
Due Date Oct 14, 2025, 11:59pm
Grade 10% of the project points
Late Submissions 30% per day
Teams Projects can be done individually or in teams of 2.

Project Proposals

By October 14, you will need to submit a ≈ 1 to 2 page project proposal which will include:

1. Your cool team name and the names & ID of the students in your team.
2. A description of the goal of the project
3. A proposed methodology to implement the project
4. The evaluation methodology
5. A timeline and role of each team member if applicable
6. All references you have consulted so far
7. In case you propose your own project, you need to *certify that the project has not and will not be used for credits in other courses or directly for your thesis.*

Below are some suggestions for your COMP 6781 Project. You are free to propose your own project, but please discuss it with me first and make sure I give you my *ok, in principle* before you submit your project proposal.

If you pick a project from the list below, it is ok if several teams work on the same project. However, in that case, there should be no communications across these teams.

Proposed Projects

1 Open-Ended Research Projects

1.1 Analysis of cultural biases in language models

Motivation: Language is more than transmission of knowledge; it is also a reflection of culture, values and beliefs. Language is a primary tool for humans to learn how the world works and how our society works. Today, LLMs are trained on large quantities of texts selected mostly based on the language. Some LMs have been developed specifically for specific domains (eg. medical), but most training corpora do not consider the social and cultural aspects conveyed. Consequently, LLMs contain an amalgam of human cultural information that do not necessarily convey the values and beliefs of any specific community. Interacting with LLMs that do not take into account the specific social environment, cultural values and beliefs of the participants may lead to suboptimal communication. Imagine for example, an LLM advising a teen in distress in a society with traditional versus western values.

Goal: This project will focus on the evaluation of cultural notions learned by LLMs by exploring how we can automatically evaluate the cultural notions learned by an LLM and compare these notions across LLMs.

Some starting points:

- Artificial intelligence needs to be trained on culturally diverse datasets to avoid bias
- CultureLLM: Incorporating Cultural Differences into Large Language Models
- Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions
- Workshop on Global AI Cultures at ICLR-2024

1.2 DISRPT-2025

Motivation: Language is more than a bag of sentences (see the slides on Introduction to NLP). Discourse parsing is the process of analyzing the structure and relationships between different units of a text beyond individual sentences in order to understand how these units are connected to form coherent meaning. Discourse parsing is typically divided into 3 sub-tasks

1. Discourse segmentation: segmenting texts into elementary units
2. Connective identification: identifying explicit connectives that signal a discourse relation (eg *but*, *however*, *furthermore*, *because*)
3. Relation classification: identifying the relation between discourse segments (eg. *cause*, *purpose*, *result* etc)

For example, in : *Mary was sick, but she went to the party.*

1. the segments are: 1) Mary was sick 2) but she went to the party.
2. the discourse connective is: "but"
3. the relation is a "contrast"

This year, the DISRPT-2025 shared task organized a competition for discourse parsing and provided a dataset and scripts to evaluate systems.

Goal: This project will focus on developing a model for one of the tasks above (segmentation, connective identification, or relation classification) and compare your model to the state-of-the-art.

2 Multi-lingual or Multi-class Classification Tasks

Motivation: Classifying texts into subjective categories such as misinformation, toxicity, or hate speech is a typical application in NLP. Several techniques can be used for this task (1) traditional word-based approaches like a MultiNomial Naive Bayes, (2) using word embeddings such as Word2Vec with a networks, or (3) fine-tuning pre-trained transformer models (e.g., BERT).

Goal: The following are interesting tasks with datasets. Pick one task, and runs at least 3 different classification models on them (note: changing the pre-trained model from BERT to another is considered the same approach). Compare the performance of these 3 approaches.

Some Datasets/Tasks:

1. Multilingual Machine Generated Text Detection, see here for plenty of datasets
2. Multilingual Promise Evaluation, see SemEval 2025 Shared Task 6 description and the training set here
3. Misinformation Detection, see this recent paper and the dataset here
4. Other datasets/classification tasks: You are free to use other tasks that interest you, as long as they are multilingual datasets or have multiple classes.