Throughout the semester-long project, we aim to give you a taste of the full pipeline of NLP research. NLP research pipelines include problem formulation, literature surveys, data annotations, model replication, experiments, and analysis, as well as report writing. Additionally, you will learn how to collaborate with your teammates and make progress on your research project. The mentors will be assigned to each team, so you will have the opportunity to collaborate and discuss with other NLP researchers including DK and TAs. You can make your project an extension of your previous homework, but the project must have novel extensions and research contributions beyond your homework.

The following document outlines instructions for your class projects, including the types of contributions, timeline and dues, types of project topics, and evaluation criteria.

1 Types of Project Contributions

Novel contributions to your project can be made in a variety of ways. The problem definition section of your proposal and final report should clearly state what types of contributions your project will make. The following are possible types of contributions you could make along with example papers:

- Critical analysis of existing model/dataset (default project), e.g., [NRS+18, KL18, RKR20]
- New benchmark results and findings judged suitable for acceptance to an NLP or ML workshop,
- Collection of your own dataset on new tasks, (complex social) problems [EZM+21] or adversarial datasets [PWGK21] that can fool the existing systems,
- An in-depth literature survey on emerging topics [FGW+21, ZKK23],
- Interactive demonstration (e.g., Chrome Extension, Flask) [DKR+22, KMWK23] or visualization of existing systems [WTW+19],
- Applying NLP tools to your own domain of research (e.g., psychology [Kos23, Ull23], law [CHMS23], education, robotics [ABB+22]),
- New open-source repository or dataset with a high impact on the community
- Others (consult your mentors as soon as possible if you wish to do other types of projects).

2 Project Deliverables and Due Dates

Your project takes up 30% of your class grade. Every group member (maximum of 4 people) should submit their report, link to code (or a zipped code), and presentation slides/poster on Canvas before the deadline. Below is the list of your deliverables with due dates and link to Canvas submission:

- In-class proposal pitch (Oct 26, 31) and proposal report (5 points, due: November 3) (canvas)
- Midterm office hour participation (5 points, due: November 21) (canvas)
- Poster presentation (5 points, due: November 30 / December 5) (canvas)
- Final report (15 points, due: December 8) (canvas)

Note that the project deliverables submitted late after all late days of your teammates have been used will receive no credit. For each deliverable, please read the specific instructions carefully and the evaluation criteria.

2.1 In-class proposal pitch and proposal report

You and your team need to decide which topic of the project you will contribute to in this step. When choosing your topic, please refer to general project tips in Section 3 and past project samples in Section 4.

In-class Proposal Pitch. Before submitting the proposal, your team needs to give a short presentation of your proposal idea. Every member of your team should present and give a 3-minutes pitch of your proposal idea. The in-class proposal pitch is scheduled on October 26 and 31. You need to follow the example template and create a slide for your own project in the slide deck (Slides for October 26 and Slides for October 31). Please think about your next steps regarding your project and some questions which we can discuss about your project during the proposal pitch.

While the proposal pitch is not counted as part of your grade, it is a good opportunity to get helpful feedback from both instructors and classmates. Your proposal should clearly address the comments and feedback.

Requirement. Your project proposal should not exceed three pages. For this report, you must use the unofficial ACL-style template (link). In case you haven't used LaTex for scientific writing, this is a great opportunity to learn how scientists and researchers write their manuscripts using this typesetting tool called LaTex. Here is a tutorial for LaTex with Overleaf. Please upload your PDF report on Canvas before the deadline.

Evaluation. The following items should be included in your proposal:

```
Rubrik (10 points) for Proposal Report:

(1 point) Title, Team name, members, and role assignment
(1 point) Motivation
(2 point) Literature survey (at least three relevant papers)
(2 point) Problem definition
(2 point) Proposed idea and hypothesis
(1 point) Broader impact
(1 point) How to address feedback from proposal pitch presentation
```

You will be assigned a project mentor with feedback on your report within two weeks of submitting your proposal.

2.2 Midterm office hour participation

Requirement. Your team must schedule an office hour meeting with your assigned mentor (15 to 20 minutes) before the due date (November 21) to discuss your intermediate results and progress. The mentor expects you to give an update on your progress, ask questions, and consult with your plan until the final presentation and report. Below is a rubric for midterm evaluation:

```
Rubrik (5 points) for Midterm Office Hour Participation:

(1 point) Recap: Motivation, Literature Survey, Problem definition (based on
↪  your proposal)
(1 point) Detailed proposed ideas
(1 point) Novel contribution compared to prior work
(1 point) Preliminary results and comparison to the baseline performance
↪  (e.g., experimental results, findings, visualization)
(1 point) Plan until the end of the semester
```

You have to summarize the feedback you receive from your mentor and explain how you plan to address it. By the deadline, your response must be uploaded on Canvas.

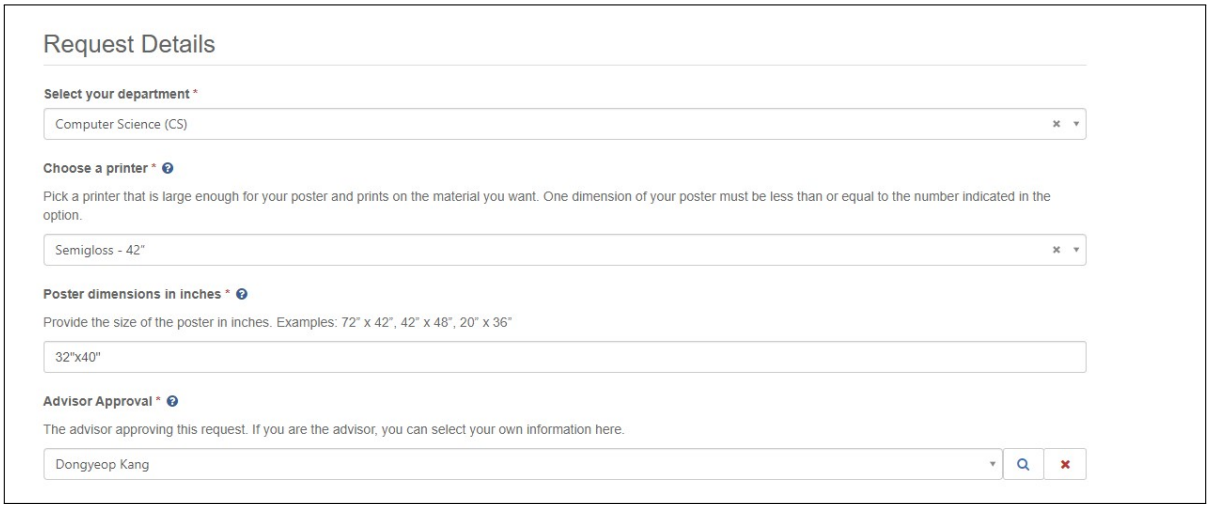
2.3 Poster presentation

Requirement. Everyone on your team should present their work at your assigned poster session on November 30 (Group A) or December 5 (Group B). Your group assignment will be provided a week in advance. You should stand by your poster during class (11:15am to 12:30pm) and communicate with instructors and other students (and possibly external guests). A PDF or PowerPoint file of your poster should be submitted to Canvas before the deadline (November 29)

Your poster clearly describes the following information: motivation, literature survey, problem definition, proposed ideas, contribution, experimental results, main findings, and plan for the final report. In order to facilitate interactive discussion, you should add conceptual figures to describe your proposed ideas, tables with your experiment results, and visualizations of your main findings. Please refer to past project posters in Section 4.

Location. Your poster presentation will be held in Shepherd 164 (aka Drone Lab). Around ten aisles will be set up for you to attach your poster for each poster session. Bring your printed poster and attach it to your designated aisle 10 minutes before the session starts.

Poster Printing. You have to print your poster by yourself, and the poster size should be 32" x 40". Please carefully read the printing instructions and request poster printing through the form. Remember, CSE-printing guarantees posters submitted 2 business days in advance, but **does not** work on the weekends. Some details on how to fill out the initial fields in the request form are given below.



Evaluation. Both instructors and your peers (i.e., classmates) will review your poster presentation. For peer group review points, every group is assigned a random team on their session day to review based on a rubric provided by instructors. You can optionally (no points for this) vote for the best poster. The team winning **best poster** will be given extra credit as a reward. Audiences will evaluate your poster based on the following questions:

```
Questions for Poster Evaluation:

Q1. How novel is the idea of the poster?
Q2. Does their progress show enough evidence for their project?
Q3. Do their plans for the rest of the semester seem feasible?
```

2.4 Final report

Requirement. There should be a maximum of eight pages in the final report with unlimited references and appendices. You can submit a PDF report and zipped code (or link to your github) on Canvas.

Evaluation. We are not evaluating if you succeeded or failed at accomplishing what you initially proposed to do. It is totally fine if your results are negative and not significant. The essential criteria we evaluate for your project are whether you put in a reasonable effort, deeply understand the nature and challenges of your problem, set a reasonable hypothesis to tackle the challenges, and can clearly communicate them with others. For detailed assessment guidelines, please see below or visit this link:

```
Rubrik (100 points) for Final Report

Introduction / Background / Motivation:
(5 points) What problem do you try to solve? Describe your objectives clearly
↪  without using any technical jargon.
(5 points) How is it done today by other researchers? What are the limitations
↪  and challenges of current practice?
(5 points) Who might be interested in your work? What kinds of impact can you
↪  make?

Approach:
(10 points) What did you do exactly? How did you solve the problem? Why did
↪  you think it would be successful? What is your hypothesis? What is the
↪  scientific novel of your approach?
(5 points) What challenges did you anticipate and/or encounter during the
↪  development of your approach? Did the very first thing you tried work?

Experiments / Results / Error Analysis:
(10 points) How did you measure success? What research questions do you want
↪  to validate? What evaluation metrics and experiments were used? What were
↪  the results, both quantitative and qualitative? Did you succeed? Did you
↪  fail?
(10 points) No matter you succeed or fail, why? Which data points are
↪  incorrectly predicted by yours but previous models can't, or vice versa?
(10 points) Are there still some failure cases? Why can't your approach
↪  address them? Any potential solutions?
(5 points) Are the ideas/problems/results presented with appropriate
↪  illustrations?

Discussion points:
(5 points) Replicability: How easily are your results able to be reproduced by
↪  others?
(10 points) Datasets: Did your dataset or annotation affect other people's
↪  choice of research or development projects to undertake?
(10 points) Ethics: Does your work have potential harm or risk to our society?
↪  What kinds? If so, how can you address them?
(10 points) Discussion: What limitations does your model have? How can you
↪  extend your work for future research?
```

```
At least some of these questions and others should be relevant to your project
↪  and should be addressed in the final report. You do not need to address
↪  all of them in full detail. Some may be irrelevant to your project and
↪  others may be standard and thus require only a brief mention.
```

3 General advice for successful projects

Don't be too ambitious. Please keep in mind that you only have a few months to work on this project, while you are taking other classes. I don't expect you to produce a publishable research

outcome. Again, we are aiming to ensure you have gone through the entire NLP research pipeline, from data selection/creation to model replication/building to experiments and error analysis, but not make a publishable research outcome. You will receive a good grade if your final report shows that you have done a reasonable amount of work on the problem you set along the way.

Use existing tools. As part of your homework, you will learn various tools and open-sourced libraries. Instead of implementing something from scratch, use existing tools for your projects. You might consider using the following tools for your project:

- Tutorials on PyTorch and HuggingFace programming
- Huggingface model and data cards
- The state-of-the-art models in different NLP tasks/datasets in PapersWithCode leaderboard
- Wandb and tensorboard for tracking your training
- Demonstrate your tool using Gradio
- Other python libraries such as transformers, diffusers, timm, datasets, safetensors, accelerate, optimum, tokenizers, evaluate, simulate, and more.
- ChatGPT, GPT3 Playground, and other LLMs inference

Communicate with your teammates and mentors on Slack. We will create a private Slack channel with you once you are assigned to each mentor. Communicate frequently with your team members and mentors via Slack. The more frequent updates you provide and the more active discussions you have with your mentor in Slack, the better able we are to support you and evaluate your contributions.

Secure computing resources. In order to conduct your experiments, you will need some computing resources. Ensure that your department has GPU computing resources, such as MSI. Additionally, you can use publicly available resources such as Google Cloud/CoLab and Amazon AWS.

The DOs for successful projects

- Clearly divide work between team members for optimal collaboration process
- Start early and work on it regularly every week rather than rush at the end
- Set up workflow – download data, verify data, set up base code on github, communicate via Slack, sharing results on Google spreadsheet, etc
- Have a clear, well-defined hypothesis to be tested (+ novel/creative hypothesis)
- Conclusions and results should provide some insights
- Meaningful tables and plots to display the key results
  - nice visualizations or interactive demos
  - novel/impressive engineering feat
  - good empirical results in both qualitative and quantitative ways.

The Don'ts for successful projects

- Data not available or hard to get access to, which stalls progress
- All experiments run with prepackaged source – no extra code written for model/data processing
- Team starts late – only draft of code up before dues
- Just ran the model once or twice on the data and reported results (not much hyper-parameter tuning and statistical significance test)
- Only a few standard graphs (loss curves, accuracy) without any analysis
- Results/Conclusion don't say much besides that it didn't work. Negative results are fine, but only with in-depth analysis and justification

4 Example Project Topics and Past Projects

The following are examples of project topics:

- Experiment with improving an architecture on a well-defined NLP task
- Case study: apply an architecture to a dataset in the real world that has not been done before
- Compete in a predefined competition (SemEval 20XX, Kaggle, etc)
- Stress test on comparison study of known models/architecture (e.g., when are LSTMs better than Transformers for task XYZ?)
- Design a novel NN layer, objective function, etc on NLP tasks
- Multi-domain/Multi-lingual NLP (RL+NLP, CV+NLP, Social Science + NLP)
- Visualization/Interpretability/controllability study of NN models
- In-depth error analysis on XYZ datasets using the state-of-the-art model
- Collection of adversarial datasets for XYZ tasks
- Human evaluation on current NLG evaluation metrics
- Collect a new dataset of interesting language variation or cognition
- Capacity of latest large language models (e.g., GPT3, chatGPT) on existing or new tasks

You can access the final reports and poster slides for previous NLP classes:

- Simulating Everyone's Voice: Exploring ChatGPTs Ability to Simulate Human Annotators, CSCI 5541 S23 (final report, poster)
- Vision & Language-guided Generalized Object Grasping, CSCI 5541 S23 (final report, poster)
- Who is speaking? Discriminating Artificial and Human-Generated Text with A Natural Language Processing Approach, CSCI 5541 S23 (final report, poster)
- Generalizability of FLAN-T5 Model Using Composite Task Prompting, CSCI 5541 S23 (final report, poster)
- Comparing the Effectiveness of Fine-tuning vs. One-Shot Learning on the Kidz Bopification Task, CSCI 5541 S23 (final report, poster)
- Exploring Hallucination in LLMs: A Study of GPT-3.5 and GPT-4 to Enhance Fact-Based Results, CSCI 5541 S23 (final report, poster)
- Understanding Narrative Transportation in Fantasy Fanfiction, CSCI 8980 S22 (final report)
- Exploring Episodic Memory through Cross-modal representations, CSCI 8980 S22 (final report)
- Generating Controllable Long-dialogue with Coherence, CSCI 5980 F22 (final report)

References

[ABB+22] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[CHMS23] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

[DKR+22] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, 2022.

[EZM+21] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[FGW+21] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August 2021. Association for Computational Linguistics.

[KL18] Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[KMWK23] Ryan Koo, Anna Martin, Linghe Wang, and Dongyeop Kang. Decoding the end-to-end writing trajectory in scholarly manuscripts, 2023.

[Kos23] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models, 2023.

[NRS+18] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[PWGK21] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online, August 2021. Association for Computational Linguistics.

[RKR20] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[Ull23] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[WTW+19] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. AllenNLP Interpret: A framework for explaining predictions of NLP models. In *Empirical Methods in Natural Language Processing*, 2019.

[ZKK23] Hao Zou, Zae Myung Kim, and Dongyeop Kang. A survey of diffusion models in natural language processing, 2023.