These are sketches of possible project ideas from NLP researchers at NYU and from other institutions that we collaborate with.

- We trust the people who proposed each idea, and we think the ideas look reasonable at first glance, but we haven't done a careful literature review ourselves for each one, and we can't guarantee that they'll be straightforward. Even if you pick one of these ideas, you'll still have to complete the mini-proposal requirement, and you'll still have to write a full proposal with your team.
- A few of these ideas are on topics that aren't part of the core of the course, like music data, but that still build on some of the ideas in the course. We're allowing teams to work on these more distant topics *if* they have an experienced mentor who can help.
- The name listed with each idea is the person who suggested it. They've agreed to meet with a team at least once or twice to help develop the ideas.
- It's okay for multiple teams to work on the same topic, but the mentors may not be interested in helping more than one team. If you're excited about a topic and want to work with the person who proposed it, it's best to reach out soon. If you find out that someone else already reached out, consider forming a team with them. However, that these ideas are available to both the graduate and undergraduate sections of the course, and we don't generally allow teams to be split across the two sections because of the different deadlines.
- These are ideas that active researchers are excited about, and many of them are fairly complex, and build on topics that will be new to you. If you take on one of these projects, you'll have to do some self-study outside of class to catch up on the background material. If you propose your own project idea, it doesn't have to be as arcane or complex as these ideas.
- Feel free to comment on this doc to compare notes with classmates, but be warned that (i) this doc is visible to both class sections, and (ii) this doc is *not* visible to most of the mentors..

Improving Human Reasoning with NLP
Ethan Perez (perez@nyu.edu)
The research startup Ought has compiled a list of research project ideas that are relevant to helping people think using machine learning (their startup's goal). This line of work is also connected with AI safety, as they described here. Some of the ideas revolve around using language models to help people answer questions or make forecasts/predictions (i.e., by providing an initial guess or by generating relevant subquestions or other considerations).

Error Analysis Analysis
Sam Bowman
Many modeling papers present rough breakdowns of the kinds of examples that models get wrong. Are these actually informative? Run a small user study among classmates: One group of users reads the error analysis from the paper, and the other doesn't. Both groups then have to look at some examples from the dataset, and try to guess which ones the model got wrong.

Does reading the error analysis make it easier for the human users to guess what errors models will make?

Roman Numeral Analysis of Bach Chorales from Generated Training Data
Cal Peyser (ccp5804@nyu.edu)
Roman numeral analysis is a longstanding and difficult task in music information processing. Recent work, including a paper I'm submitting this year to ISMIR, addresses this problem using NLP techniques from a corpus of analyzed musical pieces, although these corpora are quite small. Separately, there are good heuristics for generating realistic pieces of music together with analysis labels. Distilling the knowledge from these heuristics into a neural model by training on generated data may provide a novel way forward.

Detecting Errors in Datasets with AUM
Sam Bowman
Use the recent AUM technique from computer vision to identify examples from existing public datasets that are likely to be mislabeled. Focus on public datasets like MNLI that have multiple crowdworker label guesses included in their public releases, and see how likely errors identified by AUM compare with examples that get relatively low agreement among crowdworker annotators. Alternately, try re-training models on training data that has been filtered using AUM, and seeing if performance improves.

Crosslingual Transfer for Grammaticality Judgment
Sam Bowman
Crosslingual models like XLM-R let you train a model on data in one language, and use it for the same task in another language. This works surprisingly well for many NLU tasks. What about the task of grammaticality judgment—or deciding whether a string is a grammatical sentence? Try a fine-tuning experiment with datasets like BLiMP and CLiMP, or CoLA and a small hand-built dataset in your favorite language.

Predicting Translation Popularity
Stephen Mayhew (not available as a mentor)
In the STAPLE shared task, the goal was to produce a large number of possible translations for a given input sentence. As the ground-truth for that task, organizers provided many possible translations for input sentences as well as normalized popularity scores for each translation, calculated from real usage patterns in the Duolingo app. There are several interesting questions to ask: what governs these scores? Do language models alone explain them? Do learners prefer shorter sentences over longer? Are learners primed by the input sentence, and create translations that are most faithful to the input? Some input sentences have many more attested translations than others -- this is surely due to sentence length, but are there other factors?

Large Scale Bias Audit of Available NLP datasets
Yacine Jernite (yacine@huggingface.co)

Train a bias classifier on recently released dataset such as Social Bias Frames and Multi-Dimensional Gender Bias, run it at scale on the 300+ English language datasets on the HF hub, and analyze trends by task type and creation type (e.g. found vs machine-generated vs crowdsource). This will provide a much needed insights into the contribution of the datasets to the trained models' harmful biases and a fantastic stepping stone to start alleviating this issue.

Continually Evaluating Leaderboards
Nazneen Rajani (nazneen.rajani@salesforce.com)
Robustness Gym (https://robustnessgym.com/) proposes a shift from static to continual evaluation as a way to better evaluate NLP models and understand robustness issues. The project would be to adopt the continual evaluation paradigm and create an extension to the existing GLUE/SuperGLUE leaderboards with this additional feature. RG has the TestBench abstraction (think of them as Datasets in HF) and so the task would be to run the existing models on the TestBench in RG and generate new scores for the leaderboard.

Generalizable Domain Adaptation for Question Answering
Sara Rosenthal (sjrosenthal@us.ibm.com), Avi Sil (avi@us.ibm.com); IBM Research AI
Question Answering is an important topic that has been explored in many areas such as news, tech, finance, etc… A labor intensive data process occurs when data is needed in a new domain (eg. recently COVID-19). Domain adaptation is a useful technique for learning a model on new, target, data with minimal annotations using a model trained on another existing, source, dataset. However, domain adaptation often comes at a cost. The model adapted to the new domain suffers a loss in performance in the original source domain. Can we build a generalizable model that uses a smaller amount of data while maintaining original and strong performance in all domains?

Benchmarking Different Data Augmentation Strategies for Text Classification
Nickil Maveli (n.maveli@sms.ed.ac.uk**)**
Contrastive Learning has proven to be successful in modeling computer vision tasks lately. The main idea would be to use existing data augmentation coupled with probing strategies at sentence level and run it on the GLUE tasks. It could later be expanded to know some stats around the extent to which each combination of these augmentation strategies contribute positively/negatively to the model's performance and when the prediction starts to degrade across tasks. Benchmarking the optimum size of each augmentation that results in the best score would be a further addition.

Cross-Lingual Fact Verification
Tal Schuster (tals@csail.mit.edu)
Fact verification involved predicting the veracity of a given claim by retrieving relevant information from external sources and assessing the relation between the retrieved evidence and the examined claim--whether it supports or refutes it, or doesn't hold enough information. Most current datasets are like FEVER and others are evaluating English written claims against English information sources (e.g., English Wikipedia). A desirable extension could be to

evaluate claims in many languages against evidence from English Wikipedia. To explore this direction we can synthetically or manually create translations to claims and evaluate crosslingual models such as BERT, XLMR, alignment methods etc.

SiFT for Improved Generalization
Will Huang (will.huang@nyu.edu)
In the recent DeBERTa paper, the authors propose the use of a virtual adversarial training algorithm: Scale-invariant-Fine-Tuning (SiFT). By adding perturbations to word embeddings, the algorithm improves a model's robustness to adversarial examples while fine-tuning for downstream tasks. Further, this method should improve model generalization. The paper leaves a comprehensive study of this method for future work. Studying this method, potentially on testbeds like MRQA, to understand when and how SiFT improves generalization could inform future development of fine-tuning methods.

Syntactic analogies in contextual embeddings
Alex Warstadt (warstadt@nyu.edu)
Analogies have been studied extensively to better understand the geometry of word embedding spaces (remember $v_{king} - v_{man} + v_{woman} \approx v_{queen}$?), but the geometry of sentence embeddings and contextual word embeddings has been far less studied in this manner (see this exception). With contextual embeddings we can study *syntactic analogies* between sentences related by a particular syntactic transformation such as the passive transformation: v(I fixed the car) - v(The car was fixed) + v(You washed the dishes yesterday) ≈ v(The dishes were washed yesterday). In addition to building a dataset of examples, you will have to figure out how to measure performance, since there is not a finite vocabulary of sentences.

Social Bias in Language Models Pretrained on Scientific Corpora
Nikita Nangia (nikitanangia@nyu.edu)
Do models like BioBERT and SciBERT exhibit less social bias than their regular pretrained LM counterparts? And if so, how does their performance compare on standard NLU tasks compare? Test models pretrained only on scientific corpora, or similar non-colloquial corpora, on social bias detection tasks and datasets like CrowS-Pairs, StereoSet, WinoGender, and WinoBias. Also measure the models' perplexity on a standard LM benchmark dataset, WSJ. Lastly, compare model performance on common NLU benchmarks like GLUE and SuperGLUE.

Document Understanding and Information Retrieval
Siyuan Xiang (siyuan.xiang@hyperscience.com)
NLP techniques are most often applied to one-dimensional, text-only input. However, many real-world applications concern "documents" where the 2D structure itself has a significant importance. This is an active field of research and several leading companies proposed approaches ranging from LayoutLM (Microsoft) to CharGrid (SAP) or Attend, Copy, Parse (Tradeshift). However, these techniques typically require large scale datasets for training, whereas most enterprise document understanding models need to be learned in a low-data regime, typically leveraging semi-supervised training, which makes it a unique NLP challenge.

A good way to evaluate document understanding capabilities is the information retrieval (IR) task: given a list of fields (i.e. keys), one seeks to extract corresponding values from a document, if present. We propose the SROIE dataset for this task: it is composed of a wide range of receipts where we seek to identify several pieces of information -- total amount, merchant address, etc.