Term Project

2025 NTHU Natural Language Processing

Hung-Yu Kao

IKM Lab TAs

Project Description

In this term project, you will be divided into groups of up to five members. There will be five topics available, and each group will choose one topic to work on.

Topics	Task type	Submissions per day
Topic 1: Chatbot Arena Human Preference Predictions	Classification	5
Topic 2: FinanceRAG Challenge	RAG	5
Topic 3: LLM Prompt Recovery	NLG	5
Topic 4: Understanding of Complex Q&A Content	Regression	5
Topic 5: WattBot 2025	RAG	5

Term Project 規則

- 請各組從五個任務中自行選擇其中一項來進行 Term project
- 你們需要：
  - 統整兩個以上的 code (這五個 Kaggle 任務都已經有許多存在的code)，並**產生自己的作法**
  - 分析自己作法和已經存在的code作法相比的**優勢**與**劣勢**
  - **提出新的做法**
  - **努力取得前面的排名**，你們組在Leaderboard上的排名將納入Term Project評分，本次競賽皆為 **Code competition**，因此需要仔細閱讀競賽規則，建立好程式碼後，上傳到 Kaggle

Topic 1
Chatbot Arena Human Preference Predictions

Official Link: LINK

This task is to predict which model's response (A, B, or tie) a human judge would prefer for a given prompt and pair of LLM responses.

- **Input**: You will receive a dataset of user interactions where each row contains a prompt and the corresponding response_a and response_b from two different language models, and winner_model_a/b/tie, indicating which response the judge preferred.
- **Output**: You must predict which model's response (a, b, or tie) is more likely to be selected by the judge for each prompt in the test set.

Evaluation: Submissions are scored based on the accuracy of predicted judge preferences compared to the ground truth.

Data Description

Input: prompt and the response of A and B

output: which one does human prefer (A, B, tie)

prompt	response_a	response_b	winner_model_a	winner_model_b	winner_tie
51734 unique values	56566 unique values	56609 unique values	 0 1	 0 1	 0 1
["Is it morally right to try to have a certain percentage of females on managerial positions?","OK, ...	["The question of whether it is morally right to aim for a certain percentage of females in manageri...	["As an AI, I don't have personal beliefs or opinions. However, I can tell you that the question of ...	1	0	0
["What is the difference between marriage license and marriage certificate?","How can I get both of ...	["A marriage license is a legal document that allows a couple to get married. It is issued by a gove...	["A marriage license and a marriage certificate are two different legal documents that have separate...	0	1	0

Topic 2
FinanceRAG Challenge

Official Link: LINK

The task is to build an RAG system that, given a financial query, retrieves relevant contexts from a document collection and generates a concise, accurate answer.

Data:

- **Input**: You will receive a dataset where each row contains a query with “_id”, “title”, “text”, and other metadata. A subset of relevance labels is provided in TSV files with “query_id”, “corpus_id”, and “score”.
- **Output**: You must predict relevance scores for query–document pairs, which indicate how relevant each document is to the corresponding query.

Evaluation: The retrieval process will be evaluated using NDCG@10, which measures how well the predicted ranking of documents matches the ideal ranking for the top 10 results.

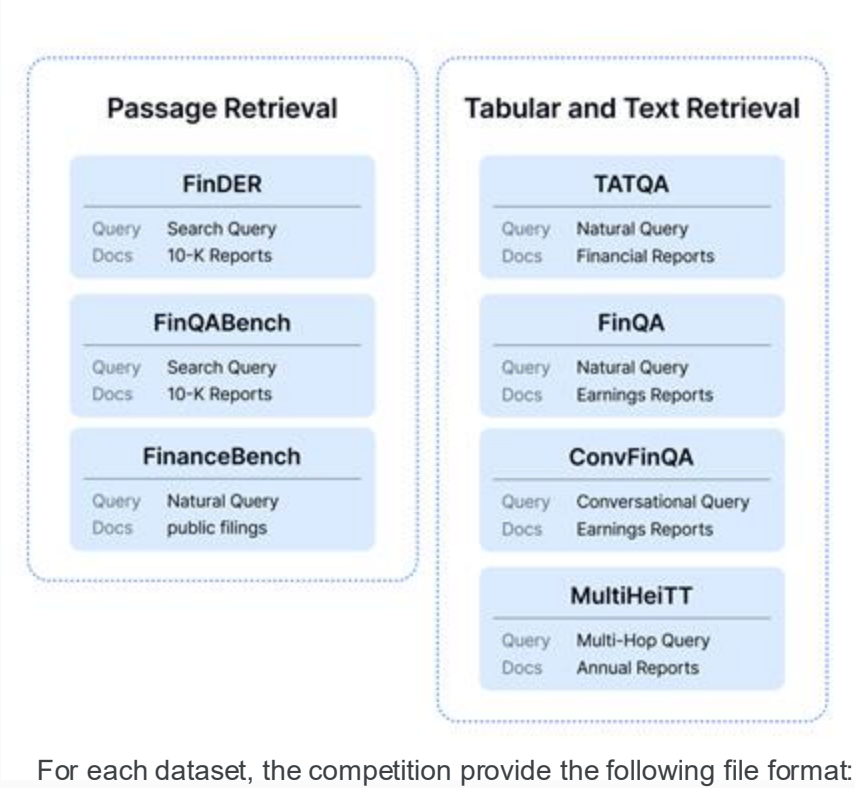
Data Description

Here's an expanded description including explanations for each line:

- **_id**: A unique identifier for the context/query.
- **title**: The title or headline of the context/query.
- **text**: The full body of the document/query, containing the main content.

To assist participants with their development, the organizer have decided to release 30% of the answer labels. These labels are provided in a `TSV` file, which follows a specific format. The `TSV` files are structured with three columns:

- **query_id**: The unique identifier for each query.
- **corpus_id**: The unique identifier for the relevant document in the corpus.
- **score**: A relevance score that indicates how relevant the `corpus_id` document is to the `query_id` query.



For each dataset, the competition provide the following file format:

{dataset_name}_qrels.tsv

Topic 3
LLM Prompt Recovery

Official Link: LINK

The goal of this competition is to recover the LLM prompt that was used to transform a given text.
Data:

- **Input**: The text before and after rewriting
- **Output**: The prompt needed to be predicted

Evaluation: **Sharpened Cosine Similarity (SCS)** between embeddings of the predicted and true prompts, computed with the sentence-t5-base model (exponent = 3).

Data Description

A unique ID for each data

Output: the prompt needed to be predicted

id	original_text	rewrite_prompt	rewritten_text
1 total values	1 unique value	1 unique value	1 unique value
-1	The competition dataset comprises text passages that have been rewritten by the Gemma LLM according ...	Convert this into a sea shanty: """The competition dataset comprises text passages that have been re...	Here is your shanty: (Verse 1) The text is rewritten, the LLM has spun, With prompts so clever, they...

Input: The text before and after rewriting

Remark:
You should generate additional data to train your model.

Topic 4
Understanding of Complex Q&A Content

Official Link: LINK

The task is to develop AI models that can predict subjective qualities—such as helpfulness, relevance, and clarity—of question–answer pairs like human.

Data:

- **Input**: You will receive a dataset where each row contains “question_title”, “question_body”, “answer”, and other metadata like “host”, “category”, etc.
- **Output**: You must predict **30** different target scores for each Q&A pair, these scores are continuous numbers between 0 and 1.

Evaluation: Mean Column-wise Spearman's Correlation Coefficient (see more in the official link)

Data Description

Input: The QA information

question_title	question_body	question_user_na...	question_user_pa...	answer	answer_user_name	answer_user_page	url	category
3583 unique values	3583 unique values	3215 unique values	3422 unique values	6079 unique values	4114 unique values	4430 unique values	3583 unique values	TECHNOLOGY 40% STACKOVERFLOW 21% Other (2385) 39%
What am I losing when using extension tubes instead of a macro lens?	After playing around with macro photography on-the-cheap (read: reversed lens, rev. lens mounted on ...	ysap	https://photo.stackexchange.com/users/1024	I just got extension tubes, so here's the skinny. ...what am I losing when using tubes...? A v...	rfusca	https://photo.stackexchange.com/users/1917	http://photo.stackexchange.com/questions/9169/what-am-i-losing-when-using-extension-tubes-instead-of...	LIFE_ARTS

output: 30 target scores

question_asker_i...	question_body_cr...	question_convers...	question_expect_...	question_fact_se...
0.33 1	0.33 1	0 1	0 1	0 1
1.0	0.3333333333333333	0.0	0.0	0.0

25 columns are omitted here

Topic 5
WattBot 2025

Official Link: LINK

The task is to build a RAG-based system that answers sustainability questions about AI's environmental impact with evidence-backed, well-cited responses.

Data:

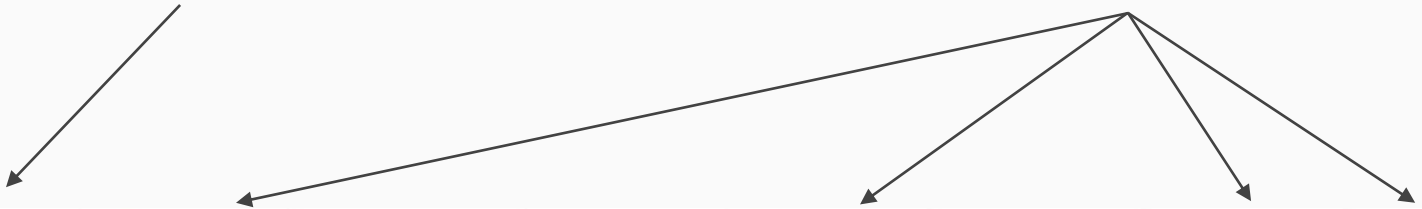
- **Input**: A sustainability-related query and an article corpus.
- **Output**: A structured response containing: "answer" (A concise natural-language answer), "ref_id" (ID(s) of cited document(s)), "supporting_materials" (Relevant excerpts, tables, or figures from the references), explanation(Reasoning linking the supporting materials to the final answer).

Evaluation: Scored with a custom WattBot score that evaluates three fields for every question and returns a weighted accuracy between 0 – 1 (see more in the official link)

Data Description

Input: The query and the corpus

output: The answer, citation, supporting material and explanation



id	question	answer	answer_value	answer_unit	ref_id	ref_url	supporting_mater...	explanation
41 unique values	41 unique values	TRUE 17% FALSE 15% Other (28) 68%	1 17% 0 15% Other (28) 68%	is_blank 41% percent 12% Other (19) 46%	['ebert2024'] 12% ['li2025b'] 10% Other (32) 78%	['https://arxiv.org... 12% ['https://arxiv.org... 10% Other (32) 78%	is_blank 5% Section 4.3 Transp... 5% Other (37) 90%	Quote 54% Figure 10% Other (15) 37%
q003	What is the name of the benchmark suite presented in a recent paper for measuring inference energy c...	The ML.ENERGY Benchmark	ML.ENERGY Benchmark	is_blank	['chung2025']	['https://arxiv.org/pdf/2505.06371']	We present the ML.ENERGY Benchmark, a benchmark suite and tool for measuring inference energy consum...	Quote
q009	What were the net CO2e emissions from training the GShard-600B model?	4.3 tCO2e	4.3	tCO2e	['patterson2021']	['https://arxiv.org/pdf/2104.10350']	"Training GShard-600B used 24 MWh and produced 4.3 net tCO 2 e."	Quote

Term Project Schedule

Term Project Schedule (changes might be made)

Every team is required to deliver an official oral presentation and submit a written report.

	Start	Deadline
Topic Announcement	11.6	/
Topic Selection	11.6	11.11
Checkpoint 1: PPT & Video submission (30%)	/	12.04 (W14)
Checkpoint 2: Peer Review (20%)	12.09 (W15)	12.11 (W15)
Checkpoint 3: Oral/Poster Presentation (selected)	12.16 (W16)	12.18 (W16)
Checkpoint 4: Report Submission (40%)	/	12.25

Performance ranking (10%)

Topic Selection

Each team is required to rank the five topics in order of preference, from highest to lowest. We will prioritize assigning teams to their top-preferred topics

However, if too many teams select the same topic, some will be reassigned based on their stated preferences. We will ensure that the number of teams per topic remains approximately balanced.

After the groups are formed, we will provide a Google Sheet where each group should indicate their topic preferences.

Checkpoint 1: PPT & Video submission (30%)

Every team should submit two materials:

1. Presentation PPT about your competition and method
2. Presentation Record (10-15 min, youtube video, the URL should be on the first page of the PPT)
3. 基本內容: 介紹資料, 分析至少兩組Kaggle codes, 資料困難點敘述
4. Key content: What did you learn from Kaggle codes? What's your new idea? New results?

	Filename rule	**Filename example**
PPT file	NLP_topic_groupName.pptx	NLP_topic1_group1.pptx

Checkpoint 2: Peer Review (**20%**)

Each team needs to review the PPTs/videos of other teams in the same competition. You need to evaluate and rank their work based on three criteria:

1. Creativity
2. Analysis Completeness
3. Presentation Quality

The evaluation form will be released on 12/5.

Scoring Criteria (reference)

	Score	Description
Creativity	0	Off-topic or shows no relevant creative contribution.
	1	Minimal effort in idea generation; no sign of originality.
	2	Mostly standard and predictable; lacks noticeable creative elements.
	3	Reasonably solid but follows conventional patterns; limited originality or new perspectives.
	4	Shows some originality or thoughtful variation in approach; includes at least one creative or non-obvious idea.
	5	Demonstrates highly original ideas or approaches; shows clear insight and innovative thinking beyond standard solutions.
Analysis Completeness	0	Incomplete, irrelevant, or incorrect analysis.
	1	Minimal analysis with little evidence of understanding.
	2	Partial or superficial analysis; several important aspects are missing or unclear.
	3	Addresses the main parts of the topic but misses some relevant points or details.
	4	Mostly complete and accurate analysis with minor gaps or small logical weaknesses.
	5	Thorough, accurate, and logically consistent analysis; all key aspects of the topic are addressed and well-supported.
Presentation Quality	0	No meaningful presentation delivered.
	1	Disorganized and unclear; audience has difficulty understanding the content.
	2	Hard to follow due to poor structure or unclear explanation.
	3	Understandable overall, though with some disorganization or inconsistent delivery.
	4	Clear and organized presentation with good pacing and minor issues in clarity or engagement.
	5	Extremely clear, well-structured, and engaging; visuals, timing, and delivery are polished and professional.

Checkpoint 3: Oral/Poster Presentation

We will select three outstanding teams for each topic. According to the leaderboard score, peer review score, and ppt score in the CP1.

Teams	Requirement
Selected Teams	Have oral presentation on class
Other Teams	Submit a poster about your work Filename rule: NLP_topic_poster_groupName.pdf Filename example: NLP_topic1_poster_group1.pdf

Checkpoint 4: Report Submission (**40%**)

Before the end of the semester, each team must submit a technical report.

The report is recommended to include the following sections:

1. Task description
2. Related works (other Kaggle codes)
3. Methods (what you have tried)
4. Experiments (including the final leaderboard score)
5. Conclusion (what you learned)
6. Reference

Performance ranking (**10%**)