

PRACTICE QUESTIONS

IABAC

Certified NLP Expert

SHARED BY [u/Misty-Server-2025](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	IABAC Certified NLP Expert

Shared by [u/Misty-Server-2025](#) on aiexamcerts

10 practice questions for the IABAC Certified NLP Expert exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/iabac/certified-nlp-expert. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Why do subword tokenizers such as byte-pair encoding help with out-of-vocabulary words?
 - A Every unknown word is replaced by one special token that carries no information about its spelling or its parts
 - B The vocabulary is made as large as every word that appears on the internet, so no unknown items remain
 - C Rare words are broken into known smaller pieces, so almost any input can still be represented
 - D Words are removed from the input before training so that the model never encounters rare terms at all

Answer: C

Because the vocabulary contains characters and frequent fragments, any string can be segmented into known pieces. A single unknown token is the tempting distractor, and it is what word-level vocabularies do, which loses information. An internet-sized vocabulary is impractical, and deleting rare words discards data.

2. Stemming and lemmatization always produce the same output for every English word.
 - A True
 - B False

Answer: B (False)

Stemming applies suffix rules and can produce non-words such as 'studi', while lemmatization returns dictionary forms. Their outputs differ on many words, such as irregular verbs. The statement is false.

3. What is the main difference between GloVe and word2vec in how they learn?
 - A GloVe uses a character-level convolution on each word, while word2vec uses only whole-word identities
 - B GloVe requires labeled sentiment data, while word2vec is fully unsupervised on raw text from the web
 - C GloVe produces contextual vectors that change with each sentence, while word2vec gives one fixed vector
 - D GloVe factorizes a global word co-occurrence matrix, while word2vec trains on local context windows

Answer: D

GloVe fits word vectors to logarithms of global co-occurrence counts, whereas word2vec samples local windows and predicts. The contextual claim is the tempting distractor, but both produce one static vector per word, and contextual vectors come from models such as BERT. The other options invent properties that neither method has.

4. The F-measure equal to the harmonic mean of precision and recall is the F _____ score.

Answer: 1 (also accepted: one, F1)

F1 weights precision and recall equally, so it is low when either one is low. The F-beta family generalizes it, which is a tempting alternative. The blank is 1.

5. Which two techniques reduce the risk of an answer being unsupported by its sources in a RAG system? (Choose two.)

- A Removing the retriever so that the model relies only on its parameters for facts
- B Lengthening each answer so that more claims can be made per response
- C Requiring the answer to cite retrieved passages by identifier
- D Increasing the sampling temperature so that the answers vary more from run to run
- E Returning no answer when retrieval scores fall below a confidence threshold

Answer: C, E

Citations make claims checkable, and a retrieval threshold prevents answers when evidence is weak. Higher temperature is the tempting distractor, since it changes outputs, but it makes unsupported content more likely. Removing retrieval and lengthening answers both increase unsupported claims.

6. Why do static word embeddings struggle with polysemy?

- A Each word gets one vector regardless of sense, so 'bank' in finance and 'bank' by a river share one point
- B The embedding table cannot store more than one hundred words, so rare senses are always dropped entirely
- C Static vectors are computed at inference time from the output layer, which makes them slow to read back
- D Polysemous words are removed from the training corpus by the tokenizer before any model sees the text

Answer: A

A static embedding table gives each token type one vector, so the senses are averaged together. The table-size claim is a tempting distractor, but table size has nothing to do with senses. Output layers and tokenizer filtering are not the cause of the problem.

7. What does scaled dot-product attention compute from queries, keys and values?
- A A product of the query and value matrices, which is then normalized to unit length for each token position
 - B A pooled average of the key vectors across all positions, with no dependence on the query at all
 - C A softmax over query-key similarities, which is then used to take a weighted sum of the values
 - D A hard selection of the single key that matches the query exactly, with all other values discarded

Answer: C

Similarities between queries and keys become attention weights through a softmax, and the weights mix the value vectors. Multiplying queries by values is the tempting distractor, since matrix products are involved, but keys are compared with queries, not values. Averaging keys and hard selection are different schemes.

8. The standard form of self-attention has a computational cost that grows quadratically with sequence length.
- A True
 - B False

Answer: A (True)

Standard attention forms a score for every pair of positions, so the score matrix has size n by n . Efficient variants aim to reduce this cost, but the standard form is quadratic. The statement is true.

9. Which model family is built mainly from the transformer decoder and trained with next-token prediction?
- A GPT-style decoder-only language models, which generate text one token at a time from left to right
 - B BERT-style encoder-only models, which are trained with masked tokens and used mostly for classification
 - C Word2vec-style shallow networks, which predict a target word from neighboring context words in the text
 - D Hidden Markov models, which estimate transition and emission probabilities over a set of tags

Answer: A

GPT-style models stack causal decoder blocks and are trained to predict the next token. BERT is the tempting distractor, since it is also transformer-based, but its encoder is bidirectional and not trained for generation. The shallow and statistical models are older families.

10. What is the standard input format for a natural language inference (NLI) model?
- A** A premise and a hypothesis, which the model classifies as entailment, contradiction or neutral
 - B** A single document and a list of keywords, which the model ranks by relevance to the query
 - C** A pair of images and a caption, which the model scores for visual alignment in the batch
 - D** A table of numeric features and a target value, which the model regresses on in the usual way

Answer: A

NLI takes two sentences and predicts whether the first supports, contradicts or is unrelated to the second. Document ranking is the tempting distractor, since it also compares text with a query, but it has no entailment labels. Image-caption and tabular regression are different input formats.