

Boost up Your Certification Score

Huawei H13-325

**Huawei Certified ICT Professional-AI-Application
Developer**



For More Information – Visit link below:

<https://www.examsboost.com/>

Product Version

- ✓ Up to Date products, reliable and verified.
- ✓ Questions and Answers in PDF Format.

Visit us at: <https://www.examsboost.com/test/h13-325>

Latest Version: 6.0

Question: 1

A team is instrumenting a live assistant. No correct answer is available at serving time, so accuracy cannot be computed on live traffic, and the team's evaluation run happens only once a month. It wants signals that would show quality slipping in between those runs.

Which three measures serve that purpose?

(Choose three.)

- A. How often a user edits, regenerates or abandons the answer they were given.
- B. A human review of a small random sample of live answers each week, scored against the same criteria.
- C. The training loss recorded at the end of the model's most recent training run.
- D. The accelerator utilization reading on the serving hosts, taken across the fleet and read as a measure of answer quality.
- E. The proportion of answers that fail the application's own conformance and policy checks.

Answer: A,B,E

Question: 2

A developer is writing the internal design note for a text generation service. One section describes an arrangement in which a smaller, cheaper model runs first and puts forward several candidate tokens. The deployed model then examines those candidates in a single pass and keeps the ones it agrees with, discarding the rest. The note needs the standard name for the arrangement and the standard names for the two models.

Which technique does the section describe?

- A. Knowledge distillation, with the smaller model as the student and the larger deployed model as the teacher that supplies its training signal.
- B. Pipeline parallelism, with the smaller model forming the first stage and the deployed model forming the second stage of the same graph.
- C. Speculative decoding, with the smaller model as the draft model and the deployed model as the target model.
- D. Continuous batching, with the smaller model admitting arriving requests and the deployed model generating for whichever group is currently assembled.

Answer: C

Question: 3

A team promised that its document assistant would answer within four seconds. In testing, each of the three model calls the feature makes — classify, then extract, then draft — returned in about one and a half seconds, and the engineers signed the design off on that basis. In production the feature routinely takes more than four seconds, and no individual call has ever been recorded above two.

What does this pattern indicate about the design?

- A. The feature is queueing behind other traffic, so the extra time is waiting rather than work and raising serving capacity is what the measurements call for.
- B. The model runs more slowly in production than in testing, so the signed-off per-call figure was not representative of the live service.
- C. The three calls duplicate work between them, so removing the overlap is what brings the total inside the promise.
- D. The three calls run one after another, so the budget the feature must meet is their sum and not the time of any single call.

Answer: D

Question: 4

A developer demonstrates a prototype assistant to a department head. The prototype runs on the developer's workstation, serves one request at a time, and answered all twelve demonstration questions within three seconds each. The department head now wants it opened to four hundred staff next month, and asks what has to be established before that date.

Which activity most directly addresses the gap the demonstration leaves?

- A. Run the same pipeline against a realistic concurrent load and record the response time and the cost per request under it.
- B. Move the prototype to a server with more memory and more cores than the workstation, since the department's load will not fit in the current footprint.
- C. Increase the number of training epochs so the model handles the wider range of questions staff will ask.
- D. Extend the demonstration set to a hundred questions and check that the answers are still judged acceptable.

Answer: A

Question: 5

A team is about to measure whether a graph optimization helped. The baseline figure it holds was taken three weeks ago from live production traffic during a busy afternoon. The optimized build is ready, and the plan is to measure it this evening on the same cluster, take the difference between the two numbers, and report that difference as the effect of the change.

What is the most important correction to this plan?

- A. Run the optimized build for a longer period than the baseline, so its figure averages over more requests and is therefore the more dependable of the two.
- B. Drive both builds with the same request trace at the same concurrency, so the difference belongs to the change.
- C. Report the effect as a percentage of the baseline rather than as an absolute figure, so the comparison holds across different traffic levels.
- D. Repeat the baseline measurement on the optimized build's hardware, since the cluster's composition is what most often differs between two runs taken weeks apart.

Answer: B

Question: 6

A team must size a deployment before release. It has measured and agreed the following:

- one replica sustains 25 requests a minute at the response length this service produces;
- peak demand reaches 130 requests a minute;
- the service must still carry peak demand with one replica out of service.

How many replicas should be provisioned?

- A. Six, because six replicas at 25 requests a minute cover the measured peak of 130 with a little capacity left over for growth.
- B. Eight, because a replica taken out of service has to be covered by two others while peak demand is being carried.
- C. Five, because 130 requests a minute divided by 25 comes to just over five, and the remainder is absorbed by queueing.
- D. Seven, because six replicas are needed to carry 130 requests a minute and a seventh covers one being out of service.

Answer: D

Question: 7

A product manual has been split into chunks at section boundaries, and long sections further split by length. Reading a sample, a developer finds chunks whose text runs along the lines of: this setting must be enabled before the procedure is run. The setting and the procedure are named in the section heading and in the section's opening paragraph, neither of which is inside the chunk being read.

Which change to the preparation addresses what the developer found?

- A. Merge each chunk with the chunk before it, so that the naming in the opening paragraph travels with the text that refers to it.
- B. Carry the document title and the section heading into every chunk drawn from that section.
- C. Rewrite the chunks so that each sentence names its own subject, replacing references back to earlier text with the terms those references stand for.

D. Increase the overlap between neighboring chunks, so that the closing text of one section appears again at the start of the next.

Answer: B

Question: 8

A team is preparing customer support transcripts for use by an AI application. Individuals must not be identifiable from the prepared text, but the team still needs to be able to tell that the same customer appears in two separate transcripts. Each customer name is therefore replaced, everywhere it occurs, by the same invented reference.

Which practice does this describe?

- A. Pseudonymization, in which an identifier is replaced by a consistent surrogate value.
- B. Anonymization, in which identifying information is removed so that no link between records survives.
- C. Encryption, in which an identifier is transformed so that a holder of the key can recover it.
- D. Redaction, in which the identifying span is struck out of the text altogether.

Answer: A

Question: 9

A service uses continuous, in-flight batching. Monitoring shows that whenever a request with a very long prompt is admitted, the step that processes its prompt lasts many times longer than usual, and every sequence already generating waits through it, so users see their responses stall. The team proposes splitting each long prompt's processing into chunks, each sharing a step with the sequences that are generating.

Which two outcomes should the team expect from the change?

(Choose two.)

- A. The memory held for the long request's cached state falls, because each chunk's state is released once the next chunk has been processed.
- B. The total prompt work performed falls, because each chunk is processed with less of the prompt in view.
- C. The stalls seen by generating sequences shrink, because each step now carries a bounded slice of prompt work beside them.
- D. The interval between tokens for the long request rises once it is generating, because its prompt was processed as several pieces.
- E. The long-prompt request waits somewhat longer for its first token, because its prompt is now processed over several steps instead of one.

Answer: C,E

Question: 10

An agent completes a purchasing task over roughly twenty steps; run lengths vary, and some exceed the transcript budget. The team observes two things on longer runs: a spending limit the user gave in the opening message is respected early and breached in the later steps, and the agent re-fetches supplier details it has already looked up. The team is deciding what the agent should carry between steps. Which approach addresses both behaviors?

- A. Give the transcript a larger share of each request, keeping more of the earlier steps and their results in view as the run continues.
- B. Summarize the transcript at each step, leaving the agent a shorter account of everything that has happened so far in the run.
- C. Repeat the user's opening instruction after each tool result, placing it most recently in what the model reads.
- D. Hold the task's constraints and the results already obtained in a task state that is supplied on every step.

Answer: D

Thank You for Trying Our Product

For More Information – **Visit link below:**

<https://www.examsboost.com/>

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ **90 Days Free Updates**
- ✓ **Money Back Pass Guarantee**
- ✓ **Instant Download or Email Attachment**
- ✓ **24/7 Live Chat Support**
- ✓ **PDF file could be used at any Platform**
- ✓ **50,000 Happy Customer**



Visit us at: <https://www.examsboost.com/test/h13-325>