FALL 2025 | DATA 266 | Homework -2

Deadline – 11.59 PM – 10/13/2025

20 Points

Question # 1 (8 pts):

Apply combinations of advanced LLMs and prompt engineering techniques on Quora Question pairs dataset to predict whether a question pair duplicate or not duplicate (https://huggingface.co/datasets/AlekseyKorshuk/quora-question-pairs). Split the dataset into 95%:5% for train test ratio (maintaining balanced class distribution as test data). Use the test data for this task. Finally, compare the performance in terms accuracy, precision, recall, and f1-score.

- LLMs: select any two LLMs released in 2024 or 2025
- Prompting techniques:
  - Zero shot prompting
  - 5 shots prompting
  - Chain-of-thought prompting with 5-shots
  - Self-consistency prompting with 5-shots
  - Tree of Thought prompting

You must provide a table including all of your prompts similar to the table 1

Table 1 Summary of the prompting techniques used in the study, along with an example for each technique

Prompting technique	Example
Zero-shot	You will be provided with a requirement statement, and your task is to classify it as either Functional or Non-functional. Return only the class label, with no additional text. Requirement: {req_text} Class →
Few-shot	You will be provided with a requirement statement, and your task is to classify it as either Functional or Non-functional. Return only the class label, with no additional text. Here are some examples: Requirement 1: The product will be able to delete conference rooms. Class → Functional Now classify: {req_text} Class →
Persona	Act as an experienced Requirements Analyst and classify the requirement as either Functional or Non-functional. Return only the class label, with no additional text. Requirement: {req_text} Class →
Chain of thought (CoT)	You will be provided with a requirement statement, and your task is to classify it as either Functional or Non-functional. Think step by step before assigning a label. Requirement: {req_text} Class →

Question # 2 (6 pts):

Use the 20 Newsgroups dataset (20 categories such as rec.sport.baseball, sci.space, talk.politics.mideast, etc.).

1. Apply BERTopic with an advanced SBERT model (e.g., all-roberta-large-v1 or instructor-xl).
2. BERTopic will cluster documents into topics and give top keywords for each cluster.
3. From the keywords, manually assign human-readable labels to the topics (e.g., Sports – Baseball, Computers – Hardware).
4. Select at least 100 test documents. Compare your assigned topic labels against the dataset's ground-truth categories. Write a prompt for an LLM to act as a Judge that evaluates whether the predicted topic label is semantically relevant to the ground-truth label (for example, ground truth = Sports, predicted = Football should be considered relevant).

```
from sklearn.datasets import fetch_20newsgroups
newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers', 'footers', 'quotes'))

```

Question # 3 (6 pts):

Use the following code – Siamese network using contrastive loss.

https://keras.io/examples/vision/siamese_contrastive/

Train a Siamese network using Quora Question pairs dataset to predict whether a question pair duplicate or not duplicate (https://huggingface.co/datasets/AlekseyKorshuk/quora-question-pairs). Use any recently published pre-trained model to generate fixed size embeddings for each question pairs, and feed both question embeddings into a cosine similarity layer (distance layer), use the distance layer as input to a classification layer (sigmoid). Train the network for minimum 50 epochs using train dataset from the question 1. Use test dataset (from question 1) for evaluation purpose. Present model accuracy, recall, precision and f1 score. Show contrastive loss learning curve. Finally, compare the model performance with different prompt engineering techniques from question 1.

You are required to follow:

1. Submit **one** MS/PDF/Scanned document:

- Include all the steps of your calculations.
- Include the summary of the model.
- Attach screenshots of your code.
- Attach screenshots – showing first few epochs of model training.
- Attach screenshots of the important code outputs such as confusion matrices, learning curves, and classification reports.

2. Source code:

a. Python (Jupyter Notebook)

b. Ensure it is well-organized with comments and proper indentation.

- **Failure to submit the source code will result in a deduction of full/partial points.**
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.

References:

Binkhonain, M., & Alfayez, R. (2025). Are prompts all you need? Evaluating prompt-based Large Language Models (LLM) s for software requirements classification. *Requirements Engineering*, 1-21.