NVIDIA NCA-GENL Practice Questions

Shared by Faulkner on 21-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Which Python library is specifically designed for working with large language models (LLMs)?

Options:

A- NumPy
B- Pandas
C- HuggingFace Transformers
D- Scikit-learn

Answer:

C

Explanation:

The HuggingFace Transformers library is specifically designed for working with large language models (LLMs), providing tools for model training, fine-tuning, and inference with transformer-based architectures (e.g., BERT, GPT, T5). NVIDIA's NeMo documentation often references HuggingFace Transformers for NLP tasks, as it supports integration with NVIDIA GPUs and frameworks like PyTorch for optimized performance. Option A (NumPy) is for numerical computations, not LLMs. Option B (Pandas) is for data manipulation, not model-specific tasks. Option D (Scikit-learn) is for traditional machine learning, not transformer-based LLMs.

NVIDIA NeMo Documentation:
https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

HuggingFace Transformers Documentation: https://huggingface.co/docs/transformers/index

Question 2

Question Type: MultipleChoice

Which of the following contributes to the ability of RAPIDS to accelerate data processing? (Pick the 2 correct responses)

Options:

A- Ensuring that CPUs are running at full clock speed.
B- Subsampling datasets to provide rapid but approximate answers.
C- Using the GPU for parallel processing of data.
D- Enabling data processing to scale to multiple GPUs.
E- Providing more memory for data analysis.

Answer:

C, D

Explanation:

RAPIDS is an open-source suite of GPU-accelerated data science libraries developed by NVIDIA to speed up data processing and machine learning workflows. According to NVIDIA's RAPIDS documentation, its key advantages include:

Option C: Using GPUs for parallel processing, which significantly accelerates computations for tasks like data manipulation and machine learning compared to CPU-based processing.

Option D: Scaling to multiple GPUs, allowing RAPIDS to handle large datasets efficiently by distributing workloads across GPU clusters.

Option A is incorrect, as RAPIDS focuses on GPU, not CPU, performance. Option B (subsampling) is not a primary feature of RAPIDS, which aims for exact results. Option E (more memory) is a hardware characteristic, not a RAPIDS feature.

NVIDIA RAPIDS Documentation: https://rapids.ai/

Question 3

Question Type: MultipleChoice

Which of the following is a key characteristic of Rapid Application Development (RAD)?

Options:

A- Iterative prototyping with active user involvement.
B- Extensive upfront planning before any development.
C- Linear progression through predefined project phases.
D- Minimal user feedback during the development process.

Answer:

A

Explanation:

Rapid Application Development (RAD) is a software development methodology that emphasizes iterative prototyping and active user involvement to accelerate development and ensure alignment with user needs. NVIDIA's documentation on AI application development, particularly in the context of NGC (NVIDIA GPU Cloud) and software workflows, aligns with RAD principles for quickly building and iterating on AI-driven applications. RAD involves creating prototypes, gathering user feedback, and refining the application iteratively, unlike traditional waterfall models. Option B is incorrect, as RAD minimizes upfront planning in favor of flexibility. Option C describes a linear waterfall approach, not RAD. Option D is false, as RAD relies heavily on user feedback.

NVIDIA NGC Documentation: https://docs.nvidia.com/ngc/ngc-overview/index.html

Question 4

Question Type: MultipleChoice

What are the main advantages of instructed large language models over traditional, small language models (< 300M parameters)? (Pick the 2 correct responses)

Options:

A- Trained without the need for labeled data.
B- Smaller latency, higher throughput.
C- It is easier to explain the predictions.
D- Cheaper computational costs during inference.
E- Single generic model can do more than one task.

Answer:

D, E

Explanation:

Instructed large language models (LLMs), such as those supported by NVIDIA's NeMo framework, have significant advantages over smaller, traditional models:

Option D: LLMs often have cheaper computational costs during inference for certain tasks because they can generalize across multiple tasks without requiring task-specific retraining, unlike smaller models that may need separate models per task.

Option E: A single generic LLM can perform multiple tasks (e.g., text generation, classification, translation) due to its broad pre-training, unlike smaller models that are typically task-specific.

Option A is incorrect, as LLMs require large amounts of data, often labeled or curated, for pre-training. Option B is false, as LLMs typically have higher latency and lower throughput due to their size. Option C is misleading, as LLMs are often less interpretable than smaller models.

NVIDIA NeMo Documentation:
https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

Brown, T., et al. (2020). 'Language Models are Few-Shot Learners.'

Question 5

Question Type: MultipleChoice

In large-language models, what is the purpose of the attention mechanism?

Options:

A- To measure the importance of the words in the output sequence.
B- To determine the order in which words are generated.
C- To capture the order of the words in the input sequence.
D- To assign weights to each word in the input sequence.

Answer:

D

Explanation:

The attention mechanism is a critical component of large language models, particularly in Transformer architectures, as covered in NVIDIA's Generative AI and LLMs course. Its primary purpose is to assign weights to each token in the input sequence based on its relevance to other tokens, allowing the model to focus on the most contextually important parts of the input when generating or interpreting text. This is achieved through mechanisms like self-attention, where each token computes a weighted sum of all other tokens' representations, with weights determined by their relevance (e.g., via scaled dot-product attention). This enables the model to

capture long-range dependencies and contextual relationships effectively, unlike traditional recurrent networks. Option A is incorrect because attention focuses on the input sequence, not the output sequence. Option B is wrong as the order of generation is determined by the model's autoregressive or decoding strategy, not the attention mechanism itself. Option C is also inaccurate, as capturing the order of words is the role of positional encoding, not attention. The course highlights: 'The attention mechanism enables models to weigh the importance of different tokens in the input sequence, improving performance in tasks like translation and text generation.'

Question 6

Question Type: MultipleChoice

What is the main consequence of the scaling law in deep learning for real-world applications?

Options:

A- With more data, it is possible to exceed the irreducible error region.
B- The best performing model can be established even in the small data region.
C- Small and medium error regions can approach the results of the big data region.
D- In the power-law region, with more data it is possible to achieve better results.

Answer:

D

Explanation:

The scaling law in deep learning, as covered in NVIDIA's Generative AI and LLMs course, describes the relationship between model performance, data size, model size, and computational resources. In the power-law region, increasing the amount of data, model parameters, or compute power leads to predictable improvements in performance, as errors decrease following a power-law trend. This has significant implications for real-world applications, as it suggests that scaling up data and resources can yield better results, particularly for large language models (LLMs). Option A is incorrect, as the irreducible error represents the inherent noise in the data, which cannot be exceeded regardless of data size. Option B is wrong, as small data regions typically yield suboptimal performance compared to scaled models. Option C is misleading, as small and medium data regimes do not typically match big data performance without scaling. The course highlights: 'In the power-law region of the scaling law, increasing data and compute resources leads to better model performance, driving advancements in real-world deep learning applications.'

Question 7

Question Type: MultipleChoice

Which calculation is most commonly used to measure the semantic closeness of two text passages?

Options:

A- Hamming distance
B- Jaccard similarity
C- Cosine similarity
D- Euclidean distance

Answer:

C

Explanation:

Cosine similarity is the most commonly used metric to measure the semantic closeness of two text passages in NLP. It calculates the cosine of the angle between two vectors (e.g., word embeddings or sentence embeddings) in a high-dimensional space, focusing on the direction rather than magnitude, which makes it robust for comparing semantic similarity. NVIDIA's documentation on NLP tasks, particularly in NeMo and embedding models, highlights cosine similarity as the standard metric for tasks like semantic search or text similarity, often using embeddings from models like BERT or Sentence-BERT. Option A (Hamming distance) is for binary data, not text embeddings. Option B (Jaccard similarity) is for set-based comparisons, not semantic content. Option D (Euclidean distance) is less common for text due to its sensitivity to vector magnitude.

NVIDIA NeMo Documentation:
https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

Question 8

Question Type: MultipleChoice

Which principle of Trustworthy AI primarily concerns the ethical implications of AI's impact on society and includes considerations for both potential misuse and unintended consequences?

Options:

A- Certification
B- Data Privacy
C- Accountability
D- Legal Responsibility

Answer:

C

Explanation:

Accountability is a core principle of Trustworthy AI that addresses the ethical implications of AI's societal impact, including potential misuse and unintended consequences. NVIDIA's guidelines on Trustworthy AI, as outlined in their AI ethics framework, emphasize accountability as ensuring that AI systems are transparent, responsible, and answerable for their outcomes. This includes mitigating risks of bias, ensuring fairness, and addressing unintended societal impacts. Option A (Certification) refers to compliance processes, not ethical implications. Option B (Data Privacy) focuses on protecting user data, not broader societal impact. Option D (Legal Responsibility) is related but narrower, focusing on liability rather than ethical considerations.

NVIDIA Trustworthy AI: https://www.nvidia.com/en-us/ai-data-science/trustworthy-ai/

Question 9

Question Type: MultipleChoice

Given preparing a multilingual dataset for fine-tuning an LLM, which preprocessing technique is most effective for handling text from diverse scripts (e.g., Latin, Cyrillic, Devanagari) to ensure consistent model performance?

Options:

A- Normalizing all text to a single script using transliteration.
B- Applying Unicode normalization to standardize character encodings.
C- Removing all non-Latin characters to simplify the input.

D- Converting text to phonetic representations for cross-lingual alignment.

Answer:

B

Explanation:

When preparing a multilingual dataset for fine-tuning an LLM, applying Unicode normalization (e.g., NFKC or NFC forms) is the most effective preprocessing technique to handle text from diverse scripts like Latin, Cyrillic, or Devanagari. Unicode normalization standardizes character encodings, ensuring that visually identical characters (e.g., precomposed vs. decomposed forms) are represented consistently, which improves model performance across languages. NVIDIA's NeMo documentation on multilingual NLP preprocessing recommends Unicode normalization to address encoding inconsistencies in diverse datasets. Option A (transliteration) may lose linguistic nuances. Option C (removing non-Latin characters) discards critical information. Option D (phonetic conversion) is impractical for text-based LLMs.

NVIDIA NeMo Documentation:
https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/intro.html

Question 10

Question Type: MultipleChoice

In the development of Trustworthy AI, what is the significance of 'Certification' as a principle?

Options:

A- It ensures that AI systems are transparent in their decision-making processes.
B- It requires AI systems to be developed with an ethical consideration for societal impacts.
C- It involves verifying that AI models are fit for their intended purpose according to regional or industry-specific standards.
D- It mandates that AI models comply with relevant laws and regulations specific to their deployment region and industry.

Answer:

C

Explanation:

In the development of Trustworthy AI, 'Certification' as a principle involves verifying that AI models are fit for their intended purpose according to regional or industry-specific standards, as discussed in NVIDIA's Generative AI and LLMs course. Certification ensures that models meet performance, safety, and ethical benchmarks, providing assurance to stakeholders about their reliability and appropriateness. Option A is incorrect, as transparency is a separate principle, not certification. Option B is wrong, as ethical considerations are broader and not specific to certification. Option D is inaccurate, as compliance with laws is related but distinct from certification's focus on fitness for purpose. The course states: ''Certification in Trustworthy AI verifies that models meet regional or industry-specific standards, ensuring they are fit for their intended purpose and reliable.''

Question 11

Question Type: MultipleChoice

Which metric is primarily used to evaluate the quality of the text generated by language models?

Options:

A- Perplexity
B- Precision
C- Recall
D- Accuracy

Answer:

A

Explanation:

Perplexity is the primary metric used to evaluate the quality of text generated by language models, as emphasized in NVIDIA's Generative AI and LLMs course. Perplexity measures how well a language model predicts a sequence of tokens, with lower values indicating better performance, as the model is less ''surprised'' by the data. It is calculated as the exponentiated average negative log-likelihood of the tokens in a test set, reflecting the model's ability to assign high probabilities to correct sequences. In generative tasks, perplexity is widely used because it directly assesses the model's fluency and coherence. Option B, Precision, and Option C, Recall, are metrics for classification tasks, not text generation. Option D, Accuracy, is also irrelevant for evaluating generative quality, as it applies to categorical predictions. The course notes:

''Perplexity is a key metric for evaluating language models, measuring how well the model predicts text sequences, with lower perplexity indicating higher-quality generation.''

Question 12

Question Type: MultipleChoice

In the evaluation of Natural Language Processing (NLP) systems, what do 'validity' and 'reliability' imply regarding the selection of evaluation metrics?

Options:

A- Validity involves the metric's ability to predict future trends in data, and reliability refers to its capacity to integrate with multiple data sources.
B- Validity ensures the metric accurately reflects the intended property to measure, while reliability ensures consistent results over repeated measurements.
C- Validity is concerned with the metric's computational cost, while reliability is about its applicability across different NLP platforms.
D- Validity refers to the speed of metric computation, whereas reliability pertains to the metric's performance in high-volume data processing.

Answer:

B

Explanation:

In evaluating NLP systems, as discussed in NVIDIA's Generative AI and LLMs course, validity and reliability are critical for selecting evaluation metrics. Validity ensures that a metric accurately measures the intended property (e.g., BLEU for translation quality or F1-score for classification performance), reflecting the system's true capability. Reliability ensures that the metric produces consistent results across repeated measurements under similar conditions, indicating stability and robustness. Together, these ensure trustworthy evaluations. Option A is incorrect, as validity is not about predicting trends, and reliability is not about data source integration. Option C is wrong, as validity and reliability are not primarily about computational cost or platform applicability. Option D is inaccurate, as validity and reliability do not focus on computation speed or high-volume processing. The course notes: ''Validity ensures NLP evaluation metrics accurately measure the intended property, while reliability ensures consistent results across repeated evaluations, critical for robust system assessment.''