CITS4012 Natural Language Processing
Project Specification

Due: 17 October 2025, 11:59PM AWST

1 Objective

This project is to be completed in **a group of 2 or 3 students (maximum)**. You're allowed to complete this project individually if you prefer not to work in a group, but please note **NO bonus mark will be given for individual submission**. Every group needs to submit the Group Registration Form by **19th September**. We strongly recommend to start working early so that you will have ample time to discover stumbling blocks.

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is the task of determining the inference relation between texts. It aims to assess whether a given textual *premise* entails or implies a given *hypothesis*. NLI is a central problem in natural language understanding as it encapsulates the fundamental challenge of linguistic variability. In this project, you will have a textual entailment dataset, which is created from multiple-choice science exams and web sentences. Each question and the correct answer choice are converted into an assertive statement to form the *hypothesis*, and relevant texts are retrieved from a large text corpus of web sentences to form the *premises*. Here are two examples for the cases of ***entails*** and ***neutral*** respectively:

***Premise*:** "Beats are the periodic and repeating fluctuations heard in the intensity of a sound when two sound waves of very similar frequencies interfere with one another."

Hypothesis: "When waves of two different frequencies interfere, beating occurs."

Label: entails

Premise: "During periods of drought, trees died and prairie plants took over previously forested regions."

Hypothesis: "Because trees add water vapor to air, cutting down forests leads to longer periods of drought."

Label: neutral

In this project, you are required to design and compare the different model structures for the given science-specific NLI task.

Note, instead of solely focusing on achieving higher performance, you should consider exploring novel architecture design and justify your decision processes, as grading is by-and-large based on your research process rather than the performance (see marking scheme at the end).

2 Dataset

You will be provided with:

1. A training set (`train.json`)
2. A validation set (`val.json`)
3. A test set (`test.json`)

Each instance in the dataset consists of the sentences of premise and hypothesis, and the label of entail or neutral. The dataset download link can be found on LMS.

3 Report Writing

You **MUST** use LaTeXfor writing your report and MUST use the ACL template. You must include your group number (not your name) under the title (using the `\author` field in LATEX and changing "review" to "final" to generate the final (sometimes called camera-ready) version: change `\usepackage[review]{acl}` to `\usepackage[final]{acl}`. We will not accept reports that are longer than the stated limits below, or otherwise violate the style requirements.

The report should be submitted as a PDF and contain no more than five(5) A4 pages of content, excluding team contribution and references. Therefore, you should consider carefully the information that you want to include in the report to build a coherent and concise narrative.

Below is the required sections:

3.1 Title

The title of your project and Group number.

3.2 Abstract

An abstract should concisely (less than 300 words) motivate the problem, describe your aims, describe your contribution, and highlight your main finding(s).

3.3 Introduction

The introduction should explain the problem and your understanding of its significance, difficulties and applications. You should give an overview of the your approach and the main results. Though an introduction covers similar contents as an abstract, a good introduction should cover more details for the problem discussions and references to existing works.

3.4 Methods

This section details your methods to the problem. This is where you describe the architecture of your neural networks, and any other input preprocessing and representation. Specifically, you are required to design and implement **three substantially different model architectures.**

- Each model must differ in **structural design** (e.g., encoder–decoder, transformer-based etc.).
- Simple substitutions of components (e.g., replacing RNN with LSTM or GRU) **do not count as different architectures**.
- Merely increasing or decreasing the number of layers (e.g., one-layer RNN vs. multi-layer RNN, or RNN vs. bi-RNN) **does not count as a different structure.**
- Your designs should demonstrate clear differences in how inputs are encoded, combined, or decoded to produce predictions.
- At least one model must include an **attention mechanism**, and you are encouraged to apply it in a novel or insightful way.

- You should provide a detailed description of each model with proper equations, notations, and architecture diagrams.
- You may take inspiration from published research, but if you do so, you must provide the appropriate references in your report.

3.5 Experiment Setup

This section should contain:

- **Dataset Description** describe the dataset and include any dataset analysis you have done.
- **Implementation Details** This should include **all the details** of how you ran your experiments including but not limited to: model configuration, learning rate, optimization methods, training time etc.

3.6 Results

This section contains the following:

- **Main Performance Comparison**: you should report, compare and analyse the performance of your three model structures on the test set. When you write results, please be aware of the following questions: Are they what you expected?; Better than you expected?; Is It worse than you expected?; Why do you think that is?; What does that tell you about your approach?
- **Ablation Study on Attention Mechanism**: You must experiment with different ways of applying attention. For example:
  - Applying attention at different positions (e.g., encoder-side, decoder-side, intermediate layers).
  - Using different types of attention (e.g., self-attention, cross-attention etc.).

  At least one ablation must be performed by varying attention within the same model structure, so you can isolate and analyse its effect. You may also explore attention variations across different model structures if you wish, but this is optional. Similarly, you should interpret and analyse the results you have.

- **Additional Ablation Study**: Beyond the attention mechanism study, you are required to perform at least one additional ablation study on a significant aspect of your models.
  - This may involve investigating the effect of important hyperparameters (e.g., embedding dimension, hidden size etc.) or the role of a particular model component (e.g., varying pooling strategy etc.).
  - Your chosen ablation must be substantive and meaningful. Trivial variations (e.g., simply increasing training epochs without justification) will not be accepted.
  - In your report, you must: clearly explain why you selected this ablation study and why it is important; present the experimental results of the ablation; provide a thoughtful analysis and interpretation of the findings.

  Note: the effective and fair comparison should keep all other settings the same, that saying, the ablation study can be done one just one of your best model structure.

3.7 Qualitative Results

You are required to conduct a qualitative analysis of your model. Specifically, you should select one or two sample instances from the test set and extract the attention weights of the sentence tokens. These weights must be visualised, and you may use any form of visualisation—either manually created diagrams or automatically generated plots with code. In your report, you should present the visualisation together with a clear analysis and interpretation of what it reveals about your model's behaviour. Your discussion should explicitly link the visualisation to your model's performance, highlighting how attention contributes to correct predictions or potential errors.

3.8 Conclusion

Summarise the main findings of your project, and what you have learnt. Highlight your achievements, and note the primary limitations of your work. If you like, you can describe avenues for future work.

3.9 Team Contribution

(doesn't count towards the page limit) If you are a multi-person team, briefly describe the contributions of each member of the team.

3.10 References

(doesn't count towards the page limit) Your references section should be produced using BibTeX.

4 Submission Method

The submission should be made via LMS (The submission portal will open on 13 October 2025). Only **ONE** group member is required to make the submission. You **MUST** submit two files:

- a PDF file, with filename: `CITS4012_YourGroupID.pdf`
- a `ipynb file`, with filename: `CITS4012_YourGroupID.ipynb`

  This file should include all your implementation of this project.

You can **optionally** submit a zip file that contains a README file describe how to run the code if it's not apparent from the documentation in your ipynb files, any of your trained models or any other files that's necessary for the marker to run your program.

5 Important Rules

You MUST follow the rules below. Any team found to break any of these rules will receive zero mark to their project.

1. You're encouraged to explore different models for the task. In terms of the sequence processing components you MUST use one of the following architectures: RNN, LSTM, GRU, and Transformer.You may use deep-learning libraries (e.g. PyTorch) to import these sequence processing components. You could read relevant publications to come up with a sensible design for your methods, but you MUST NOT copy any open-source code from any publications (in other words, you MUST implement the methods yourself).

2. The following deep-learning libraries are allowed: PyTorch, Keras, and TensorFlow. Huggingface is not allowed. Standard python libraries (e.g. numpy and matplotlib etc.) and NLP preprocessing toolkits (e.g. NLTK and spacy etc.) are allowed.

3. You could use pretrained word embeddings (e.g. Word2Vec/Glove), but you MUST NOT use any pretrained language model weights or checkpoints (e.g. BERT checkpoints), or any open-source or closed-source LLMs. **In other words, you MUST train your model from scratch using the provided data.**

4. The model described in the report MUST be faithful to the submitted code and running log that you submit. **You MUST include the running log (with the reported result/performance) in the submitted ipynb file.**

5. You are allowed to use code from the lab contents (provided that they don't conflict with any project rules), but you MUST NOT copy any open source project code from GitHub or other platforms.

6. You MUST NOT use models that cannot be run on Colab.

7. You MUST use the given code template for implementation.

8. You can clean the training dataset and not use every instances in the training data (e.g. remove duplicate instances). But please keep the test set as it is for your evaluation.

6 Late Submission of Assignment

A penalty of 5 per cent of the total mark allocated for the assessment item is deducted per day for the first 7 days (including weekends and public holidays) after which the assignment is not accepted. For the first two days there will be a penalty waiver, which means that if you submit within that 48 hours period, the assessment will be marked late but the late penalty will not be applied. After 48 hours, the accrued penalty will apply, i.e. a 15% deduction will be applied on Day 3 (after 48 hours), with an additional 5% per day after that (to day 7).

7 Brief Marking Scheme

Note on Report Length *The full report must not exceed 5 pages (excluding references and appendices, if any). Please note that writing more or longer is not necessarily better. The key is to present your work in a clear, concise, and well-structured manner, focusing on the most important content. Marks will be awarded for quality of explanation, analysis — not for volume.*

Component [Grades]	**Criterion**
Model [9]	• Implementation of three substantially different model structures (correctness, diversity, functionality). • Integration of attention mechanisms (at least one model with attention, quality/novelty of use). • Detailed method description (clear equations, proper notations, consistent terminology). • Model architecture diagrams (accurate, well-labelled, easy to understand). • Justification of design choices (why each structure/approach was chosen, grounded in reasoning or references).
Experiments and Results [13]	• Main performance comparison with clear and well-structured analysis of results. • Ablation studies implemented with proper presentation of results and thoughtful justification of the chosen experiments. • Qualitative results analysis, including extraction and visualisation of attention weights, with insightful interpretation linked to model behaviour and performance.
Other writings [5]	• Detailed dataset descriptions • Clear implementation details

	• Proper citation of related work, with meaningful discussion of existing methods and how they relate to your approach. • Tables and figures: well-organised, clearly labelled, and easy to interpret. • Insightful limitations and future works discussions
Impression Marks [3]	Impression marks will be given for extremly impressed works such as Novel design/High performance/Good design of the experiments (ablation studies)/very excellent architecture drawings etc.