UNIVERSITI MALAYA

EXAMINATION FOR THE MASTERS OF DATA SCIENCE

ACADEMIC SESSION 2024/2025 : SEMESTER I

WQD7009 : BIG DATA APPLICATIONS AND ANALYTICS

January 2025

INSTRUCTIONS TO CANDIDATES

Answer **ALL** questions. (30 marks)

Alternative assessment

This assessment consists of two alternative assessments: - an industry-based course completion and a group poster design project.

Alternative Assessment 1

- This is an individual assessment, carrying 15% of the total marks for the module.
- The objective of this assessment is to complete **one** course from the Google Cloud Data Analytics Certificate.
- **Duration:** Two weeks

Alternative Assessment 2

- This is an infographic-based poster design, conducted as a group assessment (you will remain in the same group as for the class group project).
- This group **assessment carries 15% of the total marks** for the module.
- You are required to prepare an infographic (maximum of two pages) on the given topic.
- **Duration:** Two days

(This question paper consists of 1 question on 3 pages)

Alternative Assessment Two - Duration: 18th Jan 2025 (12.00 pm) till 20th Jan 2025 (11.59 pm)

Case Study – 15 % **Duration for Task Completion:- 12 hours**

This is a group task; you will remain with the same group as in the continuous assessment group project.

You are required to prepare an infographic and a simple explanation, with a **maximum of two pages.**

Data Pipelines in Large Language Models (LLMs) for Big Data Analytics
Data pipelines play a crucial role in large language models (LLMs) by enabling efficient big data processing and analytics. By streamlining data collection, transformation, and storage, these pipelines ensure that LLMs can process vast datasets for advanced optimization, pattern recognition, and predictive modeling. Applications span across domains like healthcare, finance, and education. When effectively integrated, data pipelines enhance the scalability and real-time capabilities of LLMs, driving innovation and delivering actionable insights.

Tasks for the Case Study

1. Identify **two journal** articles that apply data pipelines, large language models (LLMs), and big data analytics within the same model or domain, such as LLaMA 3.1, Mistral 7B, Falcon 180B, or BLOOM. Ensure both articles focus on the same models in a similar field.

2. Thorough Reading
   Read and comprehend the methodologies, training data, and outcomes discussed in the articles, particularly in the context of data pipelines, LLMs, and big data or data analytics.
   Note similarities, differences, and any unique approaches related to the models or domains (e.g., LLaMA, Mistral, Falcon, BLOOM).

3. Critical Analysis
   Evaluate the selected papers for:
   - Their methodologies and effectiveness in applying data pipelines and LLMs to solve the problem.
   - Best practices and innovative solutions proposed in the studies.
   - Limitations and areas where improvements can be made
4. Poster Design
   - Summarize your findings and comparative analysis in a visually appealing poster design, focusing on the application of LLMs and data pipelines.
5. You are required to provide a comparative analysis between the studies.

Mandatory content in Infographic:-

1. Introduction
Data pipelines and large language models (LLMs) enable efficient processing of vast datasets, enhancing problem-solving in big data analytics across various domains.

2. Problem Statement
Defines domain-specific problems and explains how LLMs and data pipelines provide innovative analytical solutions for challenges in fields like finance, healthcare, or physics.

3. Methodologies
Utilizes data pipelines, pre-trained LLMs (e.g., LLaMA, Mistral, Falcon), hybrid analytical models, and optimized big data processing techniques.

4. Strength(s) and Weakness(es)

Strengths: Scalability, accuracy, flexibility, and adaptability of LLMs in big data analytics or pipeline.
Weaknesses: High computational cost, data privacy concerns, and complexity of implementation.

5. Findings on Tools & Best Practices
Highlights the tools and techniques used, such as pipelines, fine-tuned LLMs, performance metrics, and innovative solutions for handling large-scale data.

6. Future Directions
Suggests developing advanced algorithms, improving data preprocessing for LLMs, expanding applications, and enhancing scalability and computational efficiency.

Rubrics Outline for Poster Design (15 Marks)

1. Introduction (2 Marks)
2. Problem Statement (2 Marks)
3. Methodologies (3 Marks)
4. Strength(s) and Weakness(es) (2 Marks)
5. Findings & Best Practices (3 Marks)
6. Future Directions (3 Marks)
7. References.

Note:

1. Specific instructions must also be provided on Spectrum.
2. Please adhere to the instructions given.
3. Team member names must be listed on the first page.
4. Submit the file through Spectrum, naming it with **the group number and team leader’s name.** Files must be in PDF format.

TAMAT
END