Throughout the semester-long project, we aim to give you a taste of the full pipeline of NLP research, including problem formulation, literature surveys, data annotations, model replication, experiments, and analysis, as well as paper writing and presentation. Additionally, you will learn how to collaborate with your teammates and make regular progress on your research project. The mentors will be assigned to each team after you submit your team formation and brainstorming ideas, so you will have the opportunity to collaborate and discuss with other NLP researchers including DK and TAs. Please carefully read the following document that outlines instructions for your class projects, including the types of contributions, timeline and dues, types of project topics, and evaluation criteria.

Note: Please note that homeworks are generously graded but the projects are not. Therefore, students should consider the potential contribution of the projects rather than trying to play it safe. Playing it safe wont give them full marks.

1 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/webpages on Canvas before the deadline. Below is the list of your deliverables by due dates and link to Canvas submission:

- §1.1 Team formation (1 point, due: Sep 19) (canvas)
- §1.2 Project Brainstorming (1 point, due: Oct 1) (canvas)
- §1.3 In-class proposal pitch (3 points, Oct 8 and 10) (Slide deck for Group A and Group B)
- §1.4 Proposal report (5 points, due: Oct 13) (canvas)
- §1.5 Midterm office hour participation (5 points, due: Oct 31) (canvas)
- §1.6 Poster presentation (5 points, due: Dec 3) (canvas)
- §1.7 Final report (10 points, due: Dec 12) (canvas)

The late days and penalty will be applied to all team members for project deliverables. For each deliverable, please carefully read the specific instructions and the evaluation criteria below.

1.1 Team formation

Requirement. You have to form a team of two to four for the project and discuss potential ideas for your project. You need to submit your team name and member information directly to Canvas.

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

```
Rubric (1 point) for team submission :
(0.5 point) Team name
(0.5 point) Member names
```

1.2 Project brainstorming

Requirement. Your team should start brainstorming of potential ideas for your project, and submit a list of project ideas, titles, and plans (i.e., a couple of sentences for each idea) directly to Canvas. If you have multiple ideas, it is okay to submit them all up to 3 ideas. When choosing your topic, please refer to general project tips in Section 2 and past project samples in Section 3.

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

```
Rubric (1 point) Brainstorming ideas:
(0.5 point) Potential project titles and ideas
(0.5 point) Clear description of the ideas and execution plan
```

You will be assigned a project mentor with feedback within one week of submitting your ideas.

1.3 In-class proposal pitch

Requirement. Before submitting your proposal, your team needs to give a short presentation of your final proposal idea. You have to develop more concrete formulation of your problem statement and plans after getting feedback from your mentors. Every member of your team should present in person or virtually (only for UNITE/remote students) (umn.zoom.us/my/dongyeop), and give a 3-minute pitch of your proposal idea and 2-minutes QA. The in-class proposal pitch is scheduled on Oct 8 and 10. Your group assignment (group A or B) will be provided a week in advance. You need to follow the example template and create a slide for your own project in the slide deck (Slide deck for Group A and Group B). During the discussion session, you will discuss what datasets/models you intend to use, what next steps to take, and some questions about your project. This is a good opportunity to get helpful feedback from both instructors and classmates on your ideas. **Your project proposal to be done in the next step should clearly address the comments and feedback you received during the pitch.**

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

```
Rubric (3 points) for Proposal Pitch:
(1 point) Clear formulation and definition of your problem
(1 point) Specific execution plan (e.g., datasets, models, systems)
(1 point) Preliminary results if possible and questions for audience
```

1.4 Proposal report

Requirement. After project pitch, your team needs to address feedback from the pitch and write a formal proposal. Your project proposal should not exceed *three pages* excluding references. For the project-related reports including the proposal and final report, you must use this LaTex 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. For literature survey and planning, please refer to general project tips in Section 2 and past project samples in Section 3.

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

```
Rubric (5 points) for Proposal Report:
(0.5 point) Title, team name, members, and role assignment
(1 point) Clear Motivation and Problem definition
(1 point) In-depth Literature survey (at least three relevant and latest
↪  papers)
(2 point) ``Novel'' proposed idea and your execution plan (novelty: compare to
↪  the state of the art methods/systems/datasets, how novel is your
↪  approach?)
(0.5 point) Plan to address feedback from the pitch presentation
```

1.5 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 and discuss your intermediate results and progress. You have to create a project webpage (template: https://github.com/minnesotanlp/csci5541-project-template) and show them during the discussion with your mentor.

The mentor expects you to give an update on your progress results, ask questions, and consult with your plan until the final presentation. After the meeting, you have to report what intermediate progress you made, summarize the feedback you receive from your mentor and explain how you plan to address it. Your response must be uploaded directly to Canvas.

Evaluation. Below is a rubric for midterm evaluation:

```
Rubric (5 points) for Midterm Office Hour Participation:
(1 point) Additional development of your ideas after the proposal
(1 point) Submission of the project webpage
(2 points) Preliminary results and comparison to the baseline performance
↪  (e.g., experimental results, findings, visualization)
(1 point) Plan to address the mentor's feedback and plan until the end of the
↪  semester
```

1.6 Poster presentation

Requirement. Everyone on your team should present their work at your assigned poster session. The order of presentation from the two groups will be swapped from the pitch to be fair. You should stand by your poster during class and communicate with instructors and other students (and possibly external guests). A PDF file of your poster should be submitted to Canvas before the deadline, so UNITE or other remote students can access your posters virtually.

Your poster clearly describes the following information:

```
Motivation
Literature survey
Problem definition
Proposed ideas
Contribution
Experimental results and comparison with baselines
Main findings
Limitation and discussion
Plan for the final report.
```

In order to facilitate effective discussion, I encourage you to add figures, tables, examples, and other **visualizations** of your main findings. Also, if you have a working system, you are welcome to bring your laptop for a demonstration. Please refer to past project posters in Section 3.

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. Instructors will use the similar evaluation criteria used in the final report link except for the completeness of your work.

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 will 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?
```

1.7 Final report

Requirement. You have to submit three deliberables to Canvas for the final project submission:

- **Report**: Your final report should a PDF file with a maximum of eight pages and unlimited pages for references and appendices.
- **Code**: A link to your public github (preferred) or zipped code
- **Webpage**: A zipped code of your project webpage with the final results updated (all project pages will be displayed on the virtual project wall at the end)

Evaluation. 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, significanlty validate them through your experiments, and can clearly communicate them with others. Overall, the three important rubrics are:

- Novelty: Compared to the state-of-the-art methods/systems/datasets, how novel
- Significance: How strong is your result? Is your finding still holding if different setups or prompting tricks?
- Clarity: How clear and easy-to-follow is your report? Do you have well organized presentation of your results and problem definition?

For detailed assessment guidelines, please see below or visit this link. We will normalize the final scores into 10 points.

```
Rubric (100 points) for Final Report

Below are three general evaluation criteria:
(10 points) Novelty: Compared to the state-of-the-art
↪  methods/systems/datasets, how novel is your approach? Is your work
↪  publishable?
(10 points) Significance: How storng is your result? Is your finding still
↪  holding if different setups or prompting tricks?
(10 points) Clarity: How clear and easy-to-follow is your report? Do you have
↪  well organized presetnation of your results and problem definition?
Introduction / Background / Motivation:

Introduction / Background / Motivation:
(5 points) What problem do you try to solve? Describe your objectives cleraly
↪  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:
(5 points) What did you do exactly? How did you solve the problem? Why did you
↪  think it would be successful? What is your hypothesis?
(5 points) What challenges did you anticipate and/or encounter during the
↪  development of your approach? Did the very first thing you tried work?
(5 points) What is scientific novel of your approach to address the
↪  challenges?
Experiments / Results / Error Analysis:

Experiments / Results / Error Analysis:
(5 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?
(5 points) No matter you succeed or fail, why? Which data points are
↪  incorrectly predicted by yours but previous models can't, or vice versa.
(5 points) Are there still some failure cases? Why can't your approach address
↪  them? Any potential solutions?
(5 points) Are the ideas/probelm/results presented with appropriate
↪  illustration?
Additional points:

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

2 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. You will receive a good grade (8/10 points) 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

Literature Survey The most important step in research is to "re-search" what other researchers have done in the past on your topics of interest. Then you need to clearly distinguish how your approach differs from previous work. You can find relevant work with some keyword searching on Google Scholar, Semantic Scholar, ACL anthology, CS arXiv, OpenReview, or general Google search. Please ask your mentors for help on which keywords to use or which seed papers to read. If you find a similar/relevant paper, check out the other papers that recently cited it using Google Scholar or Semantic Scholar. Do not read from the beginning to the end in order. You can grasp overall ideas from title, abstract, and tables/figures at first glance and decide to closely read or not.

Regularly communicate with your teammates and mentors on Slack. We will create two private Slack channels: one with all mentors, and the other only among your team members. In the channel with mentors, you can ask specific questions that you like to get feedback from your mentors. In the channel among your team members, you can set up regular meetings, share your progress, etc. Please 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. Note that we don't count your Slack participation for project as a part of class participation though.

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 Donts 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 due date
- 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 dont say much besides that it didnt work. Negative results are fine, but only with in-depth analysis and justification.

3 Project Types/Topics and Past Projects

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. 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 are possible types of contributions you could make along with example papers:

- Critical analysis of existing model/dataset, 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).

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. Some of early class projects are now published at top-tier workshops and conferences.

- 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)
- 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.