EE416 AI with Deep Learning

Semester B 2025-2026

Assignment 3

Due: April 18, 2026, at 11:00 PM

"It is not that I'm so smart. But I stay with the questions much longer." ~ Albert Einstein

Section A [60 marks]

Question 1 [10 marks]

(a) How does the Skip-Gram model differ from the CBOW model in Word2Vec?

(b) What are the key components of an LSTM cell?

(c) Describe the gating mechanism in GRUs.

(d) What are the benefits of using Parameter-Efficient Fine-Tuning (PEFT) for LLMs?

(e) How does the Transformer model handle long-range dependencies in text?

Question 2 [12 marks]

(a) Explain the different types of attention mechanisms utilized in Transformer neural networks. Describe how these attention mechanisms contribute to the overall architecture of the Transformer model and their significance in capturing relationships and dependencies in textual data. [4 marks]

(b) Compare the architectures of GPT, BERT, and T5 models based on the Transformer. Highlight the key differences in their designs, functionalities, and how they leverage the Transformer's building blocks for various NLP tasks. [4 marks]

(c) What is the primary contribution of the T5 model to language model pretraining? [4 marks]

Question 3 [5 marks]

Based on the Sinusoidal Positional Encoding method to compute the first 6 position vectors (pos = 0, 1, 2, 3, 4, 5) with $d_{model} = 8$.

The sinusoidal positional encodings are represented as:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

<u>Question 4</u> [13 marks]

(a) Why is prompt engineering crucial for LLM performance? [4 marks]

(b) What role does temperature play in controlling LLM output? [4 marks]

(c) Describe how **Retrieval-Augmented Generation (RAG)** helps reduce hallucinations in large language models. Discuss its advantages and limitations.

[5 marks]

<u>Question 5</u> [20 marks]

(a) How do transformers improve upon traditional sequence-to-sequence (Seq2Seq) models?

[4 Marks]

(b) Explain how self-attention scores are computed in Transformers. In your answer, describe the roles of the query (**Q**), key (**K**), and value (**V**) matrices, the purpose of the scaling factor, and how the final context-aware representation is obtained. [2 Marks]

(c) Given the matrices for query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$, compute self-attention matrix $\mathbf{Z} = \text{Attention}\ (\mathbf{Q}, \mathbf{K}, \mathbf{V})$ using the scaled dot-product attention mechanism. [7 marks]

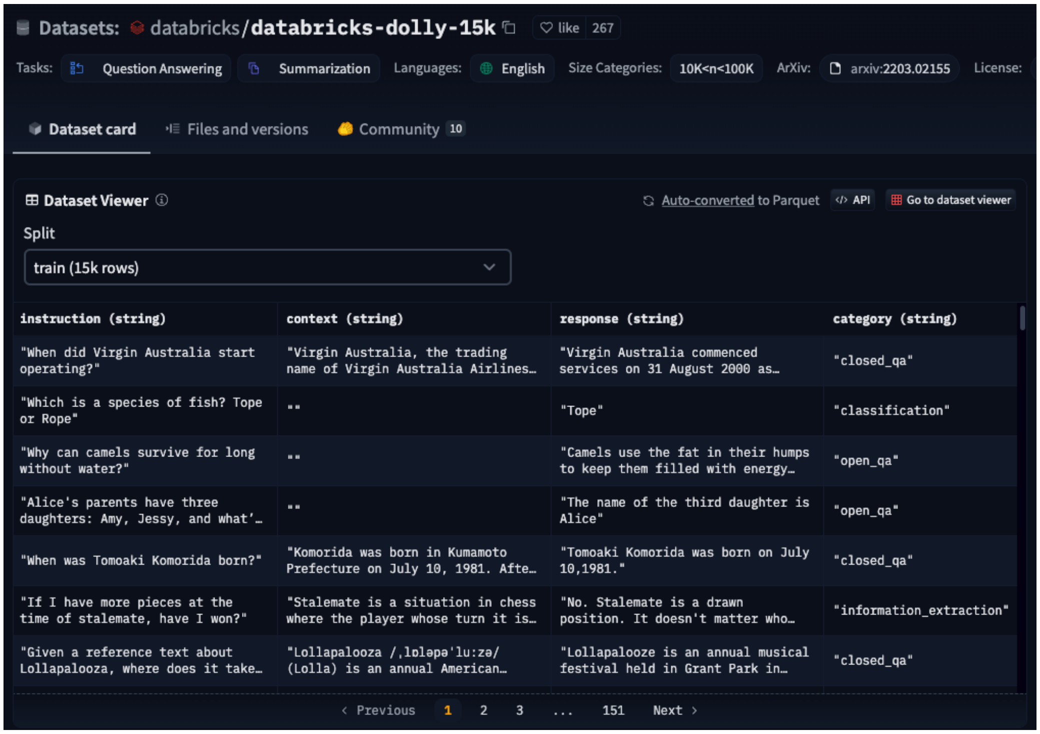
$$\mathbf{Q} = \begin{bmatrix} 4 & 2 \\ 1 & 7 \\ 8 & 1 \end{bmatrix} \qquad \mathbf{K} = \begin{bmatrix} 7 & 3 \\ 5 & 1 \\ 1 & 6 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} 1 & 2 \\ 8 & 4 \\ 9 & 3 \end{bmatrix}$$

(d) Define "hallucination" in the context of large language models (LLMs) and explain how supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) were employed between 2021 and 2022 to address this problem. [4 Marks]

(e) Explain why ChatGPT (2022) was regarded as a breakthrough in conversational AI, and identify one key advancement it offered over earlier models. [3 Marks]

Section B: Programming Assignment [40 marks]

In this section, students are required to fine-tune a small language model with instruction-following dataset that that consists of prompt/response pairs along with any contextual information that can be used as input when training the model. The databricks/databricks-dolly-15k is one such dataset that provides high-quality, human-generated prompt/response pairs.



instruction (string)	context (string)	response (string)	category (string)
"When did Virgin Australia start operating?"	"Virgin Australia, the trading name of Virgin Australia Airlines…	"Virgin Australia commenced services on 31 August 2000 as…	"closed_qa"
"Which is a species of fish? Tope or Rope"	""	"Tope"	"classification"
"Why can camels survive for long without water?"	""	"Camels use the fat in their humps to keep them filled with energy…	"open_qa"
"Alice's parents have three daughters: Amy, Jessy, and what'…	""	"The name of the third daughter is Alice"	"open_qa"
"When was Tomoaki Komorida born?"	"Komorida was born in Kumamoto Prefecture on July 10, 1981. Afte…	"Tomoaki Komorida was born on July 10,1981."	"closed_qa"
"If I have more pieces at the time of stalemate, have I won?"	"Stalemate is a situation in chess where the player whose turn it is…	"No. Stalemate is a drawn position. It doesn't matter who…	"information_extraction"
"Given a reference text about Lollapalooza, where does it take…	"Lollapalooza /ˌlɒləpəˈluːzə/ (Lolla) is an annual American…	"Lollapalooze is an annual musical festival held in Grant Park in…	"closed_qa"

A Colab notebook has been created to demonstrate how to fine-tune the "LLAMA-3.2-1B-bnb-4bit" model using a specific dataset and evaluate it using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Students are strongly encouraged to begin with this notebook and experiment with other state-of-the-art small or large language models, such as LLaMA3.2-3B, LLaMA3.1-8B, QWen2.5-0.5B, QWen2.5-1B, and others, to achieve better performance and gain experience in fine-tuning LLMs.

https://colab.research.google.com/drive/17yMJ6AnuuPA2l6mqUD9Io5It8skgO7cN?usp=sharing

- ROUGE-1 and ROUGE-2 measure the overlap of unigrams (single words) and bigrams (two consecutive words) between the generated and reference summaries, respectively.
- ROUGE-L measures the longest common subsequence (LCS) between the generated and reference summaries. It considers sentence-level structure similarity and is useful when the summaries are not necessarily consecutive or contiguous.

Higher ROUGE scores indicate better performance. Precision measures the fraction of relevant words/phrases in the generated summary, while recall measures the fraction of relevant words/phrases from the reference summary that were captured in the generated summary.

The marking scheme is as follows:

- ROUGE-L ⩾ 0.185 (10 marks)
- ROUGE-L ⩾ 0.190 (15 marks)
- ROUGE-L ⩾ 0.195 (20 marks)
- ROUGE-L ⩾ 0.200 (25 marks).
- ROUGE-L ⩾ 0.205 (30 marks).
- 10 marks will be awarded for the coding style of your Jupyter notebook and the quality of your summary and analysis of the model.

Your Jupyter notebook should include a written summary and analysis of your model architecture, training procedure, experiments, results, and conclusions for assessment.

The test set is created by initially loading the entire dataset as the 'train' split and then creating a test split from this data. This method results in a test set that comprises 5% of the original data, with the remaining 95% used for training. The seed=4016 ensures the reproducibility of the split.

Discuss the techniques that worked well and how you improved upon the baseline model. Include relevant visualizations.

The top 5 ROUGE-L will each receive an additional 10 marks. In case of a tie, preference will be given to models with fewer parameters.

Submit a zip file that consists of pdf file of your answers of Section A and the Jupyter notebook file of Section B to the Assignment 3 on CASVAS with following file name format:

- Filename format : Assignment03_StudentName_StudentID.zip
- Filename example: Assignment03_Chen_Hoi_501234567.zip