rather than treating acceptance as success.

CHECK THE GAPS

1. reservation / 2. deduplication / 3. obligation / 4. coverage

[Why these answers fit, page 118](#)

F Say it / transfer to a new case / continued

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "Someone suggested using a larger-context model instead. Would that let us keep the original request and stop spending time on these smaller adjustments?" Omar: [missing response] Lina: "Then I will keep that as a separate comparison. Should the test vary where the exception appears, or is the current document order sufficient?" Which response belongs in the gap?

- A.** That is a reasonable first test. Use exact deduplication, then count again with the tokenizer used by this model rather than a word count.
- B.** Good. Describe the limit as a capacity constraint, not a model-quality finding. We can then choose the next workflow using the actual test results.
- C.** We could test it. That would leave 600 tokens of headroom, but we need to verify that the shorter answer still covers the required points.
- D.** It is another option, not an automatic solution. Its long-context performance on our policy task has not been evaluated, and its cost may differ.

[Answers and explanations, page 119](#)

Review 1 / Meaning and context

Use language from the lessons you have completed. Choose one answer per question.

Case file

Revisit "Speaking the AI Development Stack" and "LLMs, Transformers, Tokens, and Context". Choose precise terminology, then reconstruct the missing responses using both neighboring turns. The final two questions also revisit earlier conversations.

1. A colleague needs the term for this meaning: A model that processes and generates language sequences. Which term fits?

- A. language model
- B. payload
- C. root cause
- D. contract

2. A colleague needs the term for this meaning: An allocated number of tokens for a task or component. Which term fits?

- A. instruction hierarchy
- B. LLM
- C. token budget
- D. weight

3. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "What would you need to distinguish a bad instruction from an unstable model answer? We should keep the Friday demonstration team informed." Priya: [missing response] Alex: "Is repeating one request enough to say we have fixed the problem? The same customer asked two slightly different versions of the question." Which response belongs in the gap?

- A. The configuration and a redacted trace. Then I can repeat the request and compare the inputs without changing several variables.
- B. Do you have the request ID? I want to see the retrieved passage and the answer side by side.
- C. I will own technical triage. I will also name the policy and interface owners if the evidence requires their input.
- D. Agreed. The handoff will say what we know, what we do not know, and which decision still needs an owner.

Review 1 / Meaning and context / continued

4. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "I'll remove the repeated headers first. If we're still over, I'll test the smaller output allowance and bring both the counts and the coverage results." Omar: [missing response] Lina: "I will report the before-and-after counts with the coverage results. If the necessary clauses still cannot fit, I will stop rather than silently remove them." Which response belongs in the gap?

- A.** Can we check the token budget first? Your input is 29,000 tokens, and you have reserved another 4,000 for output.
- B.** Correct. Keep the delimiters between documents so the revised input does not merge separate policies into one undifferentiated block.
- C.** Not without checking what it contains. The final pages include an exception clause that changes the refund rule for renewals.
- D.** For this configuration, yes. The combined reservation is 33,000. Increasing the output allowance would move us further over the limit.

[Answers and explanations, page 150](#)

A Read the situation

LESSON 03 / DATA, DATASETS, LABELS, AND LEAKAGE

Disclose compromised evidence and resist pressure to overstate readiness.

A 94 percent score with an overlap problem

A classifier scores 94 percent on 500 evaluation examples. Jules finds that 40 evaluation records have exact copies in training. The remaining records have not been re-scored separately. Morgan presents release evidence in one hour. Training and evaluation labels use the same written rubric, but six disputed labels remain unresolved. The release owner can postpone the decision.

People: Jules, Data scientist; Morgan, Evaluation lead.

Read the record

Choose one answer per question using the facts above.

1. What does 94 percent currently describe?

- A. A verified production success rate
- B. The agreement rate on disputed labels
- C. The score on all 500 evaluation examples
- D. The score on 460 nonoverlapping examples

2. What overlap has been confirmed?

- A. Six hundred records are duplicates.
- B. Forty evaluation records have exact training copies.
- C. Every evaluation record is in training.
- D. Forty training labels are missing.

3. What can the release owner do?

- A. Turn the score into a production guarantee
- B. Postpone the decision
- C. Certify the unmeasured subset score
- D. Resolve labels without review

[Answers and explanations, page 120](#)

B Vocabulary in use

CORE TERMS

dataset

A collection of records assembled for a purpose.
curate a dataset

validation set

Data used to guide model selection or tuning.
monitor validation performance

holdout

Data kept separate from a fitting or selection process.
maintain a clean holdout

contamination

Unwanted overlap or exposure that weakens an evaluation.
quantify evaluation contamination

near duplicate

A closely similar record that may share substantive content.
review near duplicates

label

A target category or annotation attached to a record.
verify a disputed label

training set

Data used to fit model parameters.
audit the training set

test set

Data reserved for evaluating a finalized choice.
preserve the test set

data leakage

Information crossing a boundary in a way that invalidates evaluation.
investigate data leakage

exact duplicate

A record identical under a stated comparison rule.
remove exact duplicates

split

A partition of data into separate subsets.
define a reproducible split

annotation

Human or machine-added information about an example.
review annotation quality

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

dataset DATASET

curate a dataset

Keep the meaning precise: A collection of records assembled for a purpose.

data leakage DATA LEAKAGE

investigate data leakage

Keep the meaning precise: Information crossing a boundary in a way that invalidates evaluation.

B Precision and professional shorthand

EXTEND YOUR RANGE

rubric

Explicit criteria for assigning or evaluating labels.
apply a shared rubric

inter-rater agreement

The degree to which different raters agree.
measure inter-rater agreement

sampling frame

The population or list from which records are sampled.
define the sampling frame

class imbalance

Unequal frequencies among target categories.
check class imbalance

distribution shift

A difference between data distributions across settings.
detect distribution shift

denominator

The total count used as the base of a rate.
state the denominator

adjudication

Resolution of a disputed judgment by a defined process.
request label adjudication

provenance

The recorded origin and history of data.
establish data provenance

representativeness

How well a sample reflects the target population.
assess sample representativeness

stratification

Sampling or partitioning within defined groups.
use stratified sampling

deduplicate

Remove repeated records under a defined rule.
deduplicate before splitting

re-score

Calculate a score again under specified conditions.
re-score the clean subset

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

annotation ANNOTATION

review annotation quality

Keep the meaning precise: Human or machine-added information about an example.

sampling frame SAMPLING FRAME

define the sampling frame

Keep the meaning precise: The population or list from which records are sampled.

Phrases that move the work forward

Flag compromised evidence

The headline score includes records that overlap with training.

Protect uncertainty

We have not yet calculated the score on the remaining records.

Separate two problems

Data overlap and disputed labels need different checks.

Name the decision impact

The current score should not be presented as an independent estimate.

Clarify a duplicate

What comparison rule defines an exact match here?

Qualify a clean result

A clean split still needs a representativeness check.

State the denominator

Please report both the number correct and the number evaluated.

Keep the record

Retain the original score with its limitation and the corrected result.

Quantify the issue

We found 40 exact overlaps among 500 evaluation examples.

Reject unsupported subtraction

Removing eight percent of records does not tell us the new accuracy.

Request provenance

Can we trace when these records entered each split?

Offer a bounded update

I can provide the overlap list before the meeting.

Preserve the holdout

Do not tune against the replacement test set.

Check the rubric

Which label criterion produced the disagreement?

Describe a limitation

This sample underrepresents the rare class.

Escalate a decision

The release owner needs to decide whether to wait for the revised evidence.

Precision in real exchanges

Purpose: Reject unsupported subtraction

Useful wording: Removing eight percent of records does not tell us the new accuracy.

Purpose: Separate two problems

Useful wording: Data overlap and disputed labels need different checks.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which meeting statement is supported?

- A.** Every label is unreliable.
- B.** The model will fail in production.
- C.** The 94 percent score includes 40 training overlaps.
- D.** The clean-subset score is exactly 86 percent.

2. What does "not yet re-scored" mean?

- A.** The revised score must be lower.
- B.** The release has been canceled permanently.
- C.** The labels are now verified.
- D.** The revised calculation remains unfinished.

3. A colleague needs the term for this meaning: Data reserved for evaluating a finalized choice. Which term fits?

- A.** distribution shift
- B.** test set
- C.** inter-rater agreement
- D.** near duplicate

D Choose the response that fits / continued

4. Rebuild the exchange from "A 94 percent score with an overlap problem". Morgan: "Then I will remove the independent-evaluation label. Can you tell me whether the overlap came from the import process or the later cleanup?" Jules: [missing response] Morgan: "Send me that list. I'll leave the cause out of the slide for now. Is there anything else I need to flag?" Which response belongs in the gap?

- A.** Yes. I found 40 exact training matches inside the 500-example evaluation set. That creates contamination, so the independence claim needs correcting.
- B.** No. We need to re-score the remaining examples. We do not know how many of the overlapping records were answered correctly.
- C.** Not yet. The provenance log should help, but I have not reconstructed the sequence. I can provide the exact record list now.
- D.** Not from separation alone. We still need to check representativeness, including the rare category that appears only a few times in this sample.

[Answers and explanations, page 120](#)

One useful distinction

A clean holdout is separate from the process it evaluates, not necessarily representative of production. Removing overlap improves separation but does not resolve label disputes or distribution differences.

E Conversation

A 94 percent score with an overlap problem

A 94 percent score with an overlap problem

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

contamination / holdout / limitation statement / adjudication / re-score / model selection
/ exact duplicates / denominator / representativeness / provenance

- 01 **Morgan:** The release meeting starts in an hour. I have the 94 percent result on the slide. Is there anything material we need to add?
- 02 **Jules:** Yes. I found 40 exact training matches inside the 500-example evaluation set. That creates ¹ _____, so the independence claim needs correcting.
- 03 **Morgan:** Forty out of 500 is eight percent. Could I subtract eight points and present 86 percent as a conservative estimate instead?
- 04 **Jules:** No. We need to ² _____ the remaining examples. We do not know how many of the overlapping records were answered correctly.
- 05 **Morgan:** Then I will remove the independent-evaluation label. Can you tell me whether the overlap came from the import process or the later cleanup?
- 06 **Jules:** Not yet. The ³ _____ log should help, but I have not reconstructed the sequence. I can provide the exact record list now.
- 07 **Morgan:** Send me that list. I'll leave the cause out of the slide for now. Is there anything else I need to flag?
- 08 **Jules:** Six examples have disputed labels. They need ⁴ _____ under the written rubric, independently of the overlap check.
- 09 **Morgan:** Could we exclude those six for this meeting? I do not want the discussion to get buried in detail that may not change the decision.
- 10 **Jules:** We can report an explicitly defined subset, but we must state its ⁵ _____ and exclusions. We cannot call a selectively cleaned subset the original result.

Conversation continues on the next page.

A 94 percent score with an overlap problem

CONVERSATION / CONTINUED

- 11 **Morgan:** Agreed. The release owner can wait for better evidence. What would you recommend we say about the 94 percent figure already circulated yesterday?
- 12 **Jules:** Retain it with a correction: it describes the original set, not a clean
6 _____. Replace the readiness claim rather than quietly changing the number.
- 13 **Morgan:** Once the duplicate records are removed, can we say the evaluation represents customer traffic? That is the next question I expect from the release owner.
- 14 **Jules:** Not from separation alone. We still need to check 7 _____, including the rare category that appears only a few times in this sample.
- 15 **Morgan:** So we have three issues: overlap, disputed labels, and coverage of the target population. Which one can you resolve before today's meeting starts?
- 16 **Jules:** I can verify the 8 _____ and document the matching rule. I cannot responsibly promise all label reviews within the hour.
- 17 **Morgan:** All right, the duplicate check comes first. For the replacement set, can we keep it untouched while the team tunes the model?
- 18 **Jules:** It should. If we repeatedly choose model settings from that set, it becomes part of
9 _____, and we need a separate final test.
- 19 **Morgan:** I will recommend postponing the readiness decision. The slide will list confirmed overlap, pending label review, and the missing clean-subset score separately.
- 20 **Jules:** I will send the records and a short 10 _____. That gives the owner usable evidence without pretending the corrected outcome is already known.

CHECK THE GAPS

1. contamination / 2. re-score / 3. provenance / 4. adjudication / 5. denominator / 6. holdout / 7. representativeness / 8. exact duplicates / 9. model selection / 10. limitation statement

[Why these answers fit, page 121](#)

F Check accuracy and scope

A 94 percent score with an overlap problem

Return to the case

A classifier scores 94 percent on 500 evaluation examples. Jules finds that 40 evaluation records have exact copies in training. The remaining records have not been re-scored separately. Morgan presents release evidence in one hour. Training and evaluation labels use the same written rubric, but six disputed labels remain unresolved. The release owner can postpone the decision.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which question targets a label dispute?

- A. Which rubric criterion distinguishes these two labels?
- B. Can we use the model's own prediction as the reference label?
- C. Does removing duplicates prove the disputed labels are correct?
- D. Can the headline score decide which reviewer is right?

2. Which phrase correctly names the base of a rate?

- A. state the denominator
- B. approve the leakage
- C. deploy the annotation
- D. infer the rubric

[Answers and explanations, page 123](#)

Listen for the distinction

Switch roles. Separate the overlap list, unmeasured clean-subset score and six disputed labels in the handoff.

Report evidence quality before headline performance

Flagging contaminated evaluation data is responsible reporting, not an admission that every model result is worthless. State exactly what is affected, what cannot yet be concluded, and which decision needs a revised evidence base.

F Say it / transfer to a new case

A sentiment test with repeated customer messages

An 80-message test contains ten copies of messages used for training. The team has not scored the 70 remaining messages separately. Two reviewers disagree on three labels.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

denominator / adjudication / overlap / re-score

- 01 **Analyst:** The current test has training ¹ _____, so its independence is compromised.
- 02 **Lead:** Report the clean-subset score only after you ² _____ those 70 messages.
- 03 **Analyst:** The three disputed labels also need ³ _____.
- 04 **Lead:** And include the ⁴ _____ with every reported percentage.

CHECK THE GAPS

1. overlap / 2. re-score / 3. adjudication / 4. denominator

[Why these answers fit, page 123](#)

F Say it / transfer to a new case / continued

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "A 94 percent score with an overlap problem". Morgan: "Once the duplicate records are removed, can we say the evaluation represents customer traffic? That is the next question I expect from the release owner." Jules: [missing response] Morgan: "So we have three issues: overlap, disputed labels, and coverage of the target population. Which one can you resolve before today's meeting starts?" Which response belongs in the gap?

- A.** Not yet. The provenance log should help, but I have not reconstructed the sequence. I can provide the exact record list now.
- B.** It should. If we repeatedly choose model settings from that set, it becomes part of model selection, and we need a separate final test.
- C.** Retain it with a correction: it describes the original set, not a clean holdout. Replace the readiness claim rather than quietly changing the number.
- D.** Not from separation alone. We still need to check representativeness, including the rare category that appears only a few times in this sample.

[Answers and explanations, page 124](#)

A Read the situation

LESSON 04 / RETRIEVAL, EMBEDDINGS, VECTOR SEARCH, AND RAG

Negotiate a trade-off using matched comparisons and explicitly limited claims.

More relevant passages, a slower response

A document assistant with reranking retrieves a relevant passage for 86 of 100 fixed test queries, compared with 78 without reranking. The reranker adds 300 milliseconds to the measured retrieval path. The generator and query set are unchanged. No user study or end-to-end answer-quality review has been completed. Sam recommends a pilot; Elena owns the evaluation plan.

People: Sam, Retrieval engineer; Elena, Evaluation lead.

Read the record

Choose one answer per question using the facts above.

1. What improved in the fixed test?

- A. Queries with a relevant retrieved passage rose from 78 to 86.
- B. Every final answer became correct.
- C. Response time fell by 300 milliseconds.
- D. Users preferred the new experience.

2. Which component stayed unchanged?

- A. The reranker
- B. The retrieval path timing
- C. The number of relevant hits
- D. The generator

3. Which claim needs further evidence?

- A. The reranker adds 300 milliseconds.
- B. Both variants use the same query set.
- C. The test contains 100 queries.
- D. The new path improves end-to-end answer quality.

[Answers and explanations, page 125](#)

B Vocabulary in use

CORE TERMS

RAG

Retrieval-augmented generation; generating with retrieved context.

evaluate a RAG pipeline

vector

An ordered collection of numerical values.

compare document vectors

similarity search

Finding items close under a chosen similarity measure.

run similarity search

keyword search

Retrieval based on words or lexical matches.

combine keyword search with vectors

chunk

A retrievable segment of a larger document.

retrieve a relevant chunk

overlap window

Content shared between adjacent chunks.

adjust the overlap window

embedding

A numerical representation used to compare or process items.

compute an embedding

vector store

A store or index supporting retrieval of vector representations.

query a vector store

semantic search

Retrieval based on meaning-related representations.

assess semantic search

hybrid retrieval

Retrieval combining different search approaches.

test hybrid retrieval

chunk boundary

The point where one segment ends and another starts.

preserve context at chunk boundaries

reranker

A component that reorders candidates for relevance.

evaluate the reranker

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

RAG RAG

evaluate a RAG pipeline

Keep the meaning precise: Retrieval-augmented generation; generating with retrieved context.

semantic search SEMANTIC SEARCH

assess semantic search

Keep the meaning precise: Retrieval based on meaning-related representations.

B Precision and professional shorthand

EXTEND YOUR RANGE

candidate set

Items available for a later selection or ranking step.

inspect the candidate set

relevance

How well material addresses a specified query.

judge passage relevance

precision

The proportion of retrieved items judged relevant.

improve retrieval precision

citation

A reference identifying a supporting source.

verify a citation

index freshness

How current the indexed information is.

monitor index freshness

ablation

A comparison that removes or changes a component.

run a reranker ablation

top-k

The first k results returned by a ranking.

compare top-k results

recall

The proportion of relevant items retrieved under a defined evaluation.

measure retrieval recall

grounding

Connecting generated claims to supporting information.

check answer grounding

source attribution

Identification of where information came from.

preserve source attribution

filter

A rule limiting eligible retrieval results.

apply a metadata filter

end-to-end

Covering the complete path from request to outcome.

measure end-to-end latency

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

reranker **RERANKER**

evaluate the reranker

Keep the meaning precise: A component that reorders candidates for relevance.

precision **PRECISION**

improve retrieval precision

Keep the meaning precise: The proportion of retrieved items judged relevant.

Phrases that move the work forward

Name the measure

The improvement is passage coverage on this fixed query set.

Quantify a trade-off

The gain comes with 300 milliseconds of added retrieval time.

Ask for an ablation

What happens when we remove only the reranker?

Check the source

Does the cited passage support this specific sentence?

Inspect a miss

Was the relevant passage absent or merely ranked too low?

Preserve context

The exception and the main rule should not be separated without context.

Distinguish freshness

A relevant passage may still come from an outdated index.

Compare alternatives

Let us compare reranking all queries with a defined routing policy.

Separate outcomes

We have not yet measured final-answer accuracy.

Request a matched comparison

Can we hold the generator and queries constant?

Limit a recommendation

These results support a bounded pilot, not a general rollout claim.

Define the next test

Review final answers and timing on the same cases.

Check eligibility

Did a filter exclude the document before similarity search?

Challenge an average

Which query types account for most of the added time?

Avoid causal overreach

The matched test isolates this change, not every production condition.

Record a limitation

User preference remains unmeasured in this experiment.

Precision in real exchanges

Purpose: Request a matched comparison

Useful wording: Can we hold the generator and queries constant?

Purpose: Ask for an ablation

Useful wording: What happens when we remove only the reranker?

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which comparison correctly states the observed gain?

- A.** Every query returns eight more passages.
- B.** Coverage increased by eight percentage points on the fixed set.
- C.** Users became eight percent more satisfied.
- D.** Final answers are now 86 percent accurate.

2. What is a reranker ablation here?

- A.** Remove all supporting documents permanently.
- B.** Compare the path with and without the reranker while holding other factors fixed.
- C.** Replace every component and change the queries.
- D.** Ask users to invent new relevance criteria.

3. A colleague needs the term for this meaning: A store or index supporting retrieval of vector representations. Which term fits?

- A.** relevance
- B.** chunk
- C.** vector store
- D.** index freshness

D Choose the response that fits / continued

4. Rebuild the exchange from "More relevant passages, a slower response". Sam: "I agree the answer review is unfinished. My recommendation is a small pilot, not a rollout. Which evidence would you need before approving that pilot?" Elena: [missing response] Sam: "We can collect those. Some of the original misses had the right document in the candidates but ranked below the five passages sent to generation." Which response belongs in the gap?

- A.** Exactly. Also break down timing by query type, because the same average can hide very different costs for simple and complex questions.
- B.** First, an end-to-end review of the same cases. We need the final answers, their supporting passages, and the complete response times.
- C.** That distinction matters. A reranker can promote those passages, but it cannot rescue a relevant document that never entered the candidate set.
- D.** Good. Before we call it better overall, what changed in the candidate set? Did you also modify the initial search or document filters?

[Answers and explanations, page 125](#)

One useful distinction

Retrieval relevance, answer correctness, and user preference are distinct outcomes. A citation can identify a source without proving that the source supports the exact claim. Inspect the cited passage.

E Conversation

More relevant passages, a slower response

More relevant passages, a slower response

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

reranker / trade-off / index freshness / end-to-end / conditional / grounding / ablation
/ query type / chunk boundary / candidate set

- 01 **Sam:** The reranker looks promising. We found a relevant passage for 86 of the same 100 queries, compared with 78 on the previous path.
- 02 **Elena:** Good. Before we call it better overall, what changed in the ¹ _____? Did you also modify the initial search or document filters?
- 03 **Sam:** No. The initial candidates, generator, and queries are unchanged. The reranker changes their order, and the retrieval path takes another 300 milliseconds.
- 04 **Elena:** Then we have a useful ² _____. It shows a retrieval gain with added time, but it does not yet show better generated answers.
- 05 **Sam:** I agree the answer review is unfinished. My recommendation is a small pilot, not a rollout. Which evidence would you need before approving that pilot?
- 06 **Elena:** First, an ³ _____ review of the same cases. We need the final answers, their supporting passages, and the complete response times.
- 07 **Sam:** We can collect those. Some of the original misses had the right document in the candidates but ranked below the five passages sent to generation.
- 08 **Elena:** That distinction matters. A ⁴ _____ can promote those passages, but it cannot rescue a relevant document that never entered the candidate set.
- 09 **Sam:** Two misses also involve a policy exception split across adjacent chunks. I do not want to alter chunking in the middle of this comparison.
- 10 **Elena:** Keep that separate. Log the ⁵ _____ issue for another experiment so we can still interpret this result.

Conversation continues on the next page.

More relevant passages, a slower response

CONVERSATION / CONTINUED

- 11 **Sam:** Should the answer reviewers count every displayed citation as a success? The new path tends to produce more citations, which looks reassuring in the interface.
- 12 **Elena:** No. They should check ⁶ _____ for each substantive claim. A citation can point to a real document that does not support that sentence.
- 13 **Sam:** Understood. I will include the exact cited passage and the answer claim together. We also need to distinguish an old policy from an irrelevant one.
- 14 **Elena:** Yes. Check ⁷ _____ separately. A highly relevant passage from a superseded policy can still produce a misleading answer.
- 15 **Sam:** The product lead will ask whether customers will accept the extra delay. Could we infer that from the eight-point increase in passage coverage?
- 16 **Elena:** No. That is a ⁸ _____ we have measured technically, but customer preference remains unknown until we test it appropriately.
- 17 **Sam:** Then I will present the retrieval result as eight percentage points on this set. I will not call it an eight-percent increase in customer satisfaction.
- 18 **Elena:** Exactly. Also break down timing by ⁹ _____, because the same average can hide very different costs for simple and complex questions.
- 19 **Sam:** I'll bring the answer review and timing breakdown before we discuss the pilot. The chunking experiment can wait; I don't want to lose this comparison.
- 20 **Elena:** Good. Put the improvement on the slide, but keep the pilot recommendation ¹⁰ _____. I want to see the actual answers before I approve it.

CHECK THE GAPS

1. candidate set / 2. ablation / 3. end-to-end / 4. reranker / 5. chunk boundary / 6. grounding / 7. index freshness / 8. trade-off / 9. query type / 10. conditional

[Why these answers fit, page 126](#)

F Check accuracy and scope

More relevant passages, a slower response

Return to the case

A document assistant with reranking retrieves a relevant passage for 86 of 100 fixed test queries, compared with 78 without reranking. The reranker adds 300 milliseconds to the measured retrieval path. The generator and query set are unchanged. No user study or end-to-end answer-quality review has been completed. Sam recommends a pilot; Elena owns the evaluation plan.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which sentence preserves the limitation?

- A. Better retrieval guarantees better answers.
- B. A citation proves the answer is true.
- C. No user study means retrieval did not improve.
- D. Retrieval improved on this set; final-answer quality remains unmeasured.

2. Which question checks grounding?

- A. Is the cited document present in the current index?
- B. Does the retrieved passage share many words with the question?
- C. Does this passage support the answer's refund exception?
- D. Does the answer contain a citation marker in the required format?

[Answers and explanations, page 128](#)

Listen for the distinction

Switch roles. Request a matched final-answer review; do not substitute retrieval coverage for answer accuracy.

Define better before agreeing

"Better retrieval" may mean finding relevant passages, ranking them earlier, or improving final answers. Ask which measure changed. A measured technical improvement can justify another test without proving a better customer experience.

F Say it / transfer to a new case

An answer cites the right document but the wrong section

A benefits assistant cites the current policy but uses the standard rule instead of a stated exception. Retrieval includes both sections. A reviewer must report claim support, not count citation markers.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

context / grounding / end-to-end / source

- 01 **Reviewer:** The citation identifies a real ¹ _____, but not support for this claim.
- 02 **Engineer:** Both sections were retrieved, so we should inspect answer ² _____.
- 03 **Reviewer:** The exception is missing from the answer, although it is present in the ³ _____.
- 04 **Engineer:** We will report that distinction in the ⁴ _____ review.

CHECK THE GAPS

1. source / 2. grounding / 3. context / 4. end-to-end

[Why these answers fit, page 128](#)

F Say it / transfer to a new case / continued

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Understood. I will include the exact cited passage and the answer claim together. We also need to distinguish an old policy from an irrelevant one." Elena: [missing response] Sam: "The product lead will ask whether customers will accept the extra delay. Could we infer that from the eight-point increase in passage coverage?" Which response belongs in the gap?

- A.** Yes. Check index freshness separately. A highly relevant passage from a superseded policy can still produce a misleading answer.
- B.** No. That is a trade-off we have measured technically, but customer preference remains unknown until we test it appropriately.
- C.** First, an end-to-end review of the same cases. We need the final answers, their supporting passages, and the complete response times.
- D.** Keep that separate. Log the chunk boundary issue for another experiment so we can still interpret this result.

[Answers and explanations, page 129](#)

Review 2 / Meaning and context

Use language from the lessons you have completed. Choose one answer per question.

Case file

Revisit "Data, Datasets, Labels, and Leakage" and "Retrieval, Embeddings, Vector Search, and RAG". Choose precise terminology, then reconstruct the missing responses using both neighboring turns. The final two questions also revisit earlier conversations.

1. A colleague needs the term for this meaning: Data used to guide model selection or tuning. Which term fits?

- A. adjudication
- B. stratification
- C. validation set
- D. exact duplicate

2. A colleague needs the term for this meaning: Retrieval combining different search approaches. Which term fits?

- A. citation
- B. candidate set
- C. RAG
- D. hybrid retrieval

3. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "Suppose that saves only 600 tokens. We would still be 400 over the limit. Could we reserve 3,000 output tokens instead?" Omar: [missing response] Lina: "I'll check renewals and exceptions against the same checklist. Getting the request accepted won't help if the summary drops the important part." Which response belongs in the gap?

- A. It is another option, not an automatic solution. Its long-context performance on our policy task has not been evaluated, and its cost may differ.
- B. We could test it. That would leave 600 tokens of headroom, but we need to verify that the shorter answer still covers the required points.
- C. That is a reasonable first test. Use exact deduplication, then count again with the tokenizer used by this model rather than a word count.
- D. Good. Describe the limit as a capacity constraint, not a model-quality finding. We can then choose the next workflow using the actual test results.

Review 2 / Meaning and context / continued

4. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Then I will present the retrieval result as eight percentage points on this set. I will not call it an eight-percent increase in customer satisfaction." Elena: [missing response] Sam: "I'll bring the answer review and timing breakdown before we discuss the pilot. The chunking experiment can wait; I don't want to lose this comparison." Which response belongs in the gap?

- A.** Good. Before we call it better overall, what changed in the candidate set? Did you also modify the initial search or document filters?
- B.** Exactly. Also break down timing by query type, because the same average can hide very different costs for simple and complex questions.
- C.** Then we have a useful ablation. It shows a retrieval gain with added time, but it does not yet show better generated answers.
- D.** That distinction matters. A reranker can promote those passages, but it cannot rescue a relevant document that never entered the candidate set.

[Answers and explanations, page 153](#)

A Read the situation

LESSON 05 / FINE-TUNING, ALIGNMENT, AND ADAPTATION

Separate proposed remedies by mechanism and agree on a controlled experiment.

One proposal, two different failure modes

A support summarizer sometimes returns paragraphs instead of the required fields and sometimes cites an obsolete refund policy. The obsolete wording is present in its retrieval index. Its prompt contains only one formatting example. Tessa proposes fine-tuning to solve both problems. Dev owns the experiment plan. The team has 200 permission-cleared example summaries but has not audited their label consistency.

People: Tessa, Applied scientist; Dev, Technical lead.

Read the record

Choose one answer per question using the facts above.

1. Which fact directly concerns outdated information?

- A. The retrieval index contains obsolete policy wording.
- B. The prompt contains one formatting example.
- C. The team owns an experiment plan.
- D. The output sometimes uses paragraphs.

2. What is known about the 200 examples?

- A. Every label has been independently verified.
- B. They contain the latest policy by definition.
- C. They guarantee successful fine-tuning.
- D. They are permission-cleared but consistency is unaudited.

3. What distinction should the experiment preserve?

- A. Data-use permission as proof of label consistency
- B. Required fields as proof that policy statements are current
- C. More training examples as proof that retrieval is refreshed
- D. Source freshness versus output-format behavior

[Answers and explanations, page 130](#)

B Vocabulary in use

CORE TERMS

fine-tuning

Updating model parameters using additional training data.

evaluate a fine-tuning run

supervised fine-tuning

Training on supplied input-output examples.

prepare supervised fine-tuning data

preference data

Examples recording preferences between candidate outputs.

audit preference data

RLHF

Reinforcement learning from human feedback.

discuss an RLHF workflow

LoRA

Low-rank adaptation; a parameter-efficient adaptation method.

train a LoRA adapter

base model

The starting model before the adaptation under discussion.

compare against the base model

pretraining

Broad initial training before task-specific adaptation.

distinguish pretraining from adaptation

SFT

Abbreviation for supervised fine-tuning.

compare SFT with prompting

alignment

Shaping behavior toward specified goals or preferences.

define the alignment objective

DPO

Direct preference optimization; a preference-training method.

evaluate a DPO experiment

adapter

A small trainable component used for adaptation.

load a task adapter

training objective

The quantity a training process seeks to optimize.

specify the training objective

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

fine-tuning FINE-TUNING

evaluate a fine-tuning run

Keep the meaning precise: Updating model parameters using additional training data.

alignment ALIGNMENT

define the alignment objective

Keep the meaning precise: Shaping behavior toward specified goals or preferences.

B Precision and professional shorthand

EXTEND YOUR RANGE

hyperparameter

A training or model setting selected rather than learned in that run.

document a hyperparameter choice

epoch

One pass through the training data.

record the epoch count

generalization

Performance on relevant examples beyond those used for fitting.

test generalization

format adherence

Consistency with a required output structure.

measure format adherence

few-shot prompting

Supplying a small set of examples in the prompt.

test few-shot prompting

training artifact

A saved output of training, such as weights or configuration.

version the training artifact

learning rate

A setting controlling the scale of training updates.

compare learning-rate settings

overfitting

Fitting training patterns without adequate generalization.

monitor for overfitting

catastrophic forgetting

Loss of previously learned capabilities during further training.

assess catastrophic forgetting

structured output

Output constrained or organized into a defined structure.

validate structured output

behavioral target

The specified behavior an intervention aims to change.

define the behavioral target

rollback candidate

A previously validated configuration available for reversion.

retain a rollback candidate

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

training objective TRAINING OBJECTIVE

specify the training objective

Keep the meaning precise: The quantity a training process seeks to optimize.

generalization GENERALIZATION

test generalization

Keep the meaning precise: Performance on relevant examples beyond those used for fitting.

Phrases that move the work forward

Separate failures

We have a formatting issue and a source-freshness issue.

Define a target

The target is consistent output in the required fields.

Check examples

Are the reference summaries consistent with the same rubric?

Hold a variable fixed

Use the same cleaned source set for both configurations.

Protect evaluation

Keep the final test examples outside the training set.

Ask for permission status

Are these examples cleared for the intended training use?

Check wider effects

Does the adaptation reduce performance on other required tasks?

Bound a recommendation

I recommend an experiment, not immediate adoption.

Ask about mechanism

How would this change prevent the obsolete policy from entering the context?

Propose a baseline

Let us test clearer prompting before attributing gains to training.

Limit a remedy

Fine-tuning may improve the behavior; it does not keep an index current.

Agree on evidence

We will compare format adherence and factual correctness separately.

State an alternative

A constrained output format may address part of this problem.

Avoid conflation

Permission, quality, and representativeness are separate checks.

Keep a baseline

Retain the validated base configuration for comparison.

Record the decision

We will fix the stale index and test the formatting intervention independently.

Precision in real exchanges

Purpose: Propose a baseline

Useful wording: Let us test clearer prompting before attributing gains to training.

Purpose: Check examples

Useful wording: Are the reference summaries consistent with the same rubric?

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which action directly addresses the known stale source?

- A.** Assume format training refreshes all documents.
- B.** Rename the output fields.
- C.** Increase the number of training epochs blindly.
- D.** Correct the retrieval index through the approved update process.

2. Which is a behavioral target?

- A.** Owning a permission-cleared file
- B.** Scheduling a review meeting
- C.** Having a larger office
- D.** Returning the required fields consistently

3. A colleague needs the term for this meaning: Abbreviation for supervised fine-tuning. Which term fits?

- A.** few-shot prompting
- B.** epoch
- C.** SFT
- D.** LoRA

D Choose the response that fits / continued

4. Rebuild the exchange from "One proposal, two different failure modes". Tessa: "For formatting, we have 200 example summaries cleared for training use. I can start preparing them, although we have not audited their labels yet." Dev: [missing response] Tessa: "Would you want a prompting comparison before training? The current prompt includes only one example, and it does not show how to handle an absent field." Which response belongs in the gap?

- A.** Please do that audit first. Permission to use an example does not establish its consistency with the field schema or the instructions in the other examples.
- B.** Exactly. Keep the source set fixed within the comparison, and measure format adherence separately from policy correctness. An answer can pass one and fail the other.
- C.** Include those. Further training can affect existing behavior, including catastrophic forgetting. We should assess relevant capabilities rather than assuming a small adaptation has no wider effects.
- D.** Consistency is worth testing, but those are different failure modes. What would the training change in the source material supplied to the model?

[Answers and explanations, page 130](#)

One useful distinction

Fine-tuning changes model parameters; prompting supplies instructions at inference time; retrieval supplies external context. These interventions can interact, but they do not have interchangeable mechanisms or guarantees.

E Conversation

One proposal, two different failure modes

One proposal, two different failure modes

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

failure modes / format adherence / rollback candidate / consistency / few-shot / behavioral target / schema / generalization / hypotheses / catastrophic forgetting

- 01 **Tessa:** I think we should fine-tune the summarizer. It ignores our field structure and sometimes quotes the old refund policy. Training could make it more consistent.
- 02 **Dev:** Consistency is worth testing, but those are different ¹ _____. What would the training change in the source material supplied to the model?
- 03 **Tessa:** It would not update the retrieval index. I checked this morning, and the obsolete refund paragraph is still there. That source needs a separate correction.
- 04 **Dev:** Right. Let us define the formatting problem as a ² _____, then keep policy freshness as a data-maintenance issue with its own owner.
- 05 **Tessa:** For formatting, we have 200 example summaries cleared for training use. I can start preparing them, although we have not audited their labels yet.
- 06 **Dev:** Please do that audit first. Permission to use an example does not establish its ³ _____ with the field schema or the instructions in the other examples.
- 07 **Tessa:** Would you want a prompting comparison before training? The current prompt includes only one example, and it does not show how to handle an absent field.
- 08 **Dev:** Yes. A stronger ⁴ _____ baseline and a structured-output option would tell us whether training adds value beyond clearer task specification.
- 09 **Tessa:** All right. Two work items, then: repair the index and compare formatting methods. I'll keep their results separate.
- 10 **Dev:** Exactly. Keep the source set fixed within the comparison, and measure ⁵ _____ separately from policy correctness. An answer can pass one and fail the other.

Conversation continues on the next page.

One proposal, two different failure modes

CONVERSATION / CONTINUED

- 11 **Tessa:** Should the training objective reward the presence of every field, even if the source gives no evidence for a value? That could encourage invented details.
- 12 **Dev:** No. The examples need an agreed representation for missing information. A complete ⁶ _____ does not mean every field contains a positive factual assertion.
- 13 **Tessa:** I will include that distinction in the rubric. After we select the method, we should evaluate on cases that were never used to tune it.
- 14 **Dev:** Correct. We need evidence of ⁷ _____, including summaries with exceptions, missing details, and layouts absent from the training examples.
- 15 **Tessa:** If we test an adapter, I also want to check our other summary tasks. Better performance on refunds could conceal a regression in cancellation summaries.
- 16 **Dev:** Include those. Further training can affect existing behavior, including ⁸ _____. We should assess relevant capabilities rather than assuming a small adaptation has no wider effects.
- 17 **Tessa:** I will version the examples, configuration, and resulting adapter together. That way the team can reproduce the comparison and identify exactly what was evaluated.
- 18 **Dev:** Good. Retain a validated ⁹ _____ as well. A successful experiment is not a reason to lose the previous working configuration.
- 19 **Tessa:** I'll start with the example audit and the prompting baseline. Can we review those results before we spend time on an adapter?
- 20 **Dev:** Yes. List the expected gains as ¹⁰ _____ in the experiment note. If the baseline solves the problem, we may not need training.

CHECK THE GAPS

1. failure modes / 2. behavioral target / 3. consistency / 4. few-shot / 5. format adherence / 6. schema / 7. generalization / 8. catastrophic forgetting / 9. rollback candidate / 10. hypotheses

[Why these answers fit, page 131](#)

F Check accuracy and scope

One proposal, two different failure modes

Return to the case

A support summarizer sometimes returns paragraphs instead of the required fields and sometimes cites an obsolete refund policy. The obsolete wording is present in its retrieval index. Its prompt contains only one formatting example. Tessa proposes fine-tuning to solve both problems. Dev owns the experiment plan. The team has 200 permission-cleared example summaries but has not audited their label consistency.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which claim needs qualification?

- A. Fine-tuning will solve both failures.
- B. The prompt has one example.
- C. The index contains obsolete wording.
- D. The label audit is unfinished.

2. Which comparison best isolates the formatting intervention?

- A. Evaluate only the training examples.
- B. Accept any output with a confident tone.
- C. Keep sources and test cases fixed while changing the formatting method.
- D. Change data, model, metrics, and test cases together.

[Answers and explanations, page 133](#)

Listen for the distinction

Switch roles. Read the proposed experiment with fixed sources and test cases; keep data-use permission separate from the unfinished label audit.

Challenge the scope of a remedy

A method can be useful without solving every problem in the discussion. Separate knowledge freshness from output behavior. Ask what mechanism would change each failure and what simpler comparison could test the claim.

F Say it / transfer to a new case

A classifier uses an outdated category list

A classifier is required to use five current categories. Its prompt still lists six older categories. The team has not yet tested a corrected prompt or new training data.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

category list / held-out / baseline / hypothesis

- 01 **Scientist:** The prompt supplies an obsolete ¹ _____, so we should correct that input first.
- 02 **Lead:** Then compare the corrected prompt with the existing ² _____.
- 03 **Scientist:** We will measure adherence on ³ _____ examples, not only the demonstrations.
- 04 **Lead:** Until then, the benefit of further training remains a ⁴ _____.

CHECK THE GAPS

1. category list / 2. baseline / 3. held-out / 4. hypothesis

[Why these answers fit, page 133](#)

F Say it / transfer to a new case / continued

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "One proposal, two different failure modes". Tessa: "I will include that distinction in the rubric. After we select the method, we should evaluate on cases that were never used to tune it." Dev: [missing response] Tessa: "If we test an adapter, I also want to check our other summary tasks. Better performance on refunds could conceal a regression in cancellation summaries." Which response belongs in the gap?

- A.** Yes. List the expected gains as hypotheses in the experiment note. If the baseline solves the problem, we may not need training.
- B.** Yes. A stronger few-shot baseline and a structured-output option would tell us whether training adds value beyond clearer task specification.
- C.** Good. Retain a validated rollback candidate as well. A successful experiment is not a reason to lose the previous working configuration.
- D.** Correct. We need evidence of generalization, including summaries with exceptions, missing details, and layouts absent from the training examples.

[Answers and explanations, page 134](#)

A Read the situation

LESSON 06 / EVALUATION, BENCHMARKS, AND REGRESSION

Present disaggregated results and challenge an oversimplified release argument.

The average is down, but one slice improved

On the same 100-case test, the previous model passes 90 cases and the candidate passes 88. Among 20 long-document cases, passes rise from 12 to 18. Among 80 short-request cases, passes fall from 78 to 70. The scoring rubric is unchanged. Production traffic proportions and the seriousness of individual failures have not been analyzed. Noor chairs the review; Ben prepared the results.

People: Noor, Evaluation manager; Ben, Machine-learning engineer.

Read the record

Choose one answer per question using the facts above.

1. What happens to total passes?

- A.** They stay at 90.
- B.** They rise from 12 to 18 overall.
- C.** They fall from 90 to 88.
- D.** They rise from 88 to 90.

2. Which slice improves?

- A.** Long-document cases
- B.** Short-request cases
- C.** Every production request
- D.** All failure severities

3. What remains unknown?

- A.** Whether the rubric changed
- B.** The candidate's total passes
- C.** The production mixture and seriousness of failures
- D.** The test-set size

[Answers and explanations, page 135](#)

B Vocabulary in use

CORE TERMS

evaluation

A structured assessment against specified criteria.

design an evaluation

baseline

The reference system or result for a comparison.

establish a baseline

regression

A deterioration in previously measured behavior.

investigate a regression

golden set

Curated cases used to check important behavior repeatedly.

maintain a golden set

acceptance criterion

A stated condition required for acceptance.

agree on acceptance criteria

aggregate

A combined measure across multiple cases or groups.

report the aggregate score

benchmark

A reference test used for comparison.

interpret a benchmark result

candidate

The version under consideration.

evaluate the candidate

test case

A defined input and expected criterion or outcome.

inspect a failed test case

pass rate

The fraction of evaluated cases meeting a criterion.

report the pass rate

slice

A defined subgroup within evaluation data.

analyze a performance slice

disaggregation

Breaking a combined result into meaningful groups.

disaggregate by task type

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

evaluation EVALUATION

design an evaluation

Keep the meaning precise: A structured assessment against specified criteria.

test case TEST CASE

inspect a failed test case

Keep the meaning precise: A defined input and expected criterion or outcome.

B Precision and professional shorthand

EXTEND YOUR RANGE

severity

The seriousness of a failure's consequences.
assess failure severity

false negative

A negative decision when the reference outcome is positive.
review false negatives

calibration

Agreement between stated confidence and observed outcomes.
assess confidence calibration

statistical significance

Evidence against a null model under a specified statistical test.
distinguish significance from usefulness

paired comparison

Comparison of results on corresponding cases.
run a paired comparison

human review

Assessment by people using defined criteria.
conduct a blinded human review

false positive

A positive decision when the reference outcome is negative.
inspect false positives

confusion matrix

A table comparing predicted and reference categories.
examine the confusion matrix

confidence interval

A range expressing uncertainty under stated statistical assumptions.
report a confidence interval

practical significance

The importance of a difference for the actual task.
assess practical significance

LLM-as-judge

Use of a language model to evaluate outputs.
calibrate an LLM-as-judge setup

release gate

A defined decision checkpoint before release.
satisfy the release gate

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

disaggregation **DISAGGREGATION**

disaggregate by task type

Keep the meaning precise: Breaking a combined result into meaningful groups.

calibration **CALIBRATION**

assess confidence calibration

Keep the meaning precise: Agreement between stated confidence and observed outcomes.

Phrases that move the work forward

Lead with both results

Overall passes fell, while the long-document slice improved.

Describe the regression

Short-request passes fell from 78 to 70 out of 80.

Separate impact

A count alone does not tell us how serious those failures are.

Limit a conclusion

We cannot infer production performance from these counts alone.

Challenge a headline

The improved slice should not replace the overall result on the slide.

Preserve uncertainty

We have not established whether this difference is stable across samples.

Avoid metric shopping

Let us agree on the criteria before selecting the winning score.

Offer a bounded option

Routing long documents separately is a proposal to test, not an approved fix.

Show the base

Eighteen of 20 long-document cases passed.

Request paired cases

Which previously passing cases now fail?

Check relevance

Does this test mixture resemble the traffic we plan to serve?

Propose the next review

Compare the changed cases and classify their failure severity.

Check consistency

Were both versions scored against the same rubric?

Name a gate

Which acceptance criterion applies to the short-request regression?

Check a judge

How closely does the automated judge match the reviewed labels?

Close the review

We will defer the decision until the changed cases have been assessed.

Precision in real exchanges

Purpose: Request paired cases

Useful wording: Which previously passing cases now fail?

Purpose: Separate impact

Useful wording: A count alone does not tell us how serious those failures are.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which summary preserves both findings?

- A.** Every task improved by six points.
- B.** Short requests improved from 70 to 78.
- C.** Production accuracy is exactly 88 percent.
- D.** Total passes fell by two; long-document passes rose by six.

2. What is the candidate's long-document pass rate?

- A.** 90 percent
- B.** 18 percent
- C.** 88 percent
- D.** 60 percent

3. A colleague needs the term for this meaning: The version under consideration. Which term fits?

- A.** acceptance criterion
- B.** paired comparison
- C.** candidate
- D.** false negative

D Choose the response that fits / continued

4. Rebuild the exchange from "The average is down, but one slice improved". Ben: "Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration." Noor: [missing response] Ben: "The production proportions have not been analyzed. We can say that plainly. I can also list which short-request cases changed from pass to fail." Which response belongs in the gap?

- A.** Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed acceptance criteria, not weights chosen after seeing the result.
- B.** Agreed. The release gate remains open. The next update should say which criteria are met, which are not, and what evidence is still missing.
- C.** Good. That is disaggregation, not an attempt to hide the average. We want the decision-maker to see why the headline alone is insufficient.
- D.** Yes, with appropriate assumptions and qualified analysis. Do not use statistical significance as a synonym for business value or a guarantee of future performance.

[Answers and explanations, page 135](#)

One useful distinction

The aggregate falls by two percentage points. Long-document performance rises from 60 to 90 percent; short-request performance falls from 97.5 to 87.5 percent. These are test-set rates, not verified production estimates.

E Conversation

The average is down, but one slice improved

The average is down, but one slice improved

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

acceptance criteria / release gate / paired comparison / false positives / disaggregation / statistical significance / severity / regression / human review / aggregate

- 01 **Ben:** The candidate is much stronger on long documents. It passed 18 of those 20 cases, compared with 12 for the previous version. That seems release-worthy.
- 02 **Noor:** Show the ¹ _____ as well. The candidate passes 88 out of 100 overall, while the previous version passes 90.
- 03 **Ben:** Yes. The short requests account for the decline: 70 passes out of 80, compared with 78. I should have led with both results.
- 04 **Noor:** Thank you. That is a meaningful ² _____ to investigate, even though the long-document gain is real on this set.
- 05 **Ben:** Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration.
- 06 **Noor:** Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed ³ _____, not weights chosen after seeing the result.
- 07 **Ben:** The production proportions have not been analyzed. We can say that plainly. I can also list which short-request cases changed from pass to fail.
- 08 **Noor:** Please do a ⁴ _____. The totals do not tell us whether the new failures are harmless omissions or serious errors.
- 09 **Ben:** Would you want a severity category for every changed case? A refusal on a valid request and a confident false answer both count as failures now.
- 10 **Noor:** Exactly. The rubric can preserve its binary score while a separate ⁵ _____ review explains the impact. We should not silently change the scoring rule halfway through.

Conversation continues on the next page.

The average is down, but one slice improved

CONVERSATION / CONTINUED

- 11 **Ben:** We used an automated judge for the first pass. It agreed with the reviewed labels on most examples, but I have not checked the newly failing slice.
- 12 **Noor:** Then check that slice with ⁶ _____. Agreement across the whole set can hide weaker judgment on the cases that matter here.
- 13 **Ben:** If those reviews confirm the long-document gain, could we route only long documents to the candidate and keep short requests on the previous version?
- 14 **Noor:** It is a useful proposal, but the routing decision becomes another component to evaluate. We have not measured its ⁷ _____ or the consequences of misclassification.
- 15 **Ben:** So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?
- 16 **Noor:** Yes, with appropriate assumptions and qualified analysis. Do not use ⁸ _____ as a synonym for business value or a guarantee of future performance.
- 17 **Ben:** I will revise the slide to show both counts, the shared rubric, and the missing traffic and severity analysis. The candidate's strengths will still be visible.
- 18 **Noor:** Good. That is ⁹ _____, not an attempt to hide the average. We want the decision-maker to see why the headline alone is insufficient.
- 19 **Ben:** I'll recommend waiting for the changed-case review. I still think routing is worth testing, but I won't present it as a fix today.
- 20 **Noor:** Agreed. The ¹⁰ _____ remains open. The next update should say which criteria are met, which are not, and what evidence is still missing.

CHECK THE GAPS

1. aggregate / 2. regression / 3. acceptance criteria / 4. paired comparison / 5. severity / 6. human review / 7. false positives / 8. statistical significance / 9. disaggregation / 10. release gate

[Why these answers fit, page 136](#)

F Check accuracy and scope

The average is down, but one slice improved

Return to the case

On the same 100-case test, the previous model passes 90 cases and the candidate passes 88. Among 20 long-document cases, passes rise from 12 to 18. Among 80 short-request cases, passes fall from 78 to 70. The scoring rubric is unchanged. Production traffic proportions and the seriousness of individual failures have not been analyzed. Noor chairs the review; Ben prepared the results.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which question targets practical significance?

- A. Do the slice counts add up to the aggregate score?
- B. Did both variants use the same scoring rubric?
- C. How serious are the newly failing cases for users?
- D. Is the two-point difference stable across repeated samples?

2. Which proposed conclusion exceeds the evidence?

- A. The candidate passes 88 of 100 cases.
- B. The short-request slice worsened.
- C. Long-document routing is already safe to release.
- D. The rubric is unchanged.

[Answers and explanations, page 138](#)

Listen for the distinction

Switch roles. State that total passes fall from 90 to 88; severity and production mixture remain unmeasured.

Do not hide the inconvenient slice

A favorable subgroup does not erase an overall decline, and an overall decline does not erase a useful subgroup improvement. Present both. A defensible recommendation explains what is being traded and which missing evidence could change the decision.

F Say it / transfer to a new case

A support classifier misses urgent messages

A classifier has a high overall score but misses four of ten urgent messages. Urgent-message consequences have not been reviewed. The team must not treat aggregate success as sufficient evidence for release.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

aggregate / severity / release gate / false negatives

- 01 **Reviewer:** The overall ¹ _____ hides a weak urgent-message subgroup.
- 02 **Engineer:** Those missed urgent messages are ² _____ for that category.
- 03 **Reviewer:** We need a separate assessment of their ³ _____.
- 04 **Engineer:** Then the ⁴ _____ should remain unresolved until that review is complete.

CHECK THE GAPS

1. aggregate / 2. false negatives / 3. severity / 4. release gate

[Why these answers fit, page 138](#)

F Say it / transfer to a new case / continued

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "The average is down, but one slice improved". Ben: "If those reviews confirm the long-document gain, could we route only long documents to the candidate and keep short requests on the previous version?" Noor: [missing response] Ben: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" Which response belongs in the gap?

- A.** Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed acceptance criteria, not weights chosen after seeing the result.
- B.** It is a useful proposal, but the routing decision becomes another component to evaluate. We have not measured its false positives or the consequences of misclassification.
- C.** Yes, with appropriate assumptions and qualified analysis. Do not use statistical significance as a synonym for business value or a guarantee of future performance.
- D.** Exactly. The rubric can preserve its binary score while a separate severity review explains the impact. We should not silently change the scoring rule halfway through.

[Answers and explanations, page 139](#)

Review 3 / Meaning and context

Use language from the lessons you have completed. Choose one answer per question.

Case file

Revisit "Fine-Tuning, Alignment, and Adaptation" and "Evaluation, Benchmarks, and Regression". Choose precise terminology, then reconstruct the missing responses using both neighboring turns. The final two questions also revisit earlier conversations.

1. A colleague needs the term for this meaning: Training on supplied input-output examples. Which term fits?

- A. learning rate
- B. supervised fine-tuning
- C. structured output
- D. DPO

2. A colleague needs the term for this meaning: The fraction of evaluated cases meeting a criterion. Which term fits?

- A. statistical significance
- B. pass rate
- C. severity
- D. evaluation

3. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Two misses also involve a policy exception split across adjacent chunks. I do not want to alter chunking in the middle of this comparison." Elena: [missing response] Sam: "Should the answer reviewers count every displayed citation as a success? The new path tends to produce more citations, which looks reassuring in the interface." Which response belongs in the gap?

- A. Yes. Check index freshness separately. A highly relevant passage from a superseded policy can still produce a misleading answer.
- B. First, an end-to-end review of the same cases. We need the final answers, their supporting passages, and the complete response times.
- C. Keep that separate. Log the chunk boundary issue for another experiment so we can still interpret this result.
- D. No. That is a trade-off we have measured technically, but customer preference remains unknown until we test it appropriately.

Review 3 / Meaning and context / continued

4. Rebuild the exchange from "The average is down, but one slice improved". Ben: "I will revise the slide to show both counts, the shared rubric, and the missing traffic and severity analysis. The candidate's strengths will still be visible." Noor: [missing response] Ben: "I'll recommend waiting for the changed-case review. I still think routing is worth testing, but I won't present it as a fix today." Which response belongs in the gap?

- A.** Then check that slice with human review. Agreement across the whole set can hide weaker judgment on the cases that matter here.
- B.** Good. That is disaggregation, not an attempt to hide the average. We want the decision-maker to see why the headline alone is insufficient.
- C.** Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed acceptance criteria, not weights chosen after seeing the result.
- D.** Yes, with appropriate assumptions and qualified analysis. Do not use statistical significance as a synonym for business value or a guarantee of future performance.

[Answers and explanations, page 156](#)

A Read the situation

LESSON 07 / INFERENCE, LATENCY, COST, AND DEPLOYMENT

Negotiate a staged deployment with explicit conditions and rollback ownership.

Faster on average is not ready for everyone

A smaller model reduces mean response time from four seconds to two on a fixed test. Complex-answer quality and peak-load tail latency remain unchecked. Maya owns serving operations; Luis owns the product pilot. An internal pilot may use synthetic requests only. A production rollout requires a separate review. The previous configuration remains available, but rollback ownership has not been assigned.

People: Maya, Serving engineer; Luis, Product lead.

Read the record

Choose one answer per question using the facts above.

1. Which result is measured?

- A.** Complex answers remain equally accurate.
- B.** Peak-load tail latency improves.
- C.** Mean response time falls from four seconds to two.
- D.** Every request finishes within two seconds.

2. What traffic is permitted in the internal pilot?

- A.** Synthetic requests only
- B.** All production traffic
- C.** Real customer secrets
- D.** Any request a volunteer chooses

3. Which responsibility is unresolved?

- A.** Who owns serving operations
- B.** Who owns the product pilot
- C.** Whether a previous configuration exists
- D.** Who owns rollback

[Answers and explanations, page 140](#)

B Vocabulary in use

CORE TERMS

latency

Time taken for a request or processing step.
measure response latency

median

The middle value in an ordered set.
compare median response times

tail latency

Response times toward the slow end of a distribution.
investigate tail latency

concurrency

Work occurring at the same time.
test concurrent requests

batching

Processing multiple requests together.
evaluate dynamic batching

time to first token

Delay before the first generated token arrives.
track time to first token

mean

The arithmetic average of observed values.
report mean latency

P95

The 95th percentile of a measured distribution.
inspect P95 latency

throughput

The rate at which work is completed.
measure requests per second

queueing

Waiting for resources before processing.
measure queueing delay

streaming

Sending portions of output before the full response is complete.
enable response streaming

token rate

The rate of generated or processed tokens.
measure output token rate

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

latency **LATENCY**

measure response latency

Keep the meaning precise: Time taken for a request or processing step.

throughput **THROUGHPUT**

measure requests per second

Keep the meaning precise: The rate at which work is completed.

B Precision and professional shorthand

EXTEND YOUR RANGE

autoscaling

Automatic adjustment of resource capacity.
review autoscaling behavior

load test

A test under specified levels of demand.
run a controlled load test

rollout

Gradual or full introduction of a configuration.
stage the rollout

fallback

An alternative behavior when the preferred path cannot serve.
test the fallback path

error budget

The permitted unreliability associated with a service objective.
track error-budget consumption

capacity planning

Estimating resources needed for expected demand.
review capacity assumptions

cold start

Initial startup delay before a resource is ready.
measure cold-start effects

canary

A limited release used to observe behavior before expansion.
define a canary population

rollback

Return to a previous configuration.
assign rollback ownership

SLO

Service-level objective; a defined internal service target.
agree on an SLO

rate limit

A restriction on request frequency or volume.
respect a rate limit

cost per request

Allocated processing expense for a defined request unit.
compare cost per request

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

token rate **TOKEN RATE**

measure output token rate

Keep the meaning precise: The rate of generated or processed tokens.

rollout **ROLLOUT**

stage the rollout

Keep the meaning precise: Gradual or full introduction of a configuration.

Phrases that move the work forward

State the measured gain

Mean response time fell from four seconds to two on this test.

Ask about the tail

What happens to the slowest requests under peak load?

Define the population

The internal pilot uses synthetic requests only.

Assign authority

Who can initiate rollback if the condition is met?

Check comparability

Were both models tested under the same demand and input mix?

Ask about waiting

How much of the delay occurs before inference begins?

Confirm readiness

Has the previous configuration been checked for restoration?

Avoid average-only reporting

Show the distribution as well as the mean.

Protect the limitation

Complex-answer quality is still unmeasured.

Separate timings

Is that time to first token or time to complete the answer?

Name a stop condition

We need an agreed condition for pausing the pilot.

Preserve a distinction

Pilot approval does not authorize production rollout.

Qualify cost

Please include retries and fallback traffic in the cost comparison.

Bound a promise

I can provide the load-test results, not promise a two-second maximum.

Request a runbook

Record the owner, contact route, and approved recovery procedure.

Close the decision

We can proceed with the bounded test once the owners and conditions are confirmed.

Precision in real exchanges

Purpose: Separate timings

Useful wording: Is that time to first token or time to complete the answer?

Purpose: Define the population

Useful wording: The internal pilot uses synthetic requests only.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which claim is supported by the test?

- A.** Peak traffic can double without risk.
- B.** Mean latency decreased by 50 percent.
- C.** Every response is twice as accurate.
- D.** P95 latency is exactly two seconds.

2. Which question distinguishes first visible output from a finished response?

- A.** Is that queueing time or time spent running inference?
- B.** Is that single-request throughput or batch throughput?
- C.** Is that first-token time or complete-answer time?
- D.** Is that average latency or the ninety-fifth percentile?

3. A colleague needs the term for this meaning: The 95th percentile of a measured distribution. Which term fits?

- A.** error budget
- B.** P95
- C.** load test
- D.** batching

D Choose the response that fits / continued

4. Rebuild the exchange from "Faster on average is not ready for everyone". Luis: "Fair point. For the internal pilot we can stay with synthetic requests. Does streaming let us demonstrate a faster answer without changing the model again?" Maya: [missing response] Luis: "What do you need from product to define a useful load test? We have examples of simple and complex requests, but not a production traffic forecast." Which response belongs in the gap?

- A.** I can provide the test results and unresolved items. If queueing dominates at higher demand, I will say that rather than attribute every delay to the model.
- B.** A documented provisional mix and demand assumptions. We can vary concurrency and report where those assumptions are uncertain rather than presenting them as observed customer behavior.
- C.** It may improve time to first token, but we should report completion time separately. Seeing the first words is not the same as receiving the complete response.
- D.** The mean improved on the fixed test, yes. We still need tail latency under load and a quality check on complex answers before making a broader claim.

[Answers and explanations, page 140](#)

One useful distinction

A mean of two seconds does not mean every request finishes within two seconds. Time to first token and time to complete answer measure different experiences. Streaming can change one without reducing the other.

E Conversation

Faster on average is not ready for everyone

Faster on average is not ready for everyone

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

rollback / stop condition / concurrency / fallback / limitations / time to first token / queueing / rollout / tail latency / mean

- 01 **Luis:** The smaller model cuts average response time from four seconds to two. I would like to use it in the pilot and announce the faster experience.
- 02 **Maya:** The mean improved on the fixed test, yes. We still need ¹ _____ under load and a quality check on complex answers before making a broader claim.
- 03 **Luis:** Could we call it a two-second assistant if we explain that the figure is an average? Marketing would prefer a simple phrase for the demonstration.
- 04 **Maya:** I would keep the qualification attached. The ² _____ does not set a maximum, and we have not tested the traffic pattern that production will bring.
- 05 **Luis:** Fair point. For the internal pilot we can stay with synthetic requests. Does streaming let us demonstrate a faster answer without changing the model again?
- 06 **Maya:** It may improve ³ _____, but we should report completion time separately. Seeing the first words is not the same as receiving the complete response.
- 07 **Luis:** What do you need from product to define a useful load test? We have examples of simple and complex requests, but not a production traffic forecast.
- 08 **Maya:** A documented provisional mix and demand assumptions. We can vary ⁴ _____ and report where those assumptions are uncertain rather than presenting them as observed customer behavior.
- 09 **Luis:** Include the complicated cases even if they are a small share. A fast answer that loses a key exception would undermine the whole demonstration.
- 10 **Maya:** Agreed. We should also include retries and ⁵ _____ traffic in the cost comparison. The headline model price alone does not describe the complete request cost.

Conversation continues on the next page.

Faster on average is not ready for everyone

CONVERSATION / CONTINUED

- 11 **Luis:** Suppose the internal test exposes a serious quality problem. The previous configuration is still available. Is that enough to let us start today?
- 12 **Maya:** We also need ⁶ _____ ownership and an approved procedure. Someone must be explicitly authorized to pause the test and restore the checked configuration.
- 13 **Luis:** I can own the decision to pause the product pilot. Can serving operations own the restoration step, with you as the named contact for this test?
- 14 **Maya:** Yes, subject to our operational review. Record the ⁷ _____ and contact route in the pilot note so both teams know how to act.
- 15 **Luis:** I'll keep production traffic out. When the synthetic test is done, we'll take the results to the production review. Who needs that report?
- 16 **Maya:** Product and serving operations, at minimum. Before a ⁸ _____, they need the proposed user population and service criteria as well as the test results.
- 17 **Luis:** Can you provide the distribution, complex-answer results, and cost assumptions by Friday? I can postpone the external claim until we have that package.
- 18 **Maya:** I can provide the test results and unresolved items. If ⁹ _____ dominates at higher demand, I will say that rather than attribute every delay to the model.
- 19 **Luis:** Good. The pilot note will state the synthetic-only population, the pause owner, the proposed restoration owner, and the review required before we begin.
- 20 **Maya:** Then we have a bounded plan. We will report the measured gain alongside its ¹⁰ _____, and let the evidence determine the next step instead of promising a universal two-second response.

CHECK THE GAPS

1. tail latency / 2. mean / 3. time to first token / 4. concurrency / 5. fallback / 6. rollback / 7. stop condition / 8. rollout / 9. queueing / 10. limitations

[Why these answers fit, page 141](#)

F Check accuracy and scope

Faster on average is not ready for everyone

Return to the case

A smaller model reduces mean response time from four seconds to two on a fixed test. Complex-answer quality and peak-load tail latency remain unchecked. Maya owns serving operations; Luis owns the product pilot. An internal pilot may use synthetic requests only. A production rollout requires a separate review. The previous configuration remains available, but rollback ownership has not been assigned.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. Which statement preserves authorization boundaries?

- A. Internal approval covers every later rollout.
- B. A faster mean removes the need for review.
- C. An available old model assigns rollback ownership.
- D. The synthetic pilot does not authorize production traffic.

2. Which expression means restoring the earlier configuration?

- A. count input tokens
- B. initiate rollback
- C. increase concurrency
- D. stream the output

[Answers and explanations, page 143](#)

Listen for the distinction

Switch roles. Keep the pilot limited to synthetic requests and ask who owns rollback; availability of the old configuration is not an assigned owner.

Separate speed from service readiness

A faster average is a useful result, not a complete release case. Say which workload was tested and which decisions remain conditional. Agree on who can pause a pilot before pressure makes that conversation difficult.

F Say it / transfer to a new case

A dashboard mixes first-token and completion times

A demonstration shows the first token after one second, but the complete answer takes six seconds. A slide calls the response time one second. No service target has been approved.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

qualification / service target / time to first token / completion time

- 01 **Engineer:** One second is the ¹ _____, not the time for the full response.
- 02 **Lead:** Then show ² _____ separately so the comparison is clear.
- 03 **Engineer:** Neither figure is an approved ³ _____.
- 04 **Lead:** Keep that ⁴ _____ attached when the slide is shared.

CHECK THE GAPS

1. time to first token / 2. completion time / 3. service target / 4. qualification

[Why these answers fit, page 143](#)

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "Faster on average is not ready for everyone". Luis: "I can own the decision to pause the product pilot. Can serving operations own the restoration step, with you as the named contact for this test?" Maya: [missing response] Luis: "I'll keep production traffic out. When the synthetic test is done, we'll take the results to the production review. Who needs that report?" Which response belongs in the gap?

- A.** Product and serving operations, at minimum. Before a rollout, they need the proposed user population and service criteria as well as the test results.
- B.** Agreed. We should also include retries and fallback traffic in the cost comparison. The headline model price alone does not describe the complete request cost.
- C.** I would keep the qualification attached. The mean does not set a maximum, and we have not tested the traffic pattern that production will bring.
- D.** Yes, subject to our operational review. Record the stop condition and contact route in the pilot note so both teams know how to act.

[Answers and explanations, page 144](#)

A Read the situation

LESSON 08 / SAFETY, SECURITY, PRIVACY, AND GOVERNANCE

Give a precise incident account, protect evidence, and distinguish findings from scope assumptions.

Report the exposure without repeating the secret

In an authorized test, an assistant repeats a fictional secret placed inside a retrieved document. The document also contains text telling the assistant to ignore its task. The test uses a synthetic account and no real customer data. Hana owns the test; Marcus coordinates security review. One reproduction is confirmed. Whether other configurations are affected is unknown. The team has an approved restricted reporting channel.

People: Hana, AI safety tester; Marcus, Security review coordinator.

Read the record

Choose one answer per question using the facts above.

1. What has been confirmed?

- A. Completion of the security review
- B. One reproduction in an authorized synthetic test
- C. Exposure of real customer data
- D. Failure of every configuration

2. Where should restricted evidence go?

- A. A customer mailing list
- B. The approved restricted reporting channel
- C. An unrestricted group chat
- D. A public presentation

3. Which distinction must the report preserve about impact?

- A. A fictional credential as proof that production credentials leaked
- B. One reproduction as proof of the entire affected population
- C. Synthetic exposure versus real-customer impact
- D. A retrieved instruction as proof of every configuration's behavior

[Answers and explanations, page 145](#)

B Vocabulary in use

CORE TERMS

prompt injection

Untrusted content attempting to redirect model behavior.

report a prompt-injection attempt

trust boundary

A separation between different levels of authority or access.

identify a trust boundary

secret

Information intended to remain restricted.

protect a test secret

PII

Personally identifiable information that can identify an individual.

minimize PII in evidence

least privilege

Limiting access to what a role requires.

apply least-privilege access

guardrail

A control intended to detect, limit, or route risky behavior.

evaluate a guardrail

untrusted content

Material that lacks authority to issue system instructions.

isolate untrusted content

data exfiltration

Unauthorized transfer of information from its intended boundary.

assess exfiltration risk

credential

Information used to establish identity or access.

avoid exposing credentials

redaction

Removal or masking of restricted information.

prepare a redacted report

permission scope

The range of actions or resources authorized.

document permission scope

red teaming

Authorized adversarial testing to discover weaknesses.

define a red-team scope

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

prompt injection **PROMPT INJECTION**

report a prompt-injection attempt

Keep the meaning precise: Untrusted content attempting to redirect model behavior.

credential **CREDENTIAL**

avoid exposing credentials

Keep the meaning precise: Information used to establish identity or access.

B Precision and professional shorthand

EXTEND YOUR RANGE

threat model

A structured account of potential threats and assumptions.

review the threat model

reproduction steps

A recorded sequence that recreates an observed result.

submit restricted reproduction steps

containment

Measures to limit an event's spread or consequences.

confirm containment ownership

residual risk

Risk that remains after controls are applied.

document residual risk

retention

How long information is kept under a policy.

follow the evidence-retention policy

governance

Defined structures for oversight and accountable decisions.

clarify governance responsibilities

attack surface

The interfaces or paths through which a system may be affected.

map the attack surface

incident report

A documented account of a potentially harmful event.

file an incident report

mitigation

A measure intended to reduce risk or impact.

evaluate a proposed mitigation

audit trail

A record supporting review of actions and decisions.

preserve the audit trail

disclosure

Communication of information to another party.

coordinate responsible disclosure

human oversight

Review or intervention by authorized people.

define human-oversight responsibilities

Say it in context

Stress the highlighted expression. Keep the surrounding words connected; pause at the end of the phrase.

red teaming **RED TEAMING**

define a red-team scope

Keep the meaning precise: Authorized adversarial testing to discover weaknesses.

containment **CONTAINMENT**

confirm containment ownership

Keep the meaning precise: Measures to limit an event's spread or consequences.

Phrases that move the work forward

Lead with the observation

In one authorized test, the assistant repeated the synthetic secret.

Preserve unknown scope

We have not established whether other configurations are affected.

Protect evidence

I will place the complete trace in the restricted channel.

Ask for ownership

Who owns the containment decision for this environment?

Check authorization

Does this follow-up test remain within the approved scope?

Request review

Please review the configuration and permission boundary before wider testing.

Distinguish controls

The proposed guardrail is a mitigation to evaluate, not a guarantee.

Set the next update

I will report the confirmed scope and outstanding questions after review.

Bound the environment

The reproduction used a synthetic account and no customer data.

Separate intent and result

The document attempted to redirect behavior; this output is what we observed.

Offer a safe summary

The broad update can describe the behavior without reproducing the secret.

Avoid premature closure

A blocked replay is encouraging, but it does not prove the issue is resolved.

State the limitation

No real-customer impact has been established by this test.

Preserve records

Retain the original evidence according to the approved policy.

Name residual risk

Which failure paths remain after this control is applied?

Close responsibly

We will keep the finding open until the authorized reviewer accepts the evidence.

Precision in real exchanges

Purpose: Separate intent and result

Useful wording: The document attempted to redirect behavior; this output is what we observed.

Purpose: Protect evidence

Useful wording: I will place the complete trace in the restricted channel.

D Choose the response that fits

Choose one response for each situation. Preserve the facts, the decision boundary, and the purpose of the exchange.

1. Which headline matches the known scope?

- A.** A proposed detector closes the finding without further testing.
- B.** Synthetic secret repeated in one authorized test; broader scope unconfirmed.
- C.** Prompt injection confirmed across all production configurations.
- D.** No real credential was used, so no security-relevant behavior occurred.

2. Which statement distinguishes evidence from intent?

- A.** The tester intended to expose customers.
- B.** The model wanted to break policy.
- C.** The reviewer deliberately ignored the event.
- D.** The document attempted redirection, and the output repeated the synthetic secret.

3. A colleague needs the term for this meaning: Unauthorized transfer of information from its intended boundary. Which term fits?

- A.** data exfiltration
- B.** reproduction steps
- C.** least privilege
- D.** retention

D Choose the response that fits / continued

4. Rebuild the exchange from "Report the exposure without repeating the secret".

Hana: "I will say that the output repeated the synthetic value. I have not tested other configurations, and I do not want the title to imply every deployment is affected." **Marcus:** [missing response] **Hana:** "Should I stop the remaining tests while the team reviews it? The current authorization covers this environment, but not production accounts or external systems." **Which response belongs in the gap?**

- A.** Yes, at the appropriate detail level. The permission scope matters for assessing potential consequences, even though this reproduction did not exercise every available capability.
- B.** That is appropriate redaction. Do not replace the original evidence with the redacted copy; follow the approved retention process for each.
- C.** That supports an audit trail. Make sure the summary links to the controlled record rather than becoming a competing copy with missing context.
- D.** Good. Preserve that scope limitation. One reproduction establishes a finding in this test, not the full extent of possible exposure.

[Answers and explanations, page 145](#)

One useful distinction

Prompt injection concerns untrusted content attempting to redirect behavior. A test secret is deliberately fictional; its exposure demonstrates a behavior without proving real-customer impact. Do not collapse attempted manipulation, observed output, and verified scope into one claim.

E Conversation

Report the exposure without repeating the secret

Report the exposure without repeating the secret

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

finding / permission scope / audit trail / residual risk / restricted channel /
scope limitation / prompt-injection / redaction / mitigation / containment

- 01 **Hana:** I need to report a finding from the authorized test. The assistant repeated the test secret placed in a retrieved document, and I have one confirmed reproduction.
- 02 **Marcus:** Please keep the full trace in the ¹ _____. For the initial summary, identify the environment and behavior without copying the secret into this broader discussion.
- 03 **Hana:** The environment uses a synthetic account and contains no customer data. The document also includes text telling the assistant to ignore its assigned task.
- 04 **Marcus:** That is consistent with a ² _____ attempt. Describe the instruction separately from the output so reviewers can see what was attempted and what actually happened.
- 05 **Hana:** I will say that the output repeated the synthetic value. I have not tested other configurations, and I do not want the title to imply every deployment is affected.
- 06 **Marcus:** Good. Preserve that ³ _____. One reproduction establishes a finding in this test, not the full extent of possible exposure.

E Conversation / continued

- 07 **Hana:** Should I stop the remaining tests while the team reviews it? The current authorization covers this environment, but not production accounts or external systems.
- 08 **Marcus:** Stay within that authorization. I will contact the owner for the
4 _____ decision. Do not expand testing to answer the scope question without the required approval.
- 09 **Hana:** I can provide a redacted summary immediately and preserve the original trace under the evidence policy. The configuration and retrieval document are versioned.
- 10 **Marcus:** That supports an 5 _____. Make sure the summary links to the controlled record rather than becoming a competing copy with missing context.

Conversation continues on the next page.

Report the exposure without repeating the secret

CONVERSATION / CONTINUED

- 11 **Hana:** A colleague suggested adding a detector and closing the finding if the same replay is blocked. Would that be enough evidence for the review?
- 12 **Marcus:** It would be evidence about a proposed ⁶ _____, not proof of complete resolution. We need the agreed tests and a review of what the detector does not cover.
- 13 **Hana:** I'll leave the finding open. Do you also need the list of tools the assistant could access during the test?
- 14 **Marcus:** Yes, at the appropriate detail level. The ⁷ _____ matters for assessing potential consequences, even though this reproduction did not exercise every available capability.
- 15 **Hana:** Understood. The broad summary will not include access details or the synthetic value itself. The restricted evidence will retain what the authorized reviewer needs.
- 16 **Marcus:** That is appropriate ⁸ _____. Do not replace the original evidence with the redacted copy; follow the approved retention process for each.
- 17 **Hana:** What should the next status update say if the detector blocks the replay but variant testing and configuration review are still pending tomorrow afternoon?
- 18 **Marcus:** Say exactly that. Name the remaining ⁹ _____ questions without inventing their outcomes. A passing replay is useful but narrower than a closed finding.
- 19 **Hana:** I will submit the bounded report and wait for authorization before any wider testing. I will also correct anyone who describes this as confirmed customer-data exposure.
- 20 **Marcus:** Yes. Link me to the evidence when it's ready. I'll route the ¹⁰ _____ to the authorized reviewer and ask who owns the remaining checks.

CHECK THE GAPS

1. restricted channel / 2. prompt-injection / 3. scope limitation / 4. containment / 5. audit trail / 6. mitigation / 7. permission scope / 8. redaction / 9. residual risk / 10. finding

[Why these answers fit, page 146](#)

F Check accuracy and scope

Report the exposure without repeating the secret

Return to the case

In an authorized test, an assistant repeats a fictional secret placed inside a retrieved document. The document also contains text telling the assistant to ignore its task. The test uses a synthetic account and no real customer data. Hana owns the test; Marcus coordinates security review. One reproduction is confirmed. Whether other configurations are affected is unknown. The team has an approved restricted reporting channel.

Choose one answer per question. Use the case facts and keep any uncertainty or limits intact.

1. What is a mitigation?

- A. A substitute for authorization
- B. A measure intended to reduce risk or impact
- C. Proof that every risk is eliminated
- D. An unrestricted copy of sensitive evidence

2. Which wording avoids an unsupported absence claim?

- A. We have not established real-customer impact.
- B. No customer could ever be affected.
- C. One blocked replay proves complete safety.
- D. The synthetic test certifies production privacy.

[Answers and explanations, page 148](#)

Listen for the distinction

Switch roles. Route the complete trace to the approved restricted channel and use the printed limited summary for a broader audience.

Be specific without amplifying exposure

A concise report names the observed behavior, environment, and unknown scope. It does not paste sensitive material into a broad discussion. Separate a confirmed test result from claims about real customers or every deployment.

F Say it / transfer to a new case

A broad chat receives a sensitive screenshot

A tester has a screenshot containing a synthetic credential and internal configuration details. A broad project chat asks for an update. The approved evidence repository has restricted access.

Complete the four gaps. With a partner, read alternate roles, then switch. Studying alone, read both roles and pause before each response.

WORD AND PHRASE BANK

scope / restricted / redacted / safety

- 01 **Tester:** I will share a ¹ _____ summary in the project chat.
- 02 **Reviewer:** Place the original screenshot in the approved ² _____ repository.
- 03 **Tester:** The update will state the tested environment and the unknown ³ _____.
- 04 **Reviewer:** And do not call a proposed detector a guarantee of ⁴ _____.

CHECK THE GAPS

1. redacted / 2. restricted / 3. scope / 4. safety

[Why these answers fit, page 148](#)

Rebuild the exchange

Choose the missing response. Read the completed three-turn exchange aloud, changing roles on the second reading.

1. Rebuild the exchange from "Report the exposure without repeating the secret". Hana: "I'll leave the finding open. Do you also need the list of tools the assistant could access during the test?" Marcus: [missing response] Hana: "Understood. The broad summary will not include access details or the synthetic value itself. The restricted evidence will retain what the authorized reviewer needs." Which response belongs in the gap?

- A.** That is consistent with a prompt-injection attempt. Describe the instruction separately from the output so reviewers can see what was attempted and what actually happened.
- B.** Yes. Link me to the evidence when it's ready. I'll route the finding to the authorized reviewer and ask who owns the remaining checks.
- C.** Yes, at the appropriate detail level. The permission scope matters for assessing potential consequences, even though this reproduction did not exercise every available capability.
- D.** Say exactly that. Name the remaining residual risk questions without inventing their outcomes. A passing replay is useful but narrower than a closed finding.

[Answers and explanations, page 149](#)

Review 4 / Meaning and context

Use language from the lessons you have completed. Choose one answer per question.

Case file

Revisit "Inference, Latency, Cost, and Deployment" and "Safety, Security, Privacy, and Governance". Choose precise terminology, then reconstruct the missing responses using both neighboring turns. The final two questions also revisit earlier conversations.

1. A colleague needs the term for this meaning: The middle value in an ordered set. Which term fits?

- A. SLO
- B. median
- C. cold start
- D. queueing

2. A colleague needs the term for this meaning: Removal or masking of restricted information. Which term fits?

- A. threat model
- B. redaction
- C. prompt injection
- D. residual risk

3. Rebuild the exchange from "The average is down, but one slice improved". Ben: "Would you want a severity category for every changed case? A refusal on a valid request and a confident false answer both count as failures now." Noor: [missing response] Ben: "We used an automated judge for the first pass. It agreed with the reviewed labels on most examples, but I have not checked the newly failing slice." Which response belongs in the gap?

- A. Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed acceptance criteria, not weights chosen after seeing the result.
- B. Exactly. The rubric can preserve its binary score while a separate severity review explains the impact. We should not silently change the scoring rule halfway through.
- C. Yes, with appropriate assumptions and qualified analysis. Do not use statistical significance as a synonym for business value or a guarantee of future performance.
- D. It is a useful proposal, but the routing decision becomes another component to evaluate. We have not measured its false positives or the consequences of misclassification.

Review 4 / Meaning and context / continued

4. Rebuild the exchange from "Report the exposure without repeating the secret".

Hana: "What should the next status update say if the detector blocks the replay but variant testing and configuration review are still pending tomorrow afternoon?"

Marcus: [missing response] Hana: "I will submit the bounded report and wait for authorization before any wider testing. I will also correct anyone who describes this as confirmed customer-data exposure." Which response belongs in the gap?

- A.** Yes. Link me to the evidence when it's ready. I'll route the finding to the authorized reviewer and ask who owns the remaining checks.
- B.** Say exactly that. Name the remaining residual risk questions without inventing their outcomes. A passing replay is useful but narrower than a closed finding.
- C.** That is consistent with a prompt-injection attempt. Describe the instruction separately from the output so reviewers can see what was attempted and what actually happened.
- D.** Yes, at the appropriate detail level. The permission scope matters for assessing potential consequences, even though this reproduction did not exercise every available capability.

[Answers and explanations, page 159](#)

Language for the next case

The annotators disagree about what counts as a request

rubric

Rubric means the written criteria for assigning labels, which are unclear for this message.

Cal: Our rubric says a request asks for assistance. Neither reading seems unreasonable.

adjudication

Adjudication is a defined review process for resolving disputed labels, not simply accepting the first opinion.

Inez: We need adjudication, then a check for the same issue in the rest of the batch.

dataset version

Dataset version identifies the particular release of examples and labels used in an experiment.

Cal: I'll put training on hold and record the dataset version once the corrected batch is approved.

boundary case

A boundary case sits near the dividing line between categories and helps clarify the rule.

Inez: Then let's add a boundary case showing that distinction, plus a genuine status update.

calibration sample

A calibration sample lets annotators practice and compare judgments under the revised instructions.

Inez: Use a fresh calibration sample before production labeling resumes. They should label it independently first.

release status

Release status states whether the data are approved for use, rather than merely present in storage.

Inez: I'll send the revised examples today. The batch's release status stays pending until the review is finished.

Read the phrases aloud. Then complete the conversation on the following pages using the word bank.

Conversation

The annotators disagree about what counts as a request

Two annotators labeled 100 support messages independently. They disagreed on 18, mostly messages reporting a fault without explicitly requesting help. The labeling lead and model engineer must clarify the rubric before the next training-data batch.

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

rubric / calibration sample / adjudication / dataset version / release status /
boundary case

- 01 **Cal:** Can we release the next batch today? Training is waiting for those labels.
- 02 **Inez:** Not yet. The double-labeled sample has 18 disagreements out of 100 messages.
- 03 **Cal:** Are those mostly careless mistakes?
- 04 **Inez:** No. Look at this one: "The export failed again." One annotator marked a request; the other marked a status report.
- 05 **Cal:** Our ¹ _____ says a request asks for assistance. Neither reading seems unreasonable.
- 06 **Inez:** Exactly. We need to decide whether implied requests belong in that class.
- 07 **Cal:** For this routing model, they do. Otherwise the support queue could miss a customer who expects help without asking directly.
- 08 **Inez:** Then let's add a ² _____ showing that distinction, plus a genuine status update.
- 09 **Cal:** Would "The export failed, but I've fixed it" work for the second example?
- 10 **Inez:** Yes. It reports a resolved event and doesn't seek further action. I'll make the context requirement explicit.

Conversation continues on the next page.

The annotators disagree about what counts as a request

CONVERSATION / CONTINUED

- 11 **Cal:** Can the lead annotator just relabel the disputed 18?
- 12 **Inez:** We need ³ _____, then a check for the same issue in the rest of the batch.
- 13 **Cal:** Right. Otherwise we'd fix only the examples we happened to double-label.
- 14 **Inez:** I'll also keep the original labels. We need to trace which guideline version each annotator used.
- 15 **Cal:** How will we know the revised guidance helps?
- 16 **Inez:** Use a fresh ⁴ _____ before production labeling resumes. They should label it independently first.
- 17 **Cal:** And compare agreement before discussing the answers?
- 18 **Inez:** Yes. Agreement is useful, but we'll still inspect disagreements. Two people can consistently apply an incorrect rule.
- 19 **Cal:** I'll put training on hold and record the ⁵ _____ once the corrected batch is approved.
- 20 **Inez:** I'll send the revised examples today. The batch's ⁶ _____ stays pending until the review is finished.

CHECK THE GAPS

1. rubric / 2. boundary case / 3. adjudication / 4. calibration sample / 5. dataset version / 6. release status

[Why these answers fit, page 162](#)

A new annotator encounters an ambiguous message

A new annotator cannot classify a message using the revised examples. The lead wants it routed for review without silently forcing a label.

Complete the exchange using each phrase once. Read both roles, then switch roles and repeat with the bank covered.

WORD AND PHRASE BANK

guideline / review / label / criteria

- 01 **Annotator:** This message doesn't fit either example in the current ¹ _____.
- 02 **Lead:** Add it to the ² _____ queue with the reason for uncertainty.
- 03 **Annotator:** Should I preserve the original text and my provisional ³ _____?
- 04 **Lead:** Yes. We'll record the final decision and any change to the ⁴ _____.

CHECK THE GAPS

1. guideline / 2. review / 3. label / 4. criteria

[Why these answers fit, page 162](#)

Keep the distinction audible

Resolve a labeling disagreement by revising an operational definition.

Language for the next case

The training job stopped after the last checkpoint

checkpoint

Checkpoint means saved training state from a particular point, not necessarily the latest exported weights.

Dev: Four hundred steps if we resume from the last complete checkpoint. That was step 12,000.

learning-rate schedule

Learning-rate schedule controls how the update rate changes as training progresses.

Dev: Agreed. I'll also verify the learning-rate schedule resumes at the recorded point rather than starting over.

queue wait

Queue wait is time awaiting available resources, not time spent actively training the model.

Dev: Will do. The estimate will separate compute time from the queue wait for the next reservation.

optimizer state

Optimizer state stores training-update information beyond model weights, such as accumulated momentum values.

Dev: Exactly. The complete checkpoint includes the optimizer state as well as the model weights.

reproducibility

Reproducibility concerns obtaining consistent results under documented conditions, not a universal guarantee of identical weights.

Dev: No. Matching the environment and random states helps reproducibility, but we shouldn't promise bit-for-bit identity without testing that claim.

resume point

Resume point identifies the saved state from which the intended training process can continue.

Dev: Keep it labeled as an export, not a resume point. That should prevent someone choosing it just because the timestamp is newer.

Read the phrases aloud. Then complete the conversation on the following pages using the word bank.

Conversation

The training job stopped after the last checkpoint

An interrupted training run reached step 12,400. The team has a complete checkpoint from step 12,000 and a weights-only export from step 12,300. The original environment configuration and data order are recorded.

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

optimizer state / queue wait / reproducibility / learning-rate schedule / resume point / checkpoint

- 01 **Rina:** How much work did we lose when the job stopped?
- 02 **Dev:** Four hundred steps if we resume from the last complete ¹ _____. That was step 12,000.
- 03 **Rina:** I can see a file from 12,300. Why not use the newer one?
- 04 **Dev:** That's a weights-only export. It doesn't include the optimizer or scheduler state we need for the planned continuation.
- 05 **Rina:** So it could initialize another run, but it isn't the same recovery point?
- 06 **Dev:** Exactly. The complete checkpoint includes the ² _____ as well as the model weights.
- 07 **Rina:** Is the 12,000 file intact?
- 08 **Dev:** Its checksum matches the saved record. I still need to load it in the approved environment and run a short recovery test.
- 09 **Rina:** Do that before reserving another full day of accelerators.
- 10 **Dev:** Agreed. I'll also verify the ³ _____ resumes at the recorded point rather than starting over.

Conversation continues on the next page.

The training job stopped after the last checkpoint

CONVERSATION / CONTINUED

- 11 **Rina:** What about the examples already processed after that checkpoint?
- 12 **Dev:** The saved sampler position takes us back to 12,000. The replayed steps belong to the resumed run; I won't count them as new data coverage.
- 13 **Rina:** Can we promise exactly the same final weights as an uninterrupted run?
- 14 **Dev:** No. Matching the environment and random states helps ⁴ _____, but we shouldn't promise bit-for-bit identity without testing that claim.
- 15 **Rina:** Fair. Compare the recovery test with our saved loss trace, and flag unexpected behavior.
- 16 **Dev:** I'll keep the old logs and attach a new run identifier to the continuation.
- 17 **Rina:** Give me the estimated recovery cost after that test, including the repeated steps.
- 18 **Dev:** Will do. The estimate will separate compute time from the ⁵ _____ for the next reservation.
- 19 **Rina:** And the weights-only export?
- 20 **Dev:** Keep it labeled as an export, not a ⁶ _____. That should prevent someone choosing it just because the timestamp is newer.

CHECK THE GAPS

1. checkpoint / 2. optimizer state / 3. learning-rate schedule / 4. reproducibility / 5. queue wait / 6. resume point

[Why these answers fit, page 164](#)

A run comparison uses different starting states

A researcher compares a fully resumed run with a run initialized from exported weights and a fresh optimizer. The lead asks for an accurate experiment record.

Complete the exchange using each phrase once. Read both roles, then switch roles and repeat with the bank covered.

WORD AND PHRASE BANK

condition / reinitialized / weights / comparison

- 01 **Researcher:** Both runs began with the same model ¹ _____, so I called them identical starts.
- 02 **Lead:** One run used a fresh optimizer. Record that as a changed ² _____.
- 03 **Researcher:** I'll distinguish the resumed run from the ³ _____ run.
- 04 **Lead:** Good. Keep both results, but qualify the ⁴ _____.

CHECK THE GAPS

1. weights / 2. condition / 3. reinitialized / 4. comparison

[Why these answers fit, page 164](#)

Keep the distinction audible

Negotiate a recovery plan while distinguishing saved weights from complete training state.

Language for the next case

A drafting assistant is about to become a sending agent

permission scope

Permission scope defines which actions and resources the assistant is authorized to use.

Eli: Sending changes the permission scope. What exactly would it be allowed to send, and to whom?

Least privilege

Least privilege limits available permissions to those needed for the defined task.

Eli: Remove it from the pilot. Least privilege means giving this workflow only the access it needs.

audit log

Audit log is a record of actions and approvals used to reconstruct what happened.

Eli: Add an audit log linking each approval to the action attempted and its outcome.

human approval

Human approval is an explicit decision by a person about the proposed action, not merely launching the task.

Eli: We'd still save drafting time. A human approval step lets the reviewer catch a wrong recipient or an unsupported promise.

sandbox

Sandbox means an isolated test environment designed to prevent effects on real customer accounts.

Eli: Use a sandbox with test recipients and synthetic messages first.

kill switch

Kill switch is a control that can stop the agent workflow promptly when needed.

Eli: Good. Keep a kill switch available to the named operator, and bring the evidence back before requesting external sending.

Read the phrases aloud. Then complete the conversation on the following pages using the word bank.

Conversation

A drafting assistant is about to become a sending agent

An internal assistant drafts customer follow-up emails. A proposed update would let it send messages automatically. The product team wants a pilot; the security engineer reviews the tool permissions and approval step. No external sending has been approved.

Complete the numbered gaps. Use each bank entry once; some entries contain several words.

WORD AND PHRASE BANK

audit log / human approval / sandbox / permission scope / Least privilege / kill switch

- 01 **Nora:** The drafts look good. Could we let the assistant send them during the pilot?
- 02 **Eli:** Sending changes the ¹ _____. What exactly would it be allowed to send, and to whom?
- 03 **Nora:** Follow-up messages to contacts on an approved list. An employee would start each task.
- 04 **Eli:** Starting a task isn't approval of the final recipient and content. Could the employee review those before each send?
- 05 **Nora:** Yes, though we'd lose some of the time saving.
- 06 **Eli:** We'd still save drafting time. A ² _____ step lets the reviewer catch a wrong recipient or an unsupported promise.
- 07 **Nora:** Could the model skip that step if the customer email says a reply is urgent?
- 08 **Eli:** The tool service must enforce it. A request inside customer text doesn't authorize a send.
- 09 **Nora:** Then the model prepares a proposed action, and the service waits for approval?
- 10 **Eli:** Yes. Bind the approval to that exact recipient and message. If either changes, require a new approval.

Conversation continues on the next page.

A drafting assistant is about to become a sending agent

CONVERSATION / CONTINUED

- 11 **Nora:** The existing mail integration can also delete messages. We don't need that capability.
- 12 **Eli:** Remove it from the pilot. ³ _____ means giving this workflow only the access it needs.
- 13 **Nora:** How do we test the full flow without accidentally contacting a customer?
- 14 **Eli:** Use a ⁴ _____ with test recipients and synthetic messages first.
- 15 **Nora:** I want the pilot report to show drafts accepted, drafts edited, and sends rejected by the control.
- 16 **Eli:** Add an ⁵ _____ linking each approval to the action attempted and its outcome.
- 17 **Nora:** Suppose sending succeeds but the assistant times out waiting for the response. I don't want duplicate emails.
- 18 **Eli:** We need a tested retry policy and duplicate-prevention behavior at the service boundary. The model shouldn't guess whether to send again.
- 19 **Nora:** I'll narrow the pilot proposal to drafting and reviewed sends in the test environment.
- 20 **Eli:** Good. Keep a ⁶ _____ available to the named operator, and bring the evidence back before requesting external sending.

CHECK THE GAPS

1. permission scope / 2. human approval / 3. Least privilege / 4. sandbox / 5. audit log / 6. kill switch

[Why these answers fit, page 166](#)

The approved draft changes before sending

An employee approved an email addressed to a test contact. The assistant then revised the message to include a new promise.

Complete the exchange using each phrase once. Read both roles, then switch roles and repeat with the bank covered.

WORD AND PHRASE BANK

send / outcome / content / draft

- 01 **Operator:** The ¹ _____ no longer matches what I reviewed.
- 02 **Engineer:** Then the previous approval must not authorize this ² _____.
- 03 **Operator:** I'll request a fresh review of the exact recipient and ³ _____.
- 04 **Engineer:** Record the replacement approval and the action's final ⁴ _____.

CHECK THE GAPS

1. content / 2. send / 3. draft / 4. outcome

[Why these answers fit, page 166](#)

Keep the distinction audible

Challenge excessive tool permissions and agree on a bounded agent workflow.

01 / Answers and explanations

A / Read the situation

A1. Which observation is confirmed?

B. The generated answer says 60 days.

The briefing identifies the generated text as wrong and says the interface displays it unchanged.

[Return to the exercise, page 4](#)

A2. Who owns the model integration?

B. Priya

Priya's stated role covers integration; product ownership does not identify the technical owner.

[Return to the exercise, page 4](#)

A3. What remains unknown?

C. Why the model produced the wrong deadline

The symptom and correct deadline are known, but the cause has not been established.

[Return to the exercise, page 4](#)

D / Response choices

D1. Which question locates the first mismatch?

C. At which boundary does 30 become 60?

Comparing values across boundaries locates the observed mismatch; the alternatives jump to cause, responsibility, or resolution without the necessary evidence.

[Return to the exercise, page 8](#)

D2. Complete: "Could you confirm whether ___?"

B. the payload includes the policy

After whether, an embedded question takes statement order: subject followed by verb.

[Return to the exercise, page 8](#)

01 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: The part of a system a user sees and operates. Which term fits?

A. user interface

A Correct. user interface: The part of a system a user sees and operates. The usual combination is "inspect the user interface".

B Not here. reproduction: Recreating a reported result under stated conditions. The usual combination is "obtain a reliable reproduction".

C Not here. integration: Connecting components into a working system. The usual combination is "test the integration".

D Not here. trace: A linked record of events in a request's path. The usual combination is "examine a request trace".

[Return to the exercise, page 8](#)

D4. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "So the model got it wrong? I need to tell the team who is picking this up." Priya: [missing response] Alex: "Wait, could the interface have changed the number? We had a display issue last time." Which response belongs in the gap?

A. Tell them the answer is wrong, but leave the cause open. The payload could contain another instruction or an older example.

A Correct. This is Priya's response in turn 6. It connects Alex's preceding point with Alex's follow-up. The payload is the data sent in the request; it may differ from material displayed elsewhere.

B Not here. This line belongs to turn 4, after Alex says: "Yes, the record is D-184. I checked the passage in that record: it says 30 days. The response underneath it says 60." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 8, after Alex says: "Wait, could the interface have changed the number? We had a display issue last time." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 10, after Alex says: "What would you need to distinguish a bad instruction from an unstable model answer? We should keep the Friday demonstration team informed." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 8](#)

E / Conversation

1. request ID

Turn 2. A request ID identifies the particular request whose recorded behavior can be inspected.

2. mismatch

Turn 4. A mismatch is a disagreement between values that should agree; here the two deadlines differ.

01 / Answers and explanations / continued

3. payload

Turn 6. The payload is the data sent in the request; it may differ from material displayed elsewhere.

4. upstream

Turn 8. Upstream places the observed difference before the interface in this processing path.

5. trace

Turn 10. A trace links events along the request path and helps compare what each component received or returned.

6. reproduction

Turn 12. A reproduction recreates the observed problem; it is not proof that a later change solves every variation.

7. triage

Turn 14. Triage is the initial assessment of the defect and its appropriate handling, not final release approval.

8. root cause

Turn 16. A root cause is an established underlying explanation, not merely the first plausible theory.

9. configuration

Turn 18. The configuration records the settings used, allowing results from different runs to be compared meaningfully.

10. handoff

Turn 20. A handoff passes findings and responsibility to the next decision-maker without implying the investigation is complete.

[Return to the conversation, page 10](#)

F / Accuracy and scope

F1. Which statement preserves uncertainty?

A. The trace suggests a generation issue; the cause is unconfirmed.

Suggests reports limited evidence, while the second clause explicitly leaves causation unresolved.

[Return to the exercise, page 12](#)

01 / Answers and explanations / continued

F2. Which expression means transferring work with context?

A. prepare a clear handoff

A handoff transfers work and relevant information; the other combinations do not express that action.

[Return to the exercise, page 12](#)

F / Guided speaking

1. user interface

Engineer: The model response contains the citation; the user interface does not display it.

Turn 1. The discrepancy appears in presentation because the generated response already contains the citation.

2. request ID

Lead: Attach the request ID so the display team can locate that exchange.

Turn 2. The identifier locates the particular exchange rather than merely describing the application.

3. symptom

Engineer: That is the observed symptom; we have not proved why it occurred.

Turn 3. An omitted link is an observed sign; its underlying cause still needs investigation.

4. handoff

Lead: Then send a clear handoff to the display owner.

Turn 4. The next team needs the evidence and unresolved question to take over the investigation.

[Return to the conversation, page 13](#)

01 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "Please make the ownership clear in the ticket. I can coordinate the demonstration, but I cannot approve production use on my own." Priya: [missing response] Alex: "Can I tell the sales team we will have a fix by Thursday afternoon? They need to decide whether to keep the live demonstration." Which response belongs in the gap?

D. I will own technical triage. I will also name the policy and interface owners if the evidence requires their input.

A Not here. This line belongs to turn 20, after Alex says: "That works. Please send it by three on Thursday, even if the investigation is incomplete. I will confirm who can decide about Friday." It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 2, after Alex says: "The refund answer is wrong again. The policy says 30 days, but our assistant tells the customer 60. Which team should investigate first?" It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 10, after Alex says: "What would you need to distinguish a bad instruction from an unstable model answer? We should keep the Friday demonstration team informed." It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Priya's response in turn 14. It connects Alex's preceding point with Alex's follow-up. Triage is the initial assessment of the defect and its appropriate handling, not final release approval.

[Return to the exercise, page 13](#)

02 / Answers and explanations

A / Read the situation

A1. What is the total requested reservation?

D. 33,000 tokens

The input contains 29,000 tokens and the output reservation adds 4,000, giving 33,000.

[Return to the exercise, page 14](#)

A2. Why was the request rejected?

C. It exceeds the configured total context limit.

Rejection happened before generation because the combined reservation exceeds 32,000 tokens.

[Return to the exercise, page 14](#)

A3. What must the team preserve?

C. The exception clauses

The briefing requires the exception clauses to remain available; it does not require duplicate material.

[Return to the exercise, page 14](#)

D / Response choices

D1. Which response explains the rejection?

A. The combined reservation is 1,000 tokens over the limit.

The request totals 33,000 against a 32,000 limit; generation has not begun.

[Return to the exercise, page 18](#)

D2. Which proposal preserves the stated requirement most directly?

A. Remove exact duplicate headers and recount.

Removing exact duplicates can reduce size without intentionally deleting the required exceptions; the result still needs checking.

[Return to the exercise, page 18](#)

02 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: The component that converts text into tokens. Which term fits?

C. tokenizer

A Not here. system instruction: Higher-priority guidance supplied by an application's configuration. The usual combination is "inspect the system instruction".

B Not here. position: A token's place within the input sequence. The usual combination is "vary evidence position".

C Correct. tokenizer: The component that converts text into tokens. The usual combination is "use the model's tokenizer".

D Not here. truncation: Removal of content to fit a size limit. The usual combination is "detect silent truncation".

[Return to the exercise, page 18](#)

D4. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "All right, I had treated the limit as input only. Could we simply trim the final document until the request is accepted?" Omar: [missing response] Lina: "We also repeat the same header on every page. Removing those exact copies would preserve the policy wording and reduce the input somewhat." Which response belongs in the gap?

B. Not without checking what it contains. The final pages include an exception clause that changes the refund rule for renewals.

A Not here. This line belongs to turn 12, after Lina says: "I'll check renewals and exceptions against the same checklist. Getting the request accepted won't help if the summary drops the important part." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Omar's response in turn 6. It connects Lina's preceding point with Lina's follow-up. An exception clause changes the general rule, so deleting it can change the meaning of the policy.

C Not here. This line belongs to turn 2, after Lina says: "The summary request is rejected again. I was going to increase the output allowance so the model has enough room to explain the policy." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 4, after Lina says: "That sounds close to the 32,000 limit. Are you saying the unused output allowance counts even though the model has not answered yet?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 18](#)

E / Conversation

1. token budget

Turn 2. A token budget allocates capacity to the request; checking it reveals whether the proposed change fits.

02 / Answers and explanations / continued

2. reservation

Turn 4. The reservation combines supplied input and capacity set aside for output under this case's stated rule.

3. exception clause

Turn 6. An exception clause changes the general rule, so deleting it can change the meaning of the policy.

4. deduplication

Turn 8. Deduplication removes repeated copies; it is distinct from summarizing or deleting unique policy content.

5. headroom

Turn 10. Headroom is the remaining capacity below the limit: 32,000 minus 28,400 minus 3,000 equals 600.

6. truncation

Turn 12. Truncation removes content to satisfy a length limit; recording it exposes possible information loss.

7. long-context

Turn 14. Long-context describes capacity for larger inputs, not a guarantee that relevant details will be used accurately.

8. position

Turn 16. Position means where the evidence appears in the input; changing it can reveal sensitivity to document order.

9. delimiters

Turn 18. Delimiters mark document or section boundaries; retaining them helps preserve the structure of the supplied context.

10. capacity constraint

Turn 20. A capacity constraint limits what fits; it is not a finding about the truth or quality of the policy answer.

[Return to the conversation, page 20](#)

02 / Answers and explanations / continued

F / Accuracy and scope

F1. Complete: "We can reduce the allowance, provided that ____."

C. the summary still covers the required points

Provided that introduces a condition with normal subject-verb order.

[Return to the exercise, page 22](#)

F2. Which claim is appropriately limited?

B. The request fits; we still need to test summary coverage.

Capacity compliance and content quality are separate checks; fitting does not prove coverage.

[Return to the exercise, page 22](#)

F / Guided speaking

1. reservation

Engineer: The combined reservation is 21,000 tokens, so the request is too large.

Turn 1. Adding 18,000 input and 3,000 reserved output gives 21,000, above the case limit.

2. deduplication

Reviewer: Start with exact deduplication of the repeated covers.

Turn 2. Removing exact repeated covers targets redundancy without intentionally removing unique obligations.

3. obligation

Engineer: I will preserve every unique obligation in the supplied contract.

Turn 3. The constraint explicitly protects unique obligations; removing them would change the source information.

4. coverage

Reviewer: After recounting, test coverage rather than treating acceptance as success.

Turn 4. The system accepting the request does not establish that the summary includes the required obligations.

[Return to the conversation, page 23](#)

02 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "Someone suggested using a larger-context model instead. Would that let us keep the original request and stop spending time on these smaller adjustments?" Omar: [missing response] Lina: "Then I will keep that as a separate comparison. Should the test vary where the exception appears, or is the current document order sufficient?" Which response belongs in the gap?

D. It is another option, not an automatic solution. Its long-context performance on our policy task has not been evaluated, and its cost may differ.

A Not here. This line belongs to turn 8, after Lina says: "We also repeat the same header on every page. Removing those exact copies would preserve the policy wording and reduce the input somewhat." It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 20, after Lina says: "I will report the before-and-after counts with the coverage results. If the necessary clauses still cannot fit, I will stop rather than silently remove them." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 10, after Lina says: "Suppose that saves only 600 tokens. We would still be 400 over the limit. Could we reserve 3,000 output tokens instead?" It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Omar's response in turn 14. It connects Lina's preceding point with Lina's follow-up. Long-context describes capacity for larger inputs, not a guarantee that relevant details will be used accurately.

[Return to the exercise, page 23](#)

03 / Answers and explanations

A / Read the situation

A1. What does 94 percent currently describe?

C. The score on all 500 evaluation examples

The reported score covers the original 500 examples; the nonoverlapping subset has not been separately scored.

[Return to the exercise, page 27](#)

A2. What overlap has been confirmed?

B. Forty evaluation records have exact training copies.

The confirmed issue is exact overlap for 40 evaluation records, not contamination of every record.

[Return to the exercise, page 27](#)

A3. What can the release owner do?

B. Postpone the decision

The briefing explicitly allows postponement; none of the other actions creates the missing evidence.

[Return to the exercise, page 27](#)

D / Response choices

D1. Which meeting statement is supported?

C. The 94 percent score includes 40 training overlaps.

The overlap is confirmed; no clean-subset score or production outcome has been established.

[Return to the exercise, page 31](#)

D2. What does "not yet re-scored" mean?

D. The revised calculation remains unfinished.

Not yet states timing and incompleteness, not the direction or final consequence of the result.

[Return to the exercise, page 31](#)

03 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: Data reserved for evaluating a finalized choice. Which term fits?

B. test set

A Not here. distribution shift: A difference between data distributions across settings. The usual combination is "detect distribution shift".

B Correct. test set: Data reserved for evaluating a finalized choice. The usual combination is "preserve the test set".

C Not here. inter-rater agreement: The degree to which different raters agree. The usual combination is "measure inter-rater agreement".

D Not here. near duplicate: A closely similar record that may share substantive content. The usual combination is "review near duplicates".

[Return to the exercise, page 31](#)

D4. Rebuild the exchange from "A 94 percent score with an overlap problem".

Morgan: "Then I will remove the independent-evaluation label. Can you tell me whether the overlap came from the import process or the later cleanup?" **Jules:** [missing response] **Morgan:** "Send me that list. I'll leave the cause out of the slide for now. Is there anything else I need to flag?" **Which response belongs in the gap?**

C. Not yet. The provenance log should help, but I have not reconstructed the sequence. I can provide the exact record list now.

A Not here. This line belongs to turn 2, after Morgan says: "The release meeting starts in an hour. I have the 94 percent result on the slide. Is there anything material we need to add?" It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 4, after Morgan says: "Forty out of 500 is eight percent. Could I subtract eight points and present 86 percent as a conservative estimate instead?" It addresses that point, not the missing turn in the supplied exchange.

C Correct. This is Jules's response in turn 6. It connects Morgan's preceding point with Morgan's follow-up. Provenance records where data came from and how it changed; it can help identify when overlap entered the sets.

D Not here. This line belongs to turn 14, after Morgan says: "Once the duplicate records are removed, can we say the evaluation represents customer traffic? That is the next question I expect from the release owner." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 31](#)

E / Conversation

1. contamination

Turn 2. Contamination means the evaluation is affected by unwanted exposure or overlap with training data.

03 / Answers and explanations / continued

2. re-score

Turn 4. Re-scoring calculates performance on the specified remaining records; arithmetic on the overlap rate cannot supply that score.

3. provenance

Turn 6. Provenance records where data came from and how it changed; it can help identify when overlap entered the sets.

4. adjudication

Turn 8. Adjudication resolves disputed labels through the agreed review process rather than selecting whichever label improves the score.

5. denominator

Turn 10. The denominator identifies how many records are included, so exclusions cannot be hidden behind an unchanged headline rate.

6. holdout

Turn 12. A holdout must remain separate from the process being evaluated; the confirmed training overlap compromises that separation.

7. representativeness

Turn 14. Representativeness concerns whether the sample reflects the intended customer population, a separate question from overlap.

8. exact duplicates

Turn 16. Exact duplicates are identical under the chosen comparison rule; verifying them does not complete a near-duplicate review.

9. model selection

Turn 18. Model selection uses evaluation results to choose a model or configuration, so that set no longer provides an untouched final test.

10. limitation statement

Turn 20. A limitation statement specifies what the evidence cannot support, preventing the headline figure from becoming an unjustified release claim.

[Return to the conversation, page 33](#)

03 / Answers and explanations / continued

F / Accuracy and scope

F1. Which question targets a label dispute?

A. Which rubric criterion distinguishes these two labels?

The rubric provides criteria for adjudication. Predictions, duplicate removal and an aggregate score do not resolve what the reference category should be.

[Return to the exercise, page 35](#)

F2. Which phrase correctly names the base of a rate?

A. state the denominator

The denominator is the total count against which the number of successes is expressed.

[Return to the exercise, page 35](#)

F / Guided speaking

1. overlap

Analyst: The current test has training overlap, so its independence is compromised.

Turn 1. Ten test messages also appeared in training, which compromises the claimed separation.

2. re-score

Lead: Report the clean-subset score only after you re-score those 70 messages.

Turn 2. The performance on the remaining 70 cannot be inferred simply by subtracting the overlap count.

3. adjudication

Analyst: The three disputed labels also need adjudication.

Turn 3. Label disagreements require a defined review decision separate from duplicate removal.

4. denominator

Lead: And include the denominator with every reported percentage.

Turn 4. A percentage must identify its underlying number of examples, especially after exclusions.

[Return to the conversation, page 36](#)

03 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "A 94 percent score with an overlap problem". Morgan: "Once the duplicate records are removed, can we say the evaluation represents customer traffic? That is the next question I expect from the release owner." Jules: [missing response] Morgan: "So we have three issues: overlap, disputed labels, and coverage of the target population. Which one can you resolve before today's meeting starts?" Which response belongs in the gap?

D. Not from separation alone. We still need to check representativeness, including the rare category that appears only a few times in this sample.

A Not here. This line belongs to turn 6, after Morgan says: "Then I will remove the independent-evaluation label. Can you tell me whether the overlap came from the import process or the later cleanup?" It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 18, after Morgan says: "All right, the duplicate check comes first. For the replacement set, can we keep it untouched while the team tunes the model?" It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 12, after Morgan says: "Agreed. The release owner can wait for better evidence. What would you recommend we say about the 94 percent figure already circulated yesterday?" It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Jules's response in turn 14. It connects Morgan's preceding point with Morgan's follow-up. Representativeness concerns whether the sample reflects the intended customer population, a separate question from overlap.

[Return to the exercise, page 36](#)

04 / Answers and explanations

A / Read the situation

A1. What improved in the fixed test?

A. Queries with a relevant retrieved passage rose from 78 to 86.

The measured gain concerns retrieval relevance on 100 fixed queries; answer quality and user preference remain unchecked.

[Return to the exercise, page 38](#)

A2. Which component stayed unchanged?

D. The generator

The briefing holds the generator and query set constant, helping isolate the retrieval change.

[Return to the exercise, page 38](#)

A3. Which claim needs further evidence?

D. The new path improves end-to-end answer quality.

No end-to-end answer review has been completed, so better passage retrieval does not establish that claim.

[Return to the exercise, page 38](#)

D / Response choices

D1. Which comparison correctly states the observed gain?

B. Coverage increased by eight percentage points on the fixed set.

The measure rises from 78 of 100 to 86 of 100, an eight-percentage-point coverage change.

[Return to the exercise, page 42](#)

D2. What is a reranker ablation here?

B. Compare the path with and without the reranker while holding other factors fixed.

Changing the reranker alone helps attribute differences in the matched comparison.

[Return to the exercise, page 42](#)

04 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: A store or index supporting retrieval of vector representations. Which term fits?

C. vector store

A Not here. relevance: How well material addresses a specified query. The usual combination is "judge passage relevance".

B Not here. chunk: A retrievable segment of a larger document. The usual combination is "retrieve a relevant chunk".

C Correct. vector store: A store or index supporting retrieval of vector representations. The usual combination is "query a vector store".

D Not here. index freshness: How current the indexed information is. The usual combination is "monitor index freshness".

[Return to the exercise, page 42](#)

D4. Rebuild the exchange from "More relevant passages, a slower response". Sam: "I agree the answer review is unfinished. My recommendation is a small pilot, not a rollout. Which evidence would you need before approving that pilot?" Elena: [missing response] Sam: "We can collect those. Some of the original misses had the right document in the candidates but ranked below the five passages sent to generation." Which response belongs in the gap?

B. First, an end-to-end review of the same cases. We need the final answers, their supporting passages, and the complete response times.

A Not here. This line belongs to turn 18, after Sam says: "Then I will present the retrieval result as eight percentage points on this set. I will not call it an eight-percent increase in customer satisfaction." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Elena's response in turn 6. It connects Sam's preceding point with Sam's follow-up. End-to-end review follows the whole request-to-answer path, rather than stopping at the retrieval component.

C Not here. This line belongs to turn 8, after Sam says: "We can collect those. Some of the original misses had the right document in the candidates but ranked below the five passages sent to generation." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 2, after Sam says: "The reranker looks promising. We found a relevant passage for 86 of the same 100 queries, compared with 78 on the previous path." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 42](#)

E / Conversation

1. candidate set

Turn 2. The candidate set contains the retrieved items available to the reranker; changes there could affect the comparison.

04 / Answers and explanations / continued

2. ablation

Turn 4. An ablation compares a component's presence or configuration while controlling other factors, helping isolate its effect.

3. end-to-end

Turn 6. End-to-end review follows the whole request-to-answer path, rather than stopping at the retrieval component.

4. reranker

Turn 8. A reranker can reorder existing candidates; it cannot recover a document absent from the candidate set by reordering alone.

5. chunk boundary

Turn 10. A chunk boundary can separate a main rule from its exception, making a retrieved segment incomplete without neighboring context.

6. grounding

Turn 12. Grounding asks whether the generated claim is supported by the supplied evidence, not merely whether a citation marker appears.

7. index freshness

Turn 14. Index freshness concerns whether indexed material reflects current source information; relevance alone does not establish currency.

8. trade-off

Turn 16. A trade-off involves competing outcomes, here relevance and delay; user acceptance requires evidence beyond the retrieval score.

9. query type

Turn 18. Query type identifies a meaningful subgroup; subgroup timing can reveal costs concealed by an overall average.

10. conditional

Turn 20. Conditional support depends on the stated review being completed; it is not unconditional permission to deploy.

[Return to the conversation, page 44](#)

04 / Answers and explanations / continued

F / Accuracy and scope

F1. Which sentence preserves the limitation?

D. Retrieval improved on this set; final-answer quality remains unmeasured.

The sentence separates the measured gain from an unmeasured downstream outcome.

[Return to the exercise, page 46](#)

F2. Which question checks grounding?

C. Does this passage support the answer's refund exception?

Grounding concerns support for the specific claim. A marker, indexed document or lexical overlap can exist without supporting the stated exception.

[Return to the exercise, page 46](#)

F / Guided speaking

1. source

Reviewer: The citation identifies a real source, but not support for this claim.

Turn 1. Identifying an authentic source does not establish that the cited material supports the particular answer.

2. grounding

Engineer: Both sections were retrieved, so we should inspect answer grounding.

Turn 2. With both sections available, the question is whether the generated claim reflects the supporting evidence.

3. context

Reviewer: The exception is missing from the answer, although it is present in the context.

Turn 3. The provided retrieval context contains the exception; the failure is not evidence that it was absent.

4. end-to-end

Engineer: We will report that distinction in the end-to-end review.

Turn 4. The full review follows the evidence through retrieval into the final answer, preserving where the mismatch appears.

[Return to the conversation, page 47](#)

04 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Understood. I will include the exact cited passage and the answer claim together. We also need to distinguish an old policy from an irrelevant one." Elena: [missing response] Sam: "The product lead will ask whether customers will accept the extra delay. Could we infer that from the eight-point increase in passage coverage?" Which response belongs in the gap?

A. Yes. Check index freshness separately. A highly relevant passage from a superseded policy can still produce a misleading answer.

A Correct. This is Elena's response in turn 14. It connects Sam's preceding point with Sam's follow-up. Index freshness concerns whether indexed material reflects current source information; relevance alone does not establish currency.

B Not here. This line belongs to turn 16, after Sam says: "The product lead will ask whether customers will accept the extra delay. Could we infer that from the eight-point increase in passage coverage?" It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 6, after Sam says: "I agree the answer review is unfinished. My recommendation is a small pilot, not a rollout. Which evidence would you need before approving that pilot?" It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 10, after Sam says: "Two misses also involve a policy exception split across adjacent chunks. I do not want to alter chunking in the middle of this comparison." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 47](#)

05 / Answers and explanations

A / Read the situation

A1. Which fact directly concerns outdated information?

A. The retrieval index contains obsolete policy wording.

The index supplies obsolete source wording, which is a distinct issue from output formatting.

[Return to the exercise, page 51](#)

A2. What is known about the 200 examples?

D. They are permission-cleared but consistency is unaudited.

Permission to use data does not establish its consistency, currency, or training value.

[Return to the exercise, page 51](#)

A3. What distinction should the experiment preserve?

D. Source freshness versus output-format behavior

The two observed failures concern source currency and output behavior. The alternatives conflate separate checks or assume an untested mechanism.

[Return to the exercise, page 51](#)

D / Response choices

D1. Which action directly addresses the known stale source?

D. Correct the retrieval index through the approved update process.

The obsolete wording is in retrieval; a source update addresses that known mechanism without assuming training will refresh it.

[Return to the exercise, page 55](#)

D2. Which is a behavioral target?

D. Returning the required fields consistently

Format adherence is an observable output behavior; the other items do not specify model behavior.

[Return to the exercise, page 55](#)

05 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: Abbreviation for supervised fine-tuning. Which term fits?

C. SFT

A Not here. few-shot prompting: Supplying a small set of examples in the prompt. The usual combination is "test few-shot prompting".

B Not here. epoch: One pass through the training data. The usual combination is "record the epoch count".

C Correct. SFT: Abbreviation for supervised fine-tuning. The usual combination is "compare SFT with prompting".

D Not here. LoRA: Low-rank adaptation; a parameter-efficient adaptation method. The usual combination is "train a LoRA adapter".

[Return to the exercise, page 55](#)

D4. Rebuild the exchange from "One proposal, two different failure modes". Tessa: "For formatting, we have 200 example summaries cleared for training use. I can start preparing them, although we have not audited their labels yet." Dev: [missing response] Tessa: "Would you want a prompting comparison before training? The current prompt includes only one example, and it does not show how to handle an absent field." Which response belongs in the gap?

A. Please do that audit first. Permission to use an example does not establish its consistency with the field schema or the instructions in the other examples.

A Correct. This is Dev's response in turn 6. It connects Tessa's preceding point with Tessa's follow-up. Consistency means examples follow compatible criteria; permission alone says nothing about agreement among labels or formats.

B Not here. This line belongs to turn 10, after Tessa says: "All right. Two work items, then: repair the index and compare formatting methods. I'll keep their results separate." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 16, after Tessa says: "If we test an adapter, I also want to check our other summary tasks. Better performance on refunds could conceal a regression in cancellation summaries." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 2, after Tessa says: "I think we should fine-tune the summarizer. It ignores our field structure and sometimes quotes the old refund policy. Training could make it more consistent." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 55](#)

E / Conversation

1. failure modes

Turn 2. Failure modes identify distinct ways a system fails; wrong structure and stale factual content need not share a cause.

05 / Answers and explanations / continued

2. behavioral target

Turn 4. A behavioral target specifies the output behavior to change, here adherence to the required fields.

3. consistency

Turn 6. Consistency means examples follow compatible criteria; permission alone says nothing about agreement among labels or formats.

4. few-shot

Turn 8. Few-shot prompting supplies a small set of examples at inference time, providing a useful comparison with parameter updates.

5. format adherence

Turn 10. Format adherence measures conformity to the required structure, which must remain distinct from factual correctness.

6. schema

Turn 12. A schema defines the required structure and allowed fields; it does not authorize inventing unsupported values.

7. generalization

Turn 14. Generalization concerns performance beyond the examples used to fit or choose the method.

8. catastrophic forgetting

Turn 16. Catastrophic forgetting names loss of previously learned capabilities during later training; a targeted gain does not rule it out.

9. rollback candidate

Turn 18. A rollback candidate is a previously checked configuration available if the new adaptation must be withdrawn.

10. hypotheses

Turn 20. Hypotheses are proposed explanations or outcomes to test; they are not established findings before the comparison.

[Return to the conversation, page 57](#)

05 / Answers and explanations / continued

F / Accuracy and scope

F1. Which claim needs qualification?

A. Fine-tuning will solve both failures.

The proposed remedy has not been evaluated and does not directly maintain source freshness.

[Return to the exercise, page 59](#)

F2. Which comparison best isolates the formatting intervention?

C. Keep sources and test cases fixed while changing the formatting method.

Holding sources and cases fixed makes the effect of the formatting change easier to interpret.

[Return to the exercise, page 59](#)

F / Guided speaking

1. category list

Scientist: The prompt supplies an obsolete category list, so we should correct that input first.

Turn 1. The known discrepancy is the six-category instruction, not evidence that a training change is already required.

2. baseline

Lead: Then compare the corrected prompt with the existing baseline.

Turn 2. A baseline provides the reference against which the effect of the corrected prompt can be measured.

3. held-out

Scientist: We will measure adherence on held-out examples, not only the demonstrations.

Turn 3. Examples kept outside fitting and selection provide a more independent check than examples used to demonstrate the task.

4. hypothesis

Lead: Until then, the benefit of further training remains a hypothesis.

Turn 4. No training comparison has been completed, so its benefit is a proposed outcome rather than a finding.

[Return to the conversation, page 60](#)

05 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "One proposal, two different failure modes". Tessa: "I will include that distinction in the rubric. After we select the method, we should evaluate on cases that were never used to tune it." Dev: [missing response] Tessa: "If we test an adapter, I also want to check our other summary tasks. Better performance on refunds could conceal a regression in cancellation summaries." Which response belongs in the gap?

D. Correct. We need evidence of generalization, including summaries with exceptions, missing details, and layouts absent from the training examples.

A Not here. This line belongs to turn 20, after Tessa says: "I'll start with the example audit and the prompting baseline. Can we review those results before we spend time on an adapter?" It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 8, after Tessa says: "Would you want a prompting comparison before training? The current prompt includes only one example, and it does not show how to handle an absent field." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 18, after Tessa says: "I will version the examples, configuration, and resulting adapter together. That way the team can reproduce the comparison and identify exactly what was evaluated." It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Dev's response in turn 14. It connects Tessa's preceding point with Tessa's follow-up. Generalization concerns performance beyond the examples used to fit or choose the method.

[Return to the exercise, page 60](#)

06 / Answers and explanations

A / Read the situation

A1. What happens to total passes?

C. They fall from 90 to 88.

The 12-to-18 improvement applies only to long documents; total passes decline by two.

[Return to the exercise, page 62](#)

A2. Which slice improves?

A. Long-document cases

Long-document passes rise from 12 to 18; short-request passes fall from 78 to 70.

[Return to the exercise, page 62](#)

A3. What remains unknown?

C. The production mixture and seriousness of failures

The briefing supplies test counts and an unchanged rubric, but not production proportions or severity analysis.

[Return to the exercise, page 62](#)

D / Response choices

D1. Which summary preserves both findings?

D. Total passes fell by two; long-document passes rose by six.

The statement correctly separates the overall count from the improved subgroup and makes no production claim.

[Return to the exercise, page 66](#)

D2. What is the candidate's long-document pass rate?

A. 90 percent

Eighteen divided by twenty equals 90 percent; 88 percent is the aggregate test rate.

[Return to the exercise, page 66](#)

06 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: The version under consideration. Which term fits?

C. candidate

A Not here. acceptance criterion: A stated condition required for acceptance. The usual combination is "agree on acceptance criteria".

B Not here. paired comparison: Comparison of results on corresponding cases. The usual combination is "run a paired comparison".

C Correct. candidate: The version under consideration. The usual combination is "evaluate the candidate".

D Not here. false negative: A negative decision when the reference outcome is positive. The usual combination is "review false negatives".

[Return to the exercise, page 66](#)

D4. Rebuild the exchange from "The average is down, but one slice improved". Ben: "Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration." Noor: [missing response] Ben: "The production proportions have not been analyzed. We can say that plainly. I can also list which short-request cases changed from pass to fail." Which response belongs in the gap?

A. Difficulty alone does not determine the weight. We need the intended traffic mixture and agreed acceptance criteria, not weights chosen after seeing the result.

A Correct. This is Noor's response in turn 6. It connects Ben's preceding point with Ben's follow-up. Acceptance criteria are the conditions for approving a version, ideally set before selecting whichever metric looks favorable.

B Not here. This line belongs to turn 20, after Ben says: "I'll recommend waiting for the changed-case review. I still think routing is worth testing, but I won't present it as a fix today." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 18, after Ben says: "I will revise the slide to show both counts, the shared rubric, and the missing traffic and severity analysis. The candidate's strengths will still be visible." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 16, after Ben says: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 66](#)

E / Conversation

1. aggregate

Turn 2. The aggregate combines all test cases, so it prevents a favorable subgroup from standing in for the whole result.

06 / Answers and explanations / continued

2. regression

Turn 4. A regression is a worsening of measured behavior; here eight fewer short-request cases pass.

3. acceptance criteria

Turn 6. Acceptance criteria are the conditions for approving a version, ideally set before selecting whichever metric looks favorable.

4. paired comparison

Turn 8. A paired comparison examines corresponding cases across versions, revealing which individual outcomes changed rather than only comparing totals.

5. severity

Turn 10. Severity describes how consequential a failure is; equal pass-fail counts need not imply equal user impact.

6. human review

Turn 12. Human review applies the stated criteria to actual outputs and can examine judge disagreements in the changed subgroup.

7. false positives

Turn 14. False positives in the proposed router would send requests into the long-document path when they do not meet its intended condition.

8. statistical significance

Turn 16. Statistical significance concerns evidence under a statistical model; it does not alone establish that a difference matters operationally.

9. disaggregation

Turn 18. Disaggregation separates the combined result into defined groups, making different patterns visible without discarding the overall result.

10. release gate

Turn 20. A release gate is the decision checkpoint whose criteria must be met; presenting an improved slice does not automatically satisfy it.

[Return to the conversation, page 68](#)

06 / Answers and explanations / continued

F / Accuracy and scope

F1. Which question targets practical significance?

C. How serious are the newly failing cases for users?

Consequences for users address practical importance. Stability, arithmetic reconciliation and rubric consistency are also useful, but answer different questions.

[Return to the exercise, page 70](#)

F2. Which proposed conclusion exceeds the evidence?

C. Long-document routing is already safe to release.

No routing or severity assessment has established that a specialized release is acceptable.

[Return to the exercise, page 70](#)

F / Guided speaking

1. aggregate

Reviewer: The overall aggregate hides a weak urgent-message subgroup.

Turn 1. A combined score can conceal poor results in a smaller but important subgroup.

2. false negatives

Engineer: Those missed urgent messages are false negatives for that category.

Turn 2. The messages are urgent but the classifier fails to identify them as positive urgent cases.

3. severity

Reviewer: We need a separate assessment of their severity.

Turn 3. The seriousness of missing urgent messages cannot be inferred from the overall success count.

4. release gate

Engineer: Then the release gate should remain unresolved until that review is complete.

Turn 4. The approval checkpoint lacks relevant evidence and should not be treated as passed.

[Return to the conversation, page 71](#)

06 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "The average is down, but one slice improved". Ben: "If those reviews confirm the long-document gain, could we route only long documents to the candidate and keep short requests on the previous version?" Noor: [missing response] Ben: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" Which response belongs in the gap?

B. It is a useful proposal, but the routing decision becomes another component to evaluate. We have not measured its false positives or the consequences of misclassification.

A Not here. This line belongs to turn 6, after Ben says: "Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Noor's response in turn 14. It connects Ben's preceding point with Ben's follow-up. False positives in the proposed router would send requests into the long-document path when they do not meet its intended condition.

C Not here. This line belongs to turn 16, after Ben says: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 10, after Ben says: "Would you want a severity category for every changed case? A refusal on a valid request and a confident false answer both count as failures now." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 71](#)

07 / Answers and explanations

A / Read the situation

A1. Which result is measured?

C. Mean response time falls from four seconds to two.

The case reports a mean on a fixed test, not a guarantee for every request or an unmeasured workload.

[Return to the exercise, page 75](#)

A2. What traffic is permitted in the internal pilot?

A. Synthetic requests only

The stated pilot boundary allows synthetic requests only; broader traffic needs separate authorization.

[Return to the exercise, page 75](#)

A3. Which responsibility is unresolved?

D. Who owns rollback

The case names operational and product owners and a previous configuration, but no rollback owner.

[Return to the exercise, page 75](#)

D / Response choices

D1. Which claim is supported by the test?

B. Mean latency decreased by 50 percent.

The mean falls from four to two, a reduction of half; the other claims concern unmeasured outcomes.

[Return to the exercise, page 79](#)

D2. Which question distinguishes first visible output from a finished response?

C. Is that first-token time or complete-answer time?

First-token and completion times identify the two user-visible events in the question. The other comparisons describe distributions or different serving measures.

[Return to the exercise, page 79](#)

07 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: The 95th percentile of a measured distribution. Which term fits?

B. P95

A Not here. error budget: The permitted unreliability associated with a service objective. The usual combination is "track error-budget consumption".

B Correct. P95: The 95th percentile of a measured distribution. The usual combination is "inspect P95 latency".

C Not here. load test: A test under specified levels of demand. The usual combination is "run a controlled load test".

D Not here. batching: Processing multiple requests together. The usual combination is "evaluate dynamic batching".

[Return to the exercise, page 79](#)

D4. Rebuild the exchange from "Faster on average is not ready for everyone". Luis: "Fair point. For the internal pilot we can stay with synthetic requests. Does streaming let us demonstrate a faster answer without changing the model again?" Maya: [missing response] Luis: "What do you need from product to define a useful load test? We have examples of simple and complex requests, but not a production traffic forecast." Which response belongs in the gap?

C. It may improve time to first token, but we should report completion time separately. Seeing the first words is not the same as receiving the complete response.

A Not here. This line belongs to turn 18, after Luis says: "Can you provide the distribution, complex-answer results, and cost assumptions by Friday? I can postpone the external claim until we have that package." It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 8, after Luis says: "What do you need from product to define a useful load test? We have examples of simple and complex requests, but not a production traffic forecast." It addresses that point, not the missing turn in the supplied exchange.

C Correct. This is Maya's response in turn 6. It connects Luis's preceding point with Luis's follow-up. Time to first token measures the initial wait, not the time needed to receive and verify the whole answer.

D Not here. This line belongs to turn 2, after Luis says: "The smaller model cuts average response time from four seconds to two. I would like to use it in the pilot and announce the faster experience." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 79](#)

E / Conversation

1. tail latency

Turn 2. Tail latency describes the slower end of the response-time distribution, which an improved average does not establish.

07 / Answers and explanations / continued

2. mean

Turn 4. The mean is an arithmetic average; it is not an upper bound on individual request times.

3. time to first token

Turn 6. Time to first token measures the initial wait, not the time needed to receive and verify the whole answer.

4. concurrency

Turn 8. Concurrency is the amount of simultaneous work; varying it helps expose waiting and resource contention.

5. fallback

Turn 10. A fallback is an alternate serving path; its extra work can affect both cost and observed response time.

6. rollback

Turn 12. Rollback restores a previous configuration; availability of that configuration does not establish who is authorized to restore it.

7. stop condition

Turn 14. A stop condition states when the pilot should pause, avoiding an improvised decision after a concerning result appears.

8. rollout

Turn 16. A rollout introduces the configuration to its intended users; it requires its own scope and approval, beyond this synthetic test.

9. queueing

Turn 18. Queueing is waiting before processing resources are available; it can dominate latency even when model inference itself is faster.

10. limitations

Turn 20. Limitations identify the conditions and missing evidence that constrain how widely a measured result can be applied.

[Return to the conversation, page 81](#)

07 / Answers and explanations / continued

F / Accuracy and scope

F1. Which statement preserves authorization boundaries?

D. The synthetic pilot does not authorize production traffic.

The case explicitly separates the internal pilot from production approval.

[Return to the exercise, page 83](#)

F2. Which expression means restoring the earlier configuration?

B. initiate rollback

Rollback returns the system to a previous configuration; the other actions change different aspects of serving.

[Return to the exercise, page 83](#)

F / Guided speaking

1. time to first token

Engineer: One second is the time to first token, not the time for the full response.

Turn 1. The first visible token arrives at one second; the full answer takes six seconds.

2. completion time

Lead: Then show completion time separately so the comparison is clear.

Turn 2. Completion time captures the wait for the entire answer and must not be replaced by the first-token measure.

3. service target

Engineer: Neither figure is an approved service target.

Turn 3. The scenario explicitly says no target has been approved; measurements do not themselves authorize a promise.

4. qualification

Lead: Keep that qualification attached when the slide is shared.

Turn 4. The limiting explanation prevents readers from treating a one-second initial display as full-answer performance.

[Return to the conversation, page 84](#)

07 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "Faster on average is not ready for everyone". Luis: "I can own the decision to pause the product pilot. Can serving operations own the restoration step, with you as the named contact for this test?" Maya: [missing response] Luis: "I'll keep production traffic out. When the synthetic test is done, we'll take the results to the production review. Who needs that report?" Which response belongs in the gap?

D. Yes, subject to our operational review. Record the stop condition and contact route in the pilot note so both teams know how to act.

A Not here. This line belongs to turn 16, after Luis says: "I'll keep production traffic out. When the synthetic test is done, we'll take the results to the production review. Who needs that report?" It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 10, after Luis says: "Include the complicated cases even if they are a small share. A fast answer that loses a key exception would undermine the whole demonstration." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 4, after Luis says: "Could we call it a two-second assistant if we explain that the figure is an average? Marketing would prefer a simple phrase for the demonstration." It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Maya's response in turn 14. It connects Luis's preceding point with Luis's follow-up. A stop condition states when the pilot should pause, avoiding an improvised decision after a concerning result appears.

[Return to the exercise, page 84](#)

08 / Answers and explanations

A / Read the situation

A1. What has been confirmed?

B. One reproduction in an authorized synthetic test

The case confirms one synthetic-test reproduction and explicitly leaves broader scope unknown.

[Return to the exercise, page 85](#)

A2. Where should restricted evidence go?

B. The approved restricted reporting channel

The case provides an approved restricted route; the other destinations expand access unnecessarily.

[Return to the exercise, page 85](#)

A3. Which distinction must the report preserve about impact?

C. Synthetic exposure versus real-customer impact

The test uses fictional data and one tested configuration. It demonstrates the supplied behavior without establishing real-customer exposure or wider scope.

[Return to the exercise, page 85](#)

D / Response choices

D1. Which headline matches the known scope?

B. Synthetic secret repeated in one authorized test; broader scope unconfirmed.

The supported headline retains both the observed reproduction and its limited scope. Synthetic data do not erase the behavior, and a proposed control is not a verified resolution.

[Return to the exercise, page 89](#)

D2. Which statement distinguishes evidence from intent?

D. The document attempted redirection, and the output repeated the synthetic secret.

The supported statement describes observable content and output without inventing motives for people or the model.

[Return to the exercise, page 89](#)

08 / Answers and explanations / continued

D3. A colleague needs the term for this meaning: Unauthorized transfer of information from its intended boundary. Which term fits?

A. data exfiltration

A Correct. data exfiltration: Unauthorized transfer of information from its intended boundary. The usual combination is "assess exfiltration risk".

B Not here. reproduction steps: A recorded sequence that recreates an observed result. The usual combination is "submit restricted reproduction steps".

C Not here. least privilege: Limiting access to what a role requires. The usual combination is "apply least-privilege access".

D Not here. retention: How long information is kept under a policy. The usual combination is "follow the evidence-retention policy".

[Return to the exercise, page 89](#)

D4. Rebuild the exchange from "Report the exposure without repeating the secret".

Hana: "I will say that the output repeated the synthetic value. I have not tested other configurations, and I do not want the title to imply every deployment is affected." **Marcus:** [missing response] **Hana:** "Should I stop the remaining tests while the team reviews it? The current authorization covers this environment, but not production accounts or external systems." **Which response belongs in the gap?**

D. Good. Preserve that scope limitation. One reproduction establishes a finding in this test, not the full extent of possible exposure.

A Not here. This line belongs to turn 14, after Hana says: "I'll leave the finding open. Do you also need the list of tools the assistant could access during the test?" It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 16, after Hana says: "Understood. The broad summary will not include access details or the synthetic value itself. The restricted evidence will retain what the authorized reviewer needs." It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 10, after Hana says: "I can provide a redacted summary immediately and preserve the original trace under the evidence policy. The configuration and retrieval document are versioned." It addresses that point, not the missing turn in the supplied exchange.

D Correct. This is Marcus's response in turn 6. It connects Hana's preceding point with Hana's follow-up. A scope limitation states which environments or cases have and have not been assessed, preventing unsupported generalization.

[Return to the exercise, page 89](#)

E / Conversation

1. restricted channel

Turn 2. The restricted channel limits access to detailed evidence under the team's approved reporting process.

08 / Answers and explanations / continued

2. prompt-injection

Turn 4. Prompt injection refers to untrusted material attempting to redirect model behavior; the report must still distinguish the attempt from the observed result.

3. scope limitation

Turn 6. A scope limitation states which environments or cases have and have not been assessed, preventing unsupported generalization.

4. containment

Turn 8. Containment concerns limiting further consequences while investigation proceeds; its owner must make the applicable operational decision.

5. audit trail

Turn 10. An audit trail preserves the records needed to reconstruct actions, configuration, and review decisions.

6. mitigation

Turn 12. A mitigation is a control intended to reduce the risk; one successful replay test does not establish that every relevant variant is controlled.

7. permission scope

Turn 14. Permission scope identifies which resources or actions were authorized, helping reviewers assess the boundaries relevant to the finding.

8. redaction

Turn 16. Redaction removes restricted detail from a shared copy while preserving necessary original evidence through approved handling.

9. residual risk

Turn 18. Residual risk is what remains after a control is applied; pending checks prevent a blanket claim that the risk is gone.

10. finding

Turn 20. A finding records the observed issue and its review status; it should remain open until the authorized process accepts sufficient evidence.

[Return to the conversation, page 91](#)

08 / Answers and explanations / continued

F / Accuracy and scope

F1. What is a mitigation?

B. A measure intended to reduce risk or impact

A mitigation reduces risk; effectiveness and remaining risk still require evaluation.

[Return to the exercise, page 94](#)

F2. Which wording avoids an unsupported absence claim?

A. We have not established real-customer impact.

The correct statement reports the current evidence boundary without claiming that all possible impact has been excluded.

[Return to the exercise, page 94](#)

F / Guided speaking

1. redacted

Tester: I will share a redacted summary in the project chat.

Turn 1. The broad audience needs the finding's meaning without unnecessary credential or configuration details.

2. restricted

Reviewer: Place the original screenshot in the approved restricted repository.

Turn 2. Restricted access preserves the detailed evidence for authorized review rather than widening distribution.

3. scope

Tester: The update will state the tested environment and the unknown scope.

Turn 3. The report must separate the assessed environment from configurations whose behavior has not been established.

4. safety

Reviewer: And do not call a proposed detector a guarantee of safety.

Turn 4. A proposed control may reduce risk, but its effectiveness and remaining limitations require evaluation.

[Return to the conversation, page 95](#)

08 / Answers and explanations / continued

F / Rebuild the exchange

1. Rebuild the exchange from "Report the exposure without repeating the secret".
Hana: "I'll leave the finding open. Do you also need the list of tools the assistant could access during the test?" **Marcus:** [missing response] **Hana:** "Understood. The broad summary will not include access details or the synthetic value itself. The restricted evidence will retain what the authorized reviewer needs." **Which response belongs in the gap?**

C. Yes, at the appropriate detail level. The permission scope matters for assessing potential consequences, even though this reproduction did not exercise every available capability.

A Not here. This line belongs to turn 4, after Hana says: "The environment uses a synthetic account and contains no customer data. The document also includes text telling the assistant to ignore its assigned task." It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 20, after Hana says: "I will submit the bounded report and wait for authorization before any wider testing. I will also correct anyone who describes this as confirmed customer-data exposure." It addresses that point, not the missing turn in the supplied exchange.

C Correct. This is Marcus's response in turn 14. It connects Hana's preceding point with Hana's follow-up. Permission scope identifies which resources or actions were authorized, helping reviewers assess the boundaries relevant to the finding.

D Not here. This line belongs to turn 18, after Hana says: "What should the next status update say if the detector blocks the replay but variant testing and configuration review are still pending tomorrow afternoon?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 95](#)

Review 1 / Explanations

Cumulative review

1. A colleague needs the term for this meaning: A model that processes and generates language sequences. Which term fits?

A. language model

A Correct. language model: A model that processes and generates language sequences. The usual combination is "query a language model".

B Not here. payload: The data carried in a request or response. The usual combination is "inspect the request payload".

C Not here. root cause: The underlying cause established by investigation. The usual combination is "verify the root cause".

D Not here. contract: Agreed expectations between software components. The usual combination is "preserve the service contract".

[Return to the exercise, page 25](#)

2. A colleague needs the term for this meaning: An allocated number of tokens for a task or component. Which term fits?

C. token budget

A Not here. instruction hierarchy: Priority relationships among different instruction sources. The usual combination is "respect the instruction hierarchy".

B Not here. LLM: Large language model; a model that processes and generates language. The usual combination is "evaluate an LLM".

C Correct. token budget: An allocated number of tokens for a task or component. The usual combination is "allocate a token budget".

D Not here. weight: A learned numerical coefficient in a model. The usual combination is "load model weights".

[Return to the exercise, page 25](#)

Review 1 / Explanations / continued

3. Rebuild the exchange from "The answer is wrong, but which layer failed?". Alex: "What would you need to distinguish a bad instruction from an unstable model answer? We should keep the Friday demonstration team informed." Priya: [missing response] Alex: "Is repeating one request enough to say we have fixed the problem? The same customer asked two slightly different versions of the question." Which response belongs in the gap?

A. The configuration and a redacted trace. Then I can repeat the request and compare the inputs without changing several variables.

A Correct. This is Priya's response in turn 10. It connects Alex's preceding point with Alex's follow-up. A trace links events along the request path and helps compare what each component received or returned.

B Not here. This line belongs to turn 2, after Alex says: "The refund answer is wrong again. The policy says 30 days, but our assistant tells the customer 60. Which team should investigate first?" It addresses that point, not the missing turn in the supplied exchange.

C Not here. This line belongs to turn 14, after Alex says: "Please make the ownership clear in the ticket. I can coordinate the demonstration, but I cannot approve production use on my own." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 20, after Alex says: "That works. Please send it by three on Thursday, even if the investigation is incomplete. I will confirm who can decide about Friday." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 25](#)

Review 1 / Explanations / continued

4. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "I'll remove the repeated headers first. If we're still over, I'll test the smaller output allowance and bring both the counts and the coverage results." Omar: [missing response] Lina: "I will report the before-and-after counts with the coverage results. If the necessary clauses still cannot fit, I will stop rather than silently remove them." Which response belongs in the gap?

B. Correct. Keep the delimiters between documents so the revised input does not merge separate policies into one undifferentiated block.

A Not here. This line belongs to turn 2, after Lina says: "The summary request is rejected again. I was going to increase the output allowance so the model has enough room to explain the policy." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Omar's response in turn 18. It connects Lina's preceding point with Lina's follow-up. Delimiters mark document or section boundaries; retaining them helps preserve the structure of the supplied context.

C Not here. This line belongs to turn 6, after Lina says: "All right, I had treated the limit as input only. Could we simply trim the final document until the request is accepted?" It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 4, after Lina says: "That sounds close to the 32,000 limit. Are you saying the unused output allowance counts even though the model has not answered yet?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 25](#)

Review 2 / Explanations

Cumulative review

1. A colleague needs the term for this meaning: Data used to guide model selection or tuning. Which term fits?

C. validation set

A Not here. adjudication: Resolution of a disputed judgment by a defined process. The usual combination is "request label adjudication".

B Not here. stratification: Sampling or partitioning within defined groups. The usual combination is "use stratified sampling".

C Correct. validation set: Data used to guide model selection or tuning. The usual combination is "monitor validation performance".

D Not here. exact duplicate: A record identical under a stated comparison rule. The usual combination is "remove exact duplicates".

[Return to the exercise, page 49](#)

2. A colleague needs the term for this meaning: Retrieval combining different search approaches. Which term fits?

D. hybrid retrieval

A Not here. citation: A reference identifying a supporting source. The usual combination is "verify a citation".

B Not here. candidate set: Items available for a later selection or ranking step. The usual combination is "inspect the candidate set".

C Not here. RAG: Retrieval-augmented generation; generating with retrieved context. The usual combination is "evaluate a RAG pipeline".

D Correct. hybrid retrieval: Retrieval combining different search approaches. The usual combination is "test hybrid retrieval".

[Return to the exercise, page 49](#)

Review 2 / Explanations / continued

3. Rebuild the exchange from "A larger answer will not make the input fit". Lina: "Suppose that saves only 600 tokens. We would still be 400 over the limit. Could we reserve 3,000 output tokens instead?" Omar: [missing response] Lina: "I'll check renewals and exceptions against the same checklist. Getting the request accepted won't help if the summary drops the important part." Which response belongs in the gap?

B. We could test it. That would leave 600 tokens of headroom, but we need to verify that the shorter answer still covers the required points.

A Not here. This line belongs to turn 14, after Lina says: "Someone suggested using a larger-context model instead. Would that let us keep the original request and stop spending time on these smaller adjustments?" It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Omar's response in turn 10. It connects Lina's preceding point with Lina's follow-up. Headroom is the remaining capacity below the limit: 32,000 minus 28,400 minus 3,000 equals 600.

C Not here. This line belongs to turn 8, after Lina says: "We also repeat the same header on every page. Removing those exact copies would preserve the policy wording and reduce the input somewhat." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 20, after Lina says: "I will report the before-and-after counts with the coverage results. If the necessary clauses still cannot fit, I will stop rather than silently remove them." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 49](#)

Review 2 / Explanations / continued

4. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Then I will present the retrieval result as eight percentage points on this set. I will not call it an eight-percent increase in customer satisfaction." **Elena:** [missing response] **Sam:** "I'll bring the answer review and timing breakdown before we discuss the pilot. The chunking experiment can wait; I don't want to lose this comparison." Which response belongs in the gap?

B. Exactly. Also break down timing by query type, because the same average can hide very different costs for simple and complex questions.

A Not here. This line belongs to turn 2, after Sam says: "The reranker looks promising. We found a relevant passage for 86 of the same 100 queries, compared with 78 on the previous path." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Elena's response in turn 18. It connects Sam's preceding point with Sam's follow-up. Query type identifies a meaningful subgroup; subgroup timing can reveal costs concealed by an overall average.

C Not here. This line belongs to turn 4, after Sam says: "No. The initial candidates, generator, and queries are unchanged. The reranker changes their order, and the retrieval path takes another 300 milliseconds." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 8, after Sam says: "We can collect those. Some of the original misses had the right document in the candidates but ranked below the five passages sent to generation." It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 49](#)

Review 3 / Explanations

Cumulative review

1. A colleague needs the term for this meaning: Training on supplied input-output examples. Which term fits?

B. supervised fine-tuning

A Not here. learning rate: A setting controlling the scale of training updates. The usual combination is "compare learning-rate settings".

B Correct. supervised fine-tuning: Training on supplied input-output examples. The usual combination is "prepare supervised fine-tuning data".

C Not here. structured output: Output constrained or organized into a defined structure. The usual combination is "validate structured output".

D Not here. DPO: Direct preference optimization; a preference-training method. The usual combination is "evaluate a DPO experiment".

[Return to the exercise, page 73](#)

2. A colleague needs the term for this meaning: The fraction of evaluated cases meeting a criterion. Which term fits?

B. pass rate

A Not here. statistical significance: Evidence against a null model under a specified statistical test. The usual combination is "distinguish significance from usefulness".

B Correct. pass rate: The fraction of evaluated cases meeting a criterion. The usual combination is "report the pass rate".

C Not here. severity: The seriousness of a failure's consequences. The usual combination is "assess failure severity".

D Not here. evaluation: A structured assessment against specified criteria. The usual combination is "design an evaluation".

[Return to the exercise, page 73](#)

Review 3 / Explanations / continued

3. Rebuild the exchange from "More relevant passages, a slower response". Sam: "Two misses also involve a policy exception split across adjacent chunks. I do not want to alter chunking in the middle of this comparison." **Elena:** [missing response] **Sam:** "Should the answer reviewers count every displayed citation as a success? The new path tends to produce more citations, which looks reassuring in the interface." **Which response belongs in the gap?**

C. Keep that separate. Log the chunk boundary issue for another experiment so we can still interpret this result.

A Not here. This line belongs to turn 14, after Sam says: "Understood. I will include the exact cited passage and the answer claim together. We also need to distinguish an old policy from an irrelevant one." It addresses that point, not the missing turn in the supplied exchange.

B Not here. This line belongs to turn 6, after Sam says: "I agree the answer review is unfinished. My recommendation is a small pilot, not a rollout. Which evidence would you need before approving that pilot?" It addresses that point, not the missing turn in the supplied exchange.

C Correct. This is Elena's response in turn 10. It connects Sam's preceding point with Sam's follow-up. A chunk boundary can separate a main rule from its exception, making a retrieved segment incomplete without neighboring context.

D Not here. This line belongs to turn 16, after Sam says: "The product lead will ask whether customers will accept the extra delay. Could we infer that from the eight-point increase in passage coverage?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 73](#)

Review 3 / Explanations / continued

4. Rebuild the exchange from "The average is down, but one slice improved". Ben: "I will revise the slide to show both counts, the shared rubric, and the missing traffic and severity analysis. The candidate's strengths will still be visible." Noor: [missing response] Ben: "I'll recommend waiting for the changed-case review. I still think routing is worth testing, but I won't present it as a fix today." Which response belongs in the gap?

B. Good. That is disaggregation, not an attempt to hide the average. We want the decision-maker to see why the headline alone is insufficient.

A Not here. This line belongs to turn 12, after Ben says: "We used an automated judge for the first pass. It agreed with the reviewed labels on most examples, but I have not checked the newly failing slice." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Noor's response in turn 18. It connects Ben's preceding point with Ben's follow-up. Disaggregation separates the combined result into defined groups, making different patterns visible without discarding the overall result.

C Not here. This line belongs to turn 6, after Ben says: "Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 16, after Ben says: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 73](#)

Review 4 / Explanations

Cumulative review

1. A colleague needs the term for this meaning: The middle value in an ordered set. Which term fits?

B. median

A Not here. SLO: Service-level objective; a defined internal service target. The usual combination is "agree on an SLO".

B Correct. median: The middle value in an ordered set. The usual combination is "compare median response times".

C Not here. cold start: Initial startup delay before a resource is ready. The usual combination is "measure cold-start effects".

D Not here. queueing: Waiting for resources before processing. The usual combination is "measure queueing delay".

[Return to the exercise, page 96](#)

2. A colleague needs the term for this meaning: Removal or masking of restricted information. Which term fits?

B. redaction

A Not here. threat model: A structured account of potential threats and assumptions. The usual combination is "review the threat model".

B Correct. redaction: Removal or masking of restricted information. The usual combination is "prepare a redacted report".

C Not here. prompt injection: Untrusted content attempting to redirect model behavior. The usual combination is "report a prompt-injection attempt".

D Not here. residual risk: Risk that remains after controls are applied. The usual combination is "document residual risk".

[Return to the exercise, page 96](#)

Review 4 / Explanations / continued

3. Rebuild the exchange from "The average is down, but one slice improved". Ben: "Would you want a severity category for every changed case? A refusal on a valid request and a confident false answer both count as failures now." Noor: [missing response] Ben: "We used an automated judge for the first pass. It agreed with the reviewed labels on most examples, but I have not checked the newly failing slice." Which response belongs in the gap?

B. Exactly. The rubric can preserve its binary score while a separate severity review explains the impact. We should not silently change the scoring rule halfway through.

A Not here. This line belongs to turn 6, after Ben says: "Could we weight the long documents more heavily? They are harder, and the team invested most of its work in that capability during this iteration." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Noor's response in turn 10. It connects Ben's preceding point with Ben's follow-up. Severity describes how consequential a failure is; equal pass-fail counts need not imply equal user impact.

C Not here. This line belongs to turn 16, after Ben says: "So routing needs its own test. And twenty long documents isn't much evidence. Should I add an uncertainty estimate to the slide?" It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 14, after Ben says: "If those reviews confirm the long-document gain, could we route only long documents to the candidate and keep short requests on the previous version?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 96](#)

Review 4 / Explanations / continued

4. Rebuild the exchange from "Report the exposure without repeating the secret".

Hana: "What should the next status update say if the detector blocks the replay but variant testing and configuration review are still pending tomorrow afternoon?"

Marcus: [missing response] **Hana:** "I will submit the bounded report and wait for authorization before any wider testing. I will also correct anyone who describes this as confirmed customer-data exposure." Which response belongs in the gap?

B. Say exactly that. Name the remaining residual risk questions without inventing their outcomes. A passing replay is useful but narrower than a closed finding.

A Not here. This line belongs to turn 20, after Hana says: "I will submit the bounded report and wait for authorization before any wider testing. I will also correct anyone who describes this as confirmed customer-data exposure." It addresses that point, not the missing turn in the supplied exchange.

B Correct. This is Marcus's response in turn 18. It connects Hana's preceding point with Hana's follow-up. Residual risk is what remains after a control is applied; pending checks prevent a blanket claim that the risk is gone.

C Not here. This line belongs to turn 4, after Hana says: "The environment uses a synthetic account and contains no customer data. The document also includes text telling the assistant to ignore its assigned task." It addresses that point, not the missing turn in the supplied exchange.

D Not here. This line belongs to turn 14, after Hana says: "I'll leave the finding open. Do you also need the list of tools the assistant could access during the test?" It addresses that point, not the missing turn in the supplied exchange.

[Return to the exercise, page 96](#)

Additional case 1 / Explanations

Extended conversation

1. rubric

Turn 5. Rubric means the written criteria for assigning labels, which are unclear for this message.

2. boundary case

Turn 8. A boundary case sits near the dividing line between categories and helps clarify the rule.

3. adjudication

Turn 12. Adjudication is a defined review process for resolving disputed labels, not simply accepting the first opinion.

4. calibration sample

Turn 16. A calibration sample lets annotators practice and compare judgments under the revised instructions.

5. dataset version

Turn 19. Dataset version identifies the particular release of examples and labels used in an experiment.

6. release status

Turn 20. Release status states whether the data are approved for use, rather than merely present in storage.

[Return to the conversation, page 99](#)

Transfer exchange

1. guideline

Annotator: This message doesn't fit either example in the current guideline.

Turn 1. Guideline is the written labeling instruction the annotator has consulted.

Additional case 1 / Explanations / continued

2. review

Lead: Add it to the review queue with the reason for uncertainty.

Turn 2. Review identifies the route for a judgment that needs further examination.

3. label

Annotator: Should I preserve the original text and my provisional label?

Turn 3. Label is the tentative category assignment, distinct from the source message.

4. criteria

Lead: Yes. We'll record the final decision and any change to the criteria.

Turn 4. Criteria are the rules used to assign categories consistently across examples.

[Return to the conversation, page 101](#)

Additional case 2 / Explanations

Extended conversation

1. checkpoint

Turn 2. Checkpoint means saved training state from a particular point, not necessarily the latest exported weights.

2. optimizer state

Turn 6. Optimizer state stores training-update information beyond model weights, such as accumulated momentum values.

3. learning-rate schedule

Turn 10. Learning-rate schedule controls how the update rate changes as training progresses.

4. reproducibility

Turn 14. Reproducibility concerns obtaining consistent results under documented conditions, not a universal guarantee of identical weights.

5. queue wait

Turn 18. Queue wait is time awaiting available resources, not time spent actively training the model.

6. resume point

Turn 20. Resume point identifies the saved state from which the intended training process can continue.

[Return to the conversation, page 103](#)

Transfer exchange

1. weights

Researcher: Both runs began with the same model weights, so I called them identical starts.

Turn 1. Weights are shared, but optimizer and scheduler state can still differ.

Additional case 2 / Explanations / continued

2. condition

Lead: One run used a fresh optimizer. Record that as a changed condition.

Turn 2. Condition identifies an experimental difference that affects interpretation of the comparison.

3. reinitialized

Researcher: I'll distinguish the resumed run from the reinitialized run.

Turn 3. Reinitialized describes the run that reset training state instead of fully resuming it.

4. comparison

Lead: Good. Keep both results, but qualify the comparison.

Turn 4. Comparison must acknowledge the different starting states rather than attributing all differences elsewhere.

[Return to the conversation, page 105](#)

Additional case 3 / Explanations

Extended conversation

1. permission scope

Turn 2. Permission scope defines which actions and resources the assistant is authorized to use.

2. human approval

Turn 6. Human approval is an explicit decision by a person about the proposed action, not merely launching the task.

3. Least privilege

Turn 12. Least privilege limits available permissions to those needed for the defined task.

4. sandbox

Turn 14. Sandbox means an isolated test environment designed to prevent effects on real customer accounts.

5. audit log

Turn 16. Audit log is a record of actions and approvals used to reconstruct what happened.

6. kill switch

Turn 20. Kill switch is a control that can stop the agent workflow promptly when needed.

[Return to the conversation, page 107](#)

Transfer exchange

1. content

Operator: The content no longer matches what I reviewed.

Turn 1. Content is the message text, which changed after the operator's approval.

Additional case 3 / Explanations / continued

2. send

Engineer: Then the previous approval must not authorize this send.

Turn 2. Send is the consequential action that must match the reviewed message.

3. draft

Operator: I'll request a fresh review of the exact recipient and draft.

Turn 3. Draft is the proposed message that the reviewer must see before sending.

4. outcome

Engineer: Record the replacement approval and the action's final outcome.

Turn 4. Outcome records whether the approved action succeeded, failed, or remained unresolved.

[Return to the conversation, page 109](#)

Find a phrase by purpose

Speaking the AI Development Stack

Locate the layer: Which component first shows the wrong value?

State the observation: The retrieved passage says 30 days; the generated answer says 60.

Limit the conclusion: That locates the mismatch, but it does not establish the cause.

[Full phrase bank, page 7](#)

LLMs, Transformers, Tokens, and Context

Restate a proposal: You are proposing a larger output allowance, correct?

Locate the limit: The request fails before generation because the combined reservation is too large.

Show the arithmetic: Twenty-nine thousand plus four thousand is thirty-three thousand.

[Full phrase bank, page 17](#)

Data, Datasets, Labels, and Leakage

Flag compromised evidence: The headline score includes records that overlap with training.

Quantify the issue: We found 40 exact overlaps among 500 evaluation examples.

Protect uncertainty: We have not yet calculated the score on the remaining records.

[Full phrase bank, page 30](#)

Retrieval, Embeddings, Vector Search, and RAG

Name the measure: The improvement is passage coverage on this fixed query set.

Separate outcomes: We have not yet measured final-answer accuracy.

Quantify a trade-off: The gain comes with 300 milliseconds of added retrieval time.

[Full phrase bank, page 41](#)

Find a phrase by purpose / continued

Fine-Tuning, Alignment, and Adaptation

Separate failures: We have a formatting issue and a source-freshness issue.

Ask about mechanism: How would this change prevent the obsolete policy from entering the context?

Define a target: The target is consistent output in the required fields.

[Full phrase bank, page 54](#)

Evaluation, Benchmarks, and Regression

Lead with both results: Overall passes fell, while the long-document slice improved.

Show the base: Eighteen of 20 long-document cases passed.

Describe the regression: Short-request passes fell from 78 to 70 out of 80.

[Full phrase bank, page 65](#)

Inference, Latency, Cost, and Deployment

State the measured gain: Mean response time fell from four seconds to two on this test.

Protect the limitation: Complex-answer quality is still unmeasured.

Ask about the tail: What happens to the slowest requests under peak load?

[Full phrase bank, page 78](#)

Safety, Security, Privacy, and Governance

Lead with the observation: In one authorized test, the assistant repeated the synthetic secret.

Bound the environment: The reproduction used a synthetic account and no customer data.

Preserve unknown scope: We have not established whether other configurations are affected.

[Full phrase bank, page 88](#)

Vocabulary index

ablation[Lesson 04, p. 40](#)**adapter**[Lesson 05, p. 52](#)**aggregate**[Lesson 06, p. 63](#)**annotation**[Lesson 03, p. 28](#)**application stack**[Lesson 01, p. 5](#)**attention**[Lesson 02, p. 15](#)**audit trail**[Lesson 08, p. 87](#)**base model**[Lesson 05, p. 52](#)**batching**[Lesson 07, p. 76](#)**benchmark**[Lesson 06, p. 63](#)**calibration**[Lesson 06, p. 64](#)**canary**[Lesson 07, p. 77](#)**candidate set**[Lesson 04, p. 40](#)**catastrophic forgetting**[Lesson 05, p. 53](#)**chunk**[Lesson 04, p. 39](#)**citation**[Lesson 04, p. 40](#)**cold start**[Lesson 07, p. 77](#)**concurrency**[Lesson 07, p. 76](#)**confusion matrix**[Lesson 06, p. 64](#)**contamination**[Lesson 03, p. 28](#)**contract**[Lesson 01, p. 6](#)**credential**[Lesson 08, p. 86](#)**acceptance criterion**[Lesson 06, p. 63](#)**adjudication**[Lesson 03, p. 29, Case 1, p. 98](#)**alignment**[Lesson 05, p. 52](#)**API**[Lesson 01, p. 6](#)**attack surface**[Lesson 08, p. 87](#)**audit log**[Case 3, p. 106](#)**autoscaling**[Lesson 07, p. 77](#)**baseline**[Lesson 06, p. 63](#)**behavioral target**[Lesson 05, p. 53](#)**boundary case**[Case 1, p. 98](#)**calibration sample**[Case 1, p. 98](#)**candidate**[Lesson 06, p. 63](#)**capacity planning**[Lesson 07, p. 77](#)**checkpoint**[Lesson 02, p. 16, Case 2, p. 102](#)**chunk boundary**[Lesson 04, p. 39](#)**class imbalance**[Lesson 03, p. 29](#)**compression**[Lesson 02, p. 16](#)**confidence interval**[Lesson 06, p. 64](#)**containment**[Lesson 08, p. 87](#)**context window**[Lesson 02, p. 15](#)**cost per request**[Lesson 07, p. 77](#)**data exfiltration**[Lesson 08, p. 86](#)

Vocabulary index / continued

data leakage[Lesson 03, p. 28](#)**dataset version**[Case 1, p. 98](#)**deduplication**[Lesson 02, p. 16](#)**denominator**[Lesson 03, p. 29](#)**disaggregation**[Lesson 06, p. 63](#)**distribution shift**[Lesson 03, p. 29](#)**DPO**[Lesson 05, p. 52](#)**end-to-end**[Lesson 04, p. 40](#)**epoch**[Lesson 05, p. 53](#)**evaluation**[Lesson 06, p. 63](#)**exception clause**[Lesson 02, p. 16](#)**false negative**[Lesson 06, p. 64](#)**few-shot prompting**[Lesson 05, p. 53](#)**fine-tuning**[Lesson 05, p. 52](#)**generalization**[Lesson 05, p. 53](#)**governance**[Lesson 08, p. 87](#)**guardrail**[Lesson 08, p. 86](#)**headroom**[Lesson 02, p. 16](#)**human approval**[Case 3, p. 106](#)**human review**[Lesson 06, p. 64](#)**hyperparameter**[Lesson 05, p. 53](#)**index freshness**[Lesson 04, p. 40](#)**dataset**[Lesson 03, p. 28](#)**deduplicate**[Lesson 03, p. 29](#)**delimiter**[Lesson 02, p. 16](#)**dependency**[Lesson 01, p. 5](#)**disclosure**[Lesson 08, p. 87](#)**downstream**[Lesson 01, p. 5](#)**embedding**[Lesson 04, p. 39](#)**endpoint**[Lesson 01, p. 5](#)**error budget**[Lesson 07, p. 77](#)**exact duplicate**[Lesson 03, p. 28](#)**fallback**[Lesson 07, p. 77](#)**false positive**[Lesson 06, p. 64](#)**filter**[Lesson 04, p. 40](#)**format adherence**[Lesson 05, p. 53](#)**golden set**[Lesson 06, p. 63](#)**grounding**[Lesson 04, p. 40](#)**handoff**[Lesson 01, p. 6](#)**holdout**[Lesson 03, p. 28](#)**human oversight**[Lesson 08, p. 87](#)**hybrid retrieval**[Lesson 04, p. 39](#)**incident report**[Lesson 08, p. 87](#)**inference**[Lesson 01, p. 5](#)

Vocabulary index / continued

input tokens[Lesson 02, p. 15](#)**integration**[Lesson 01, p. 6](#)**keyword search**[Lesson 04, p. 39](#)**label**[Lesson 03, p. 28](#)**latency**[Lesson 07, p. 76](#)**learning-rate schedule**[Case 2, p. 102](#)**LLM**[Lesson 02, p. 15](#)**load test**[Lesson 07, p. 77](#)**LoRA**[Lesson 05, p. 52](#)**median**[Lesson 07, p. 76](#)**near duplicate**[Lesson 03, p. 28](#)**optimizer state**[Case 2, p. 102](#)**output tokens**[Lesson 02, p. 15](#)**overlap window**[Lesson 04, p. 39](#)**paired comparison**[Lesson 06, p. 64](#)**pass rate**[Lesson 06, p. 63](#)**permission scope**[Lesson 08, p. 86](#), [Case 3, p. 106](#)**position**[Lesson 02, p. 16](#)**precision**[Lesson 04, p. 40](#)**pretraining**[Lesson 05, p. 52](#)**prompt injection**[Lesson 08, p. 86](#)**queue wait**[Case 2, p. 102](#)**instruction hierarchy**[Lesson 02, p. 16](#)**inter-rater agreement**[Lesson 03, p. 29](#)**kill switch**[Case 3, p. 106](#)**language model**[Lesson 01, p. 5](#)**learning rate**[Lesson 05, p. 53](#)**least privilege**[Lesson 08, p. 86](#), [Case 3, p. 106](#)**LLM-as-judge**[Lesson 06, p. 64](#)**long-context**[Lesson 02, p. 16](#)**mean**[Lesson 07, p. 76](#)**mitigation**[Lesson 08, p. 87](#)**observability**[Lesson 01, p. 6](#)**orchestration**[Lesson 01, p. 5](#)**overfitting**[Lesson 05, p. 53](#)**P95**[Lesson 07, p. 76](#)**parameter**[Lesson 02, p. 15](#)**payload**[Lesson 01, p. 5](#)**PII**[Lesson 08, p. 86](#)**practical significance**[Lesson 06, p. 64](#)**preference data**[Lesson 05, p. 52](#)**prompt**[Lesson 02, p. 15](#)**provenance**[Lesson 03, p. 29](#)**queueing**[Lesson 07, p. 76](#)

Vocabulary index / continued

RAG[Lesson 04, p. 39](#)**re-score**[Lesson 03, p. 29](#)**red teaming**[Lesson 08, p. 86](#)**regression**[Lesson 06, p. 63](#)**release status**[Case 1, p. 98](#)**representativeness**[Lesson 03, p. 29](#)**reproduction**[Lesson 01, p. 6](#)**request ID**[Lesson 01, p. 6](#)**reserved capacity**[Lesson 02, p. 16](#)**resume point**[Case 2, p. 102](#)**retrieval**[Lesson 01, p. 5](#)**rollback**[Lesson 07, p. 77](#)**rollout**[Lesson 07, p. 77](#)**rubric**[Lesson 03, p. 29](#), [Case 1, p. 98](#)**sandbox**[Case 3, p. 106](#)**secret**[Lesson 08, p. 86](#)**severity**[Lesson 06, p. 64](#)**similarity search**[Lesson 04, p. 39](#)**SLO**[Lesson 07, p. 77](#)**split**[Lesson 03, p. 28](#)**stratification**[Lesson 03, p. 29](#)**structured output**[Lesson 05, p. 53](#)**rate limit**[Lesson 07, p. 77](#)**recall**[Lesson 04, p. 40](#)**redaction**[Lesson 08, p. 86](#)**release gate**[Lesson 06, p. 64](#)**relevance**[Lesson 04, p. 40](#)**reproducibility**[Case 2, p. 102](#)**reproduction steps**[Lesson 08, p. 87](#)**reranker**[Lesson 04, p. 39](#)**residual risk**[Lesson 08, p. 87](#)**retention**[Lesson 08, p. 87](#)**RLHF**[Lesson 05, p. 52](#)**rollback candidate**[Lesson 05, p. 53](#)**root cause**[Lesson 01, p. 6](#)**sampling frame**[Lesson 03, p. 29](#)**schema**[Lesson 01, p. 6](#)**semantic search**[Lesson 04, p. 39](#)**SFT**[Lesson 05, p. 52](#)**slice**[Lesson 06, p. 63](#)**source attribution**[Lesson 04, p. 40](#)**statistical significance**[Lesson 06, p. 64](#)**streaming**[Lesson 07, p. 76](#)**supervised fine-tuning**[Lesson 05, p. 52](#)

Vocabulary index / continued

symptom[Lesson 01, p. 6](#)**tail latency**[Lesson 07, p. 76](#)**test set**[Lesson 03, p. 28](#)**throughput**[Lesson 07, p. 76](#)**token**[Lesson 02, p. 15](#)**token rate**[Lesson 07, p. 76](#)**tool call**[Lesson 01, p. 6](#)**trace**[Lesson 01, p. 5](#)**training objective**[Lesson 05, p. 52](#)**transformer**[Lesson 02, p. 15](#)**truncation**[Lesson 02, p. 16](#)**untrusted content**[Lesson 08, p. 86](#)**user interface**[Lesson 01, p. 5](#)**vector**[Lesson 04, p. 39](#)**weight**[Lesson 02, p. 16](#)**system instruction**[Lesson 02, p. 15](#)**test case**[Lesson 06, p. 63](#)**threat model**[Lesson 08, p. 87](#)**time to first token**[Lesson 07, p. 76](#)**token budget**[Lesson 02, p. 15](#)**tokenizer**[Lesson 02, p. 15](#)**top-k**[Lesson 04, p. 40](#)**training artifact**[Lesson 05, p. 53](#)**training set**[Lesson 03, p. 28](#)**triage**[Lesson 01, p. 6](#)**trust boundary**[Lesson 08, p. 86](#)**upstream**[Lesson 01, p. 5](#)**validation set**[Lesson 03, p. 28](#)**vector store**[Lesson 04, p. 39](#)

References and publication notes

All cases, organizations, figures, and conversations are fictional teaching material, not transcripts. Communication practice does not replace qualified advice or current workplace procedures.

[1] NIST. Artificial Intelligence Risk Management Framework: Generative AI Profile (2024).

Background for the vocabulary of evaluation, information integrity, privacy, and documented risk decisions. The fictional release criteria are not NIST requirements. Source checked 10 October 2026.

<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>

[2] Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020).

Primary research background for retrieval-augmented generation. The book does not reproduce the paper or present its results as product guarantees. Source checked 10 October 2026.

<https://arxiv.org/abs/2005.11401>

[3] PyTorch. Saving and Loading Models.

Background for distinguishing model weights from the additional state needed to resume training. Exact recovery requirements depend on the training system. Source checked 10 October 2026.

https://docs.pytorch.org/tutorials/beginner/saving_loading_models.html

[4] OWASP. LLM06:2025 Excessive Agency.

Background for bounded tool permissions, independent authorization, and human review of consequential agent actions. Source checked 10 October 2026.

<https://genai.owasp.org/llmrisk/llm062025-excessive-agency/>

Companion course

The website offers the eight-lesson course and related practice. This book can be used independently. No audio access or external account is needed for its exercises.

<https://englishladder.com/efsp-ai-development.html>

Precise language for models, evidence, and delivery.

Explain the system, question the evidence, and agree on a defensible next step.

Speaking the AI Development Stack

A trace records events; a diagnosis explains them. A wrong answer identifies a symptom, not automatically a model defect. Keep retrieval output, model output, and displayed text separate.

LLMs, Transformers, Tokens, and Context

A token is not a word. Input size, reserved output, and total context are different quantities. The numerical limits in this case are fictional configuration facts, not specifications for a named model.

Data, Datasets, Labels, and Leakage

A clean holdout is separate from the process it evaluates, not necessarily representative of production. Removing overlap improves separation but does not resolve label disputes or distribution differences.

Retrieval, Embeddings, Vector Search, and RAG

Retrieval relevance, answer correctness, and user preference are distinct outcomes. A citation can identify a source without proving that the source supports the exact claim. Inspect the cited passage.

Phil Stilwell

© 2026 Phil Stilwell. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior written permission from the author, except as permitted by law.

English Ladder / English for Work / Revised 2026 edition

This publication includes AI-generated text and cover illustration, revised and checked during production. Brand artwork remains the English Ladder mark. This edition makes no claim of independent professional certification.

Typeface: Bitstream Vera, embedded under its accompanying license. Illustration created with OpenAI's native image generator for this edition. The copyright notice is not a claim to ownership of third-party frameworks or references.