CS 601.471/671 NLP: Self-supervised Models[*]
Final Project Description

1 Overview

The aim of the final project is to apply the knowledge acquired throughout the course and to explore a challenging problem related to the course's theme with a degree of autonomy. What constitutes "a challenging problem"? Typically, homework assignments are narrowly focused and well-defined, which is necessary given the limited time available to address a specific issue. However, in real-world scenarios (such as industrial projects), challenges are often more open-ended. You generally have a broad understanding of the problem's scope, requiring **you** to make decisions on how to define the problem's boundaries. Making these decisions involves becoming adept at navigating unstructured environments. This project is designed to provide you with the opportunity to experience such a partially open-ended setting. **What is the anticipated level of effort for the project?** Final project effort roughly equals (4-6) × the effort for a homework assignment. The final project constitutes 20% of the total grade.

2 Topics and Tracks

There are two tracks to choose from:

- **Default Project:** This track is tailored for participants who prefer a clear and structured challenge. The objective is to develop the most effective next-token predictor (i.e., model with the lowest perplexity) within a compute-constrained" environment. Here, compute-constrained" means being restricted to GPUs with limited VRAM—specifically, the 20GB MIG GPUs you've had access to during the course. We provide a development set (2k sentences) and a public test set (2k sentences) for you to iterate on. The development set and test set can be found at this link. In the end, we will evaluate the perplexity based on a hidden test set. We will run this code on a MIG GPU to evaluate your model.

  Your deliverables include **a checkpoint due on the morning 9am of the final report deadline**. Your report should describe the different strategies you have already (plan to, for proposal), explored (e.g., architectural tweaks or data preprocessing), and which approaches yielded the best results. Refer to the Transformer LM slides for ideas on architectural or data-centric changes relevant to LM pre-training. You can also invent your own set of modifications if you'd like to explore novel changes to Transformer or *any other architecture* that you'd like to try. The final report should document your methods, present results (perplexity and any additional performance metrics), and provide comparisons across approaches. For each method, include the training curve to illustrate learning progress. There are no constraints on the dataset or training time (note: MIG GPUs may timeout, but checkpoints can be saved and resumed). Final evaluation: **On the deadline day, we will evaluate all submitted checkpoints using our private test set and rank models based on performance.** A *grand* prize awaits the model achieving the lowest perplexity!

- **Custom Project:** If you are feeling adventurous and wish to delve into a problem that excites you, this track offers an open-ended topic. This project could involve demonstrating systemic limitations of previous work, identifying weaknesses, suggesting improvements to existing methods, or even creating something entirely new. The only requirement is that it must significantly involve human language. If you want to improve upon a paper, you think about improving it with a critical perspective. (What are the main novel contributions or points? Is the key to its success something general and reusable, or is it a special case? Are there any flaws or interesting details in their approach?) A key issue here making sure that you have a coherent and feasible plan that fit within your limited time (do you have the necessary data? do you have a viable method to evaluate your work? Are there suitable baselines or proposed ablation studies for comparison?) A *grand* prize awaits the project with the most innovative focus or comprehensive analysis!

3 Milestones

Each group is required to submit a document for each of the final project milestones: (1) a project proposal, (2) a midway project report, (3) a final report, and (4) a final project poster that summarizes the technical aspects of the project. Please refer to the course calendar for submission deadlines.

[*] https://self-supervised.cs.jhu.edu/sp2025/

4 Submission and Template

All documents, including the proposal, midway report, and final report, should be formatted using this template.

Important: The title of your project report must begin with "Default project" or "Custom project" to clearly indicate the track you are following.

5 Project Proposals

The project proposal should be a 2-page document outlining your intended project. This document serves as your opportunity to communicate your project plans. We expect the following elements to be included:

- Motivation: What approach are you taking to address this problem, and why?
- Related Work: What other related works have you encountered in this area? How do they relate to the problem you are investigating?
- Hypothesis: What hypothesis motivates the problem/solution you are exploring?
- Datasets: Specifically, which datasets do you plan to use for your experiments?
- Methods: Specifically, what approaches do you plan to use or compare against in your experiments?
- Experiments: Specifically, what experiments do you plan to conduct?
- Expected Outcome: What do you hope to conclude, and why?

If these details are missing, you will not receive credit for the proposal, and you will be asked to revise it. Here are examples of project proposals from previous years:

- Adversarial Prompting of Unlearned Language Models
- Ensemble Domain-Specific Knowledge Distillation

6 Midway Progress Reports

The midway progress report, which should be no more than 5 pages and use the same template as the proposals, should detail the progress made so far and outline the remaining tasks. This report should include a description of the progress achieved, experiments conducted, preliminary results obtained, and your plan for the remaining time. Although termed “midway,” this milestone should reflect more than half of the project’s completion. By this point, you are expected to have implemented *something* and have some experimental results to present. The report should build upon your initial proposal, maintaining a similar structure but with more depth, including new results and improved writing. This is an opportunity to refine your writing. Additionally, the report should outline your plan for the remaining project time.

7 Final Report

Students are expected to write code, conduct additional experiments, and document the results in a standard conference paper format, with a maximum of 8 pages (using the same template as the proposals). References are not included in the page limit. Note that longer reports are not necessarily better. Groups must include a “Contributions” section that clearly lists each author’s contributions (see Section 8 of this paper, for example).

The quality of your writing is crucial for your grade! The final report should succinctly summarize your findings and address the following questions:

- Motivation: Why is this problem important? Provide a rationale.
- Related Work: What other related works have you encountered in this area? How do they relate to the problem you are investigating?
- Hypothesis: What hypothesis motivates the problem/solution you are exploring?

- Datasets: Specifically, which datasets did you use?
- Methods: Specifically, what approaches did you use or compare against in your experiments?
- Evaluation: How did you assess the performance of the approach(es) you investigated?
- Findings: What worked, what did not, and why? Discuss the significance of these findings.
- Conclusion and Future Work

Here are examples of final reports from the previous year:

- Adversarial Prompting of Unlearned Language Models
- Ensemble Domain-Specific Knowledge Distillation
- Efficient Distillation of Transformers via Self-Teaching

8 Final Poster Presentations

All students are required to present their findings during a poster session held in the final exam period.

9 Project Grading

The aim of the project is to showcase the group's understanding of the tools and challenges associated with using self-supervised models. Grading will be based on the quality of the approach, the thoroughness of the evaluation, and the reasoning behind successes and failures. Additional grading criteria include the project's completeness, the clarity of the write-up, the complexity and difficulty of the approach, and the ability to justify the choices made. The grading breakdown is on the course page.

10 Group Work

Students are highly encouraged to work in groups for the final project, with team sizes limited to a maximum of 4 people. Working in a team is recommended. Larger team projects or projects used for multiple classes should be broader and involve exploring more models, tasks, or analyses.

For multi-person teams, include a brief statement detailing each teammate's contributions. Generally, all team members receive the same score, but we reserve the right to differentiate in cases where one or more members do not adequately participate in the team effort.

11 Use of Existing Code

Using existing code and resources is acceptable and encouraged. Avoid reinventing the wheel. You *must document* any external code used and give appropriate credit. Your grade will be based on your *value-added* contributions. You may use any language or framework for your project, though most are expected to use PyTorch.

12 Mentorship

The course staff are available to assist you with any aspect of your projects. Attend office hours for support. If office hours do not fit your schedule, it is your responsibility to coordinate alternative times. Ideally, I (Daniel) would like to meet with each team, but my time is limited, and there is only one of me!

13 Computational Resources

All teams are highly encouraged to utilize the GPU cluster that was provided for homework assignments throughout the course of the project.