Spring 2025

Extra Credit Assignment (20 points)
Due: April 25, 2025, 11.59 PM

Seminar Analysis Report
(2-page maximum, 12pt Times New Roman, 1" margins)

Student Information

Name: Adarsh Mohan

ASU ID: 1233380241

Date: 04/25/2025

- Seminar Title attended:

☐ "Perspectives on AI Risk" (Dr. William Regli)
☐ "LLM Data Preprocessing Pipeline Lecture" (YouTube)
☐ Other (provide title and speaker's name and qualification)

For students selecting "Other Approved Talk," here are relevant options:

1. **Data Processing at Scale for Training LLMs** (NVIDIA AI Summit 2024) - – watch and summarize in 2 pages
   Covers NeMo Curator and RAPIDS library optimizations1
   Link: NVIDIA AI Summit Talk

2. **Large-Scale Data Process and ML Pipelines** (Ray AIR Deep Dive) - – watch and summarize in 2 pages
   Explores distributed data loading/preprocessing for GPU saturation
   Link: YouTube

3. **Scaling Unstructured Data Processing with Ray Data** (Anyscale 2024) - – watch and summarize in 2 pages
   Demonstrates streaming batch models and adaptive scheduling5
   Link: YouTube

4. **Large-Scale Data Processing with DataChain** (Emerging Tools 2024) – watch and summarize in 2 pages
   Combines warehouse power with distributed clusters
   Link: YouTube

5. **Data @Scale Conference Talks** (May 2024) – watch any one and summarize in 2 pages
   Covers GenAI infrastructure challenges and solutions26
   Link: Conference Agenda

LLM Data Preprocessing Pipeline Lecture

Name- Adarsh Mohan

ASU ID- 1233310241

1. ***Key Thesis Identification***
   What central argument or framework did the presenter establish about data processing? Identify 2-3 core propositions.

   To train a large language model, building a data processing pipeline is crucial. This is how we take raw text from thousands of sources and convert it into input that can be fed to train the large language model.

   We also need to consider what the output will be. As we know LLMs under the hood is a classification task, which selects the next most probable word based on a defined vocabulary. Sp to predict the next word, capturing the meaning is important.

2. ***Technical-Social Intersection (answer the relevant one)***
   For Data Pipeline talk: What 3 pre-processing steps are most critical for LLM performance, and why? Relate to scale challenges.

   The presenter has identified 3 key steps while preprocessing the data for training an LLM. These steps are

   1. Tokenization – To find an efficient and manageable way to identify words and punctuations in a text document.
   2. Token Embeddings – The goal here is plot the words into a vector of multiples dimensions, to capture the meaning and semantical relationships.
   3. Positional Embeddings – LLMs also need to capture the position of the token in an input sequence. This is also done via positional embedding.

   Both positional and token embedding vectors are generated by training a neural neural. Additionally, to create the final input embeddings, the addition of token embedding and positional embedding is done to encode both the position and meaning of the word in context.

3. ***Critical Engagement***
   Identify one claim you found underdeveloped or problematic. Explain your reasoning using:
   a) Course concepts (e.g., distributed processing constraints)
   b) Real-world implementation challenges

   In byte pair encoding (used by GPT-2), the moment where we stop identifying the byte pairs and stop the tokenizing is ambiguous and based on trail and error. It will never be able to perfectly capture the root words as it intended to do, but will do the

job sufficiently well, as it worked for GPT-2. GPT-2 was trained with 50,000 tokens while GPT-4 was trained with around 100,000 tokens, but GPT-4 used different tokenizing technique.

4. ***Scalability Implications***
   How do the seminar's insights inform large-scale data processing architectures? Discuss either:
   a) AI risk mitigation infrastructure needs, or
   b) Preprocessing pipeline optimization strategies

   Since LLMs use large unstructured text data, having a versatile document store is important. The presenter mentioned adding a "|end of text|" token to capture different sources of data. To help with the preprocessing, this can be done in databases like No-SQL which is also follows a key value pair model just like token and token id, we can add the "|end of text|" token to each document in the database. This can help optimize the strategy while processing.

5. ***Course Synthesis***
   Connect 2 specific techniques/topics from CSE 511 lectures to the seminar's content. Example:
   "The speaker's emphasis on _____ aligns with our MapReduce optimization strategies through _____."

   Databases like No-SQL also follows a key value pair model just like token and token id, we see while traing LLMs.
   Also, we can define number of workers/nodes in dataloader which fetches an input-output target pairs using a sliding window approach. This can be done by directly defining the number of threads we want to for parallel processing. This is a direct application of lectures we have seen in class of parallel data processing.