Generative AI Masterclass:
Responsible AI

H2O.ai

Exam Questions:

What is a main conclusion of TrustLLM?

A. Trustworthiness and utility were positively correlated.
B. LLMs cannot be trusted
C. Utility is more important than trustworthiness
D. Trustworthiness is more important than utility

Answer: A

How can counterfactual analysis be applied in the context of LLMs?

A. Either a matched control group must be used or an RCT.
B. Measuring consistency across prompts and RAG applications.
C. Ex post analysis using a difference-in-differences estimator.
D. Measuring accuracy across prompts and RAG applications.

Answer: B

What are the four key components of the proposed Gen AI Evaluation Framework?

A. Models, Documents, Evaluators, Tests.
B. Encoders, Decoders, Transformers, Accuracy Metrics.
C. Models, Leaderboards, Documents, Tests.
D. RAGAS, Hallucination Index, Answer correctness, Context similarity.

Answer: A

What are influence functions?

A. A function to understand how important social media influences the training data
B. A function to influence how embeddings from a decoder are used in the encoder part of a Transformer-based architecture
C. A function to measure the influence that the context has an LLM's generated response
D. A function to measure the influence of including a data point in the training set on model response.

Answer: D

What are possible options for addressing problematic behavior in the case of LLMS

A. Prompt Engineering
B. Model Monitoring
C. Choosing a different foundational model
D. All of the above

Answer: D

Which of the following are common benchmarks for open source leaderboards?

A. BLEU, ROUGE, ELO, HellaSwag
B. MMLU, HellaSwag, A12 Reasoning Challenge, Truthful QA
C. MMLU, ROUGE, ELO, Truthful QA
D. BLEU, ELO, HellaSwag, MMLU

Answer: B

Which of the following are not key steps in chain-of-verification?

A. Initial Baseline Response
B. Verification Question Generation
C. Execute Verification
D. Generate Embeddings

Answer: D

For Large Language Models, what might constitute "Conceptual Soundness"?

A. Model Architecture
B. Training Data
C. Explanations for choices of Training Data and Model Architecture
D. Explanations for why choices of Training Data and Model Architecture are reasonable for the use case that the model will be applied to
E. All of the above

Answer: E

Which of the following are examples of guardrails?

A. Content Filter Guardrails
B. Privacy Guardrails
C. Explainability Guardrails
D. Bias Mitigation Guardrails
E. All of the above

Answer: E

Which of the following are not types of attacks against a LLM?

A. Hijack Response
B. Trick Response
C. Riddle Response
D. Simulation

E. Life Threat

Answer: C

Why might publically available leaderboards not be entirely trustworthy?

A. Benchmarks are not task-specific
B. Some model entries may be fraudulent
C. Results aren't reproducible
D. All of the above

Answer: D