Natural Language Processing **Tel Aviv University**

Final Project: Instructions & Ideas

Due Date: *TBD* Lecturer: Maor Ivgi

Contents

1 Goals

- The goal of the research project is to conduct independent research related to topics discussed in class. You will pick a paper (or a couple of papers) of your choice and focus on some of its aspects: either a possible extension, a related or modified model, a novel evaluation method, etc. Alternatively, you can pick a topic you are interested in, and expand the scientific knowledge about it.

- While it would be great to come up with new results (in which case we will consider pursuing publication in an upcoming NLP conference), to get full credit it is enough to describe what you wanted to achieve, how you tried to achieve it, and what were the challenges you faced, as well as how it fits into the existing knowledge around this topic.

- The most important goal: experience research and the joy of new discoveries!

2 Project Instructions

2.1 Proposal submission

(a) Choose your partner(s): work in groups of up to four students. Working in large teams may involve some logistical efforts, but it is often worth the tradeoff, as more teammates mean more working hands and more heads to brainstorm with.

(b) Pick one to three (but no more than three!) papers related to the course topics, that were published in a related conferences (EMNLP, *ACL, ICLR, NeurIPS) ideally between 2019–2024 (inclusive). Other papers/venues can be considered but need my approval.

(c) Submit a proposal as described below for your project by **August 16th, 2024** and get a binary approval from me.

Important - while the proposal does not affect your final score, it is not a recommendation and subject to the same late-penalties as all tasks (affecting your final project grading).

Proposal The project proposal should consist of up to a single page with the following information:

1. Project title
2. Names and e-mails of the contributing students
3. A brief description of the proposed project: what is the main topic, what is already known and what you are hoping to achieve. It should be clear from reading the proposal what you intend to do and roughly how, and how you will evaluate your success. The main goal for me is to understand that the proposal is sufficiently—but also not overly—ambitious.
4. Make sure to describe your assumptions and requirements for the project (e.g. access to compute, access to api resources, access to specific data, human evaluators model weights etc.) and how you plan to attain them.
5. During the project, you will have access to the GPU cluster of Tel-Aviv Data center. If you have any specific needs that I may be able to accommodate (such as access to cloud API endpoints), please include them as well and we will see if it is possible.
6. If you work on a default project idea (see §3), you can simply submit that you will work on that, or extend the description if you have any additional thoughts about it.

2.2 Computational resources

All students will have access to the SLURM cluster using the partition *studentkillable* (see usage instructions in `https://www.cs.tau.ac.il/system/slurm`), and can run jobs on the GPUs there. Please read/watch through the tips I give in the last lecture and follow some basic tutorials before you start experimenting with slurm, it will save you a lot of time.

In addition, you will have access to dedicated storage where you can host your models and datasets. This storage is persistent (though not backed-up so make sure you back up all of your code and intermediate critical results on GitHub/google drive) and will be available for you you under

`/home/joberant/NLP_2324B/<your_user_name>` anywhere on the system machines, starting from the end of the semester and until the project submission due date. **Make sure you use it for data, cache, virtual env, default directory for HuggingFace etc.** See my tips in the wrap-up slides and check out the tutorials I linked to. You can also consider linking e.g. `~/.cache` to a directory there with `ln -s /home/joberant/NLP_2324B/<your_user_name>/.cache ~/.cache`'. Please avoid overusing it. For example, If you fine-tune a model, make sure you don't save a new checkpoint every 500 steps as it will quickly block the storage space.

For your work, in addition to the above, you may find you need some special resources such as OepnAI API credits or GPUs beyond what is available to you via slurm. If you do, please email me ASAP, detailing your needs with the following details:

1. Group names, IDs and emails
2. Project title
3. As much as possible, exact estimates of your needs . For example, if you need API access, include how much and to which models (you can use OpenAI online tokenizer and prcing to compute that). If you need GPU access, to which GPUs and how many hours to think you will need. Same goes for extra storage etc.
4. A short textual justification for this need, for me to pass on for approval.

These are not guaranteed to be fulfilled, but I will do my best. Always have a plan B, and design your research plan so that you do most of the work on the resources you already have, saving the limited resources to the final experiments as much as possible

2.3 Final submission

The final project will be handed in as a "paper" in ACL format; see:
https://www.overleaf.com/latex/templates/acl-2023-proceedings-template/qjdgcrdwcnwp

The report length is limited to 8 pages (not including references and appendix). It can be shorter, but make sure you cover all the relevant topics needed (see §4). If needed, after submission, I will schedule a short meeting with some groups to discuss their report.

3 Project ideas

3.1 General directions

There are many interesting questions to answer with the tools you have learned in class. Try to avoid ideas that require significant engineering (e.g. building an online platform), significant human efforts (e.g. large scale annotations) or significant compute (e.g. reproducing GPT-4). Instead, you can think of ways to improve the evaluation of existing methods, coming up with new interesting tasks (and working on a proof-of-concept for them), improving a model for a specific task or extending our knowledge into tools that are widely used.

While the grading guideline section might sound daunting, the scope of the project is a white-paper with a proof-of-concept presentation of an idea, and not a scientific paper ready to be submitted to an international conference. Choosing a good idea is key to make sure your research and implementation phase is both interesting, and can be completed in a few days of work, leaving you with meaningful results to be compiled into a white paper.

3.2 Default project ideas

Below, you can find some ideas I have gathered, which you can choose to work on. Before choosing any of them, please make sure you have a good understating how you will approach the subject, and that you have access to any required resources to complete what you plan.

Related papers prediction The idea for the research project is to use information from existing arxiv/semantic scholar papers to predict what papers should have been cited, in order to address the problem of identifying related work and ensuring all required citations have been made. This would involve using textual information from the introduction section and potentially network analysis. The goal is to improve the accuracy of citations, particularly in the context of the increasing number of papers being published in open source. In this project, you can choose to focus on predicting a local citation inline in a sentence, or look at the task globally, for example predicting probable related work given a paper abstract or introduction. You can also choose to suggest a novel approach and try to improve the results on the new CiteME benchmark.

Can frankenmodels really be a free-lunch? A recent trend arising from Reddit suggests that using *Frankenmodels* can enhance a pretrained LM's ability without introducing new parameters or incurring additional training costs. Namely, people found that repeating frozen layers and artificially increasing the model size is helpful. In this project, you will take the role of a proper academic researcher, investigating the efficacy of the model. You can analyze it across model scales, model architectures and families, benchmark types, or properties of the frankenmerges themselves (such as which layers to repeat and how many times to do so).

Building an LLM identifiability challenge As LLMs become more ubiquitous, the distinction between human-generated and machine-generated text becomes less clear. Even resources like OpenAI's classifier are no longer reliably distinguishing between them. In this project, you will design a pipeline to gather diverse human-generated texts in specific domains (e.g., tweets, Facebook posts, newspaper articles, and any other domain you find interesting). Then, your pipeline should demonstrate how to collect LLM-based texts mimicking the same diversity and domains as the natural ones. Finally, you will use these settings to set up a challenge to build the best classifier capable of handling as many domains, topics, diversities, and ranges of LLMs as possible. You are not expected to gather a large dataset as part of the project, but you do need to demonstrate a working proof-of-concept and clearly describe how your approach can be scaled up, along with the benefits of using it.

Detecting hallucinations from multiple-timesteps logits Recent work such as 1, 2 suggest cases to induce and detect hallucinations by LMs. While there are abundant of hallucinations detectors, this research project aims to take a different view to detect hallucinations. Namely, instead of treating hallucination as a local behavior that it detectable through the activations at the respective timestep, here you will treat it as a time-series prediction problem. Namely, given some input that causes an hallucination to occur at timestep t, you will take the top-k logits/probabilities from timesteps $t-l, \dots, t+l$ and try to train a classifier to predict which such 2d sequences contain an hallucination and which are truthful. The project would involve using a small dataset of questions with gold answers and comparing the model's activations for both gold and random answers to determine if it is correct or wrong. The goal is to expand this to detection in long outputs and to define what constitutes hallucination, such as distinguishing between a novel conclusion and something made up and then justified. If successful, such research can have great impact on deployed LLMs.

Training a strong Hebrew Sentence Encoder from a pretrained Decoder While recent years have brought many additions to the open-source set of pretrained LMs in high-resource languages such as English, most of these tools are not directly useful for use on Hebrew Inputs. Recently, a new project aiming to bridge this gap has introduced new tools and most importantly benchmarks for Herbrew LMs. Concurrently, some new open-source strong models have been trained on Hebrew text, most recently, the DictaLM 2.0. In this project, you will modify the DictaLM model to be a strong Encoder-model using the LLM2Vec method. To evaluate the result, you will train linear classifier for a Hebrew sentiment analysis task on top of embeddings from your trained model, and against some baselines. Such baselines can be strong English and multilingual pretrained models, and existing pretrained Hebrew encoders (for example, AlephBERT and AlephBERTGimmel).

Assessing LLM Susceptibility to Induced Doubt This research project aims to investigate the susceptibility of Language Models (LLMs) to induced doubt. The core idea is to explore whether repeated questioning and suggestions, such as "think again" or "are you sure," can influence LLMs to change their initially correct responses. Additionally, you will evaluate the impact of assertively stating incorrect answers to see if the models can be coerced into adopting those incorrect answers. This will involve designing a series of experiments where the models are subjected to various forms of induced doubt and analyzing their responses for changes in accuracy and consistency. The goal is to understand the robustness of LLMs to external influence and to identify strategies to improve their resilience against such manipulations.

Evaluating Expertise in Mixture of Experts Models This project focuses on analyzing the performance and expertise of individual components within Mixture of Experts (MoE) models. Specifically, you will take each expert in a model like Mixtral (8x7B) and evaluate its performance on various benchmarks to determine its true area of expertise. The project involves designing a comprehensive evaluation framework to assess each expert's strengths and weaknesses, and identifying the contexts in which each expert is utilized by the MoE model. By conducting this analysis, you will gain insights into the specialization of experts within MoE models and the effectiveness of their deployment in different scenarios. The ultimate goal is to understand how expertise is distributed and leveraged within these models, providing valuable information for optimizing their performance.

Evaluating long-form open-ended generations As large language models that are trained using reinforcement learning with human feedback (RLHF) become more widespread, the generated outputs are becoming increasingly long. However, existing techniques for evaluating these outputs are becoming less useful. In response, preference models have been developed based on human comparisons of multiple possible outputs. This research project aims to define more rigorous methods for evaluating the quality of generated output, focusing on factors such as coherence, consistency, toxicity, and factuality. The project will explore what types of annotations on potential output passages are needed to improve model evaluations. Researchers will experiment with using these annotations to fine-tune or evaluate models. Ultimately, the goal of this project is to develop more principled approaches for evaluating the quality of generated outputs from large language models trained with RLHF, which can help improve the performance of these models in practical applications.

LLMs phrasing tone sensitivity The idea for the research project is to evaluate Language Models (LLMs) such as ChatGPT to determine whether the tone of the prompt, particularly the assertiveness and politeness, affects the performance of the model. This would involve designing a set of prompts with varying levels of assertiveness and politeness and evaluating the performance of the LLMs on these prompts. The evaluation could include metrics such as accuracy, fluency, and coherence. The goal of the project is to gain insights into how the tone of the prompt affects the performance of LLMs and to identify best practices for designing prompts that optimize performance.

Newspapers analysis It is well-known that different newspapers present the same information in vastly different light, that different authors use unique phrasings and that the language used when reporting different topics may vary greatly. This project is a way for you to explore this phenomenon yourself, showcase our creativity and perform real-world exploratory data analysis on raw texts. Specially, you can use the *Newspaper3k*, the data in *The GDELT Project* or build your own scraping pipeline to retrieve raw articles and some of their metadata. You will then analyze these documents using whatever tools you want to try and draw conclusions on the similarities and differences between them, and any other surprising fact you find. Don't forget to give nice visualization of what you find.

Evaluating Text Embeddings in CLIP Models for Visual-Language Tasks CLIP models are widely used in visual-language and text-to-image models, but it remains unclear which aspects of the textual conditions are effectively encoded. For instance, text-to-image models often struggle to generate the correct number of objects specified in the text. This project aims to develop a benchmark for testing text embeddings on tasks such as restoring object attributes (e.g., color, shape, size), counting objects, and understanding relations between objects (e.g., next to, right of). Building on this benchmark, you will examine several CLIP models, including a pretrained CLIP model and the one used in Stable-Diffusion 1.5. The goal is to determine whether linear probes can reveal what information is accessible to the vision component of these models and what gets lost during encoding. This investigation will help clarify the strengths and limitations of CLIP models in encoding textual conditions and improve their application in various visual-language tasks.

Multi-lingual knowledge transfer The concept for this research project is to examine the ability of Large Language Models (LLMs) like ChatGPT to share and transfer knowledge across different languages. Specifically, the study aims to ascertain whether a model that can accurately respond to a query in one language fails to do so in a mother language, and if this discrepancy is observed, we aim to delve into the underlying reasons. This would involve crafting a series of queries in diverse languages and evaluating the proficiency of the LLMs in responding accurately to these prompts. If failure is observed, the research would then experiment with potential strategies to overcome this, such as integrating an intermediary translation phase, or instructing the model to conduct a chain-of-thought process that includes translation back and forth.

Replicating Discrimination Assessment in LMs with Focus on Jewish People and Israel-Associated Individuals This project aims to adapt the methodology used in the referenced paper to specifically investigate how LMs handle decisions involving Jewish people and Israel-associated individuals. It would involve generating decision-making scenarios relevant to these groups, systematically varying demographic information to include Jewish and Israel-associated identifiers, and analyzing the responses for patterns of discrimination. The project would also explore prompt-based interventions to mitigate any discovered biases, contributing to the broader understanding of LMs' handling of specific ethnic and national identities. For more details on the original paper, you can access it here.

3.3 Mentored projects

The projects listed below were suggested by practicing NLP researchers at the university and can be pursued under the guidance of these researchers with the goal of achieving a publication. These projects are more extensive and may demand substantial effort and dedication, but they also offer the advantages of collaborating with experienced researchers. While **you are free to work on any of these ideas independently**, mentorship is contingent upon their consent. If you are interested in pursuing any of these ideas, please email me, and I will connect you with the mentor to arrange an initial meeting to assess compatibility.

Also, if you feel you have a good idea and are willing to work hard on pursuing it but could benefit from working with a mentor, **clearly note it in your proposal**. I will distribute these proposals among additional mentors to see if any of them have interests aligned with the proposal, and if so, I will make the connection.

Automating Sentence Generation for Psycholinguistic Experiments One of the time-consuming tasks in psycholinguistic experiments is creating sentences that adhere to a strictly defined structure. Current state-of-the-art Language Models (LLMs) often fail to generate experiment sets that meet these specific criteria. In this project, you will design and implement a more sophisticated pipeline that combines advanced sentence generation and validation models. The goal is to create sentences that follow the required experimental templates accurately. Your task will include developing or refining models to generate sentences and evaluating their adherence to the templates using another model designed to judge compliance. This project aims to streamline the process of sentence creation for psycholinguistic experiments, ultimately saving time and ensuring high-quality, structurally accurate sentences for research. **The project will be mentored by Ph.D. student Samuel Amouyal**

Building a Verb Property Database for Psycholinguistic Experiments Creating sentences for psycholinguistic experiments often requires verbs with specific properties, such as transitivity or reflexivity. Currently, there is no comprehensive database that provides this information about verbs. This project involves building such a database by scraping web data and parsing the usage of various verbs to determine their properties. This task is primarily an engineering challenge but also offers a research component suitable for students interested in both fields. You will design a web scraping and parsing pipeline to collect and analyze verb usage, classifying verbs based on their syntactic properties. The goal is to create a valuable resource that can significantly aid in the design of psycholinguistic experiments, providing researchers with easy access to verb properties and improving the efficiency and accuracy of sentence creation. **The project will be mentored by Ph.D. student Samuel Amouyal**

Investigating Cognitive Capabilities in NLP Models Continuing the work led by three students last year, this project aims to further explore the cognitive capabilities of Natural Language Processing (NLP) models and work towards a publication. NLP models have demonstrated remarkable progress in various language-related tasks, such as question answering, sentence comprehension, summarization, common sense reasoning, and translation. Despite these advancements, understanding the extent of their cognitive abilities remains challenging. A critical question is to what degree the mechanisms underlying human language comprehension correspond to those used by language models. In this study, you will investigate the similarities and differences between human cognitive skills and those developed by trained language models, focusing on models like MultiBERTs. This involves analyzing the cognitive skills of these models in interpreting complex sentence structures and comparing their predictions with human performance. Preliminary results suggest that models outperform human participants in most linguistic scenarios tested. This project aims to build on these findings, refine the analysis, and contribute towards a comprehensive understanding of the cognitive capabilities of language models. All related code can be found in our GitHub repository. **The project will be mentored by Ph.D. student Samuel Amouyal**

3.4 Mor's mentored projects

Important note: I am planning to mentor not more than 3 groups. If you are interested in working on this project, please reach out by August 8th with a project plan that includes a concrete research question and proposed analyses or experiments. I will announce the groups I will mentor later in August. Priority will be given to groups who wish to turn their project into a scientific paper, provided the project has significant research potential.

Project description: Sparse auto-encoders (SAE) have been emerged as a powerful interpretability tool for extracting features from latent representations of large language models (LLMs) in an unsupervised manner. However, only little has been done in evaluating and analyzing the extracted features, and using the capabilities of SAEs to tackle outstanding questions regarding the inner-workings of LLMs, such as how factual knowledge is encoded and how memorized sequences are captured. Groups working on this project could leverage recent repositories of trained SAEs for open-source models, such as this or this.

References and resources:

Scaling and evaluating sparse autoencoders

Towards Monosemanticity: Decomposing Language Models With Dictionary Learning

SAELens

4 NLP Research project grading guideline

Each project will be graded according to the following guideline. While not every mentioned point is applicable for all works, you should take a look at them at least once to give you a sense of what should be considered when writing your project.

Quality of research question (5pt) The project should start with a clear and concise research question that is relevant to the subject of the class. The question should be original and significant. This is mainly done during your project proposal phase.

Ambitiousness and effort (5pt) This considers the complexity and novelty of the chosen research question, the innovation in methodology or approach, the potential impact of the research, and the demonstrated effort in pursuing an ambitious and challenging project. This criterion rewards students for intellectual risk-taking and innovation, and for pushing the boundaries of their knowledge and skills, even if the final results may not be as polished or conclusive as less ambitious projects.

Literature review (10pt) The project should include a well-written and comprehensive literature review, providing a background of existing research on the topic and contextualize your work within it. It should critically evaluate previous research and identify gaps in the literature that the project addressed.

Methodology (25pt) Assess the methodology used to answer the research question. The methods should be appropriate for the question and be rigorously executed. Depending on the nature of the project, this might involve an evaluation of experimental design, data collection and analysis, or theoretical argumentation. This clause also asserts that your methodology is aligned with the project's expected scope.

Some points for consideration

- When performing experiments, always make sure you follow best-practices such as avoiding data-leakage and overfitting (remember train/validation/test splits?), using appropriate models and sensible hyper-parameters. Also, always compare to relevant baselines. For example, a naive majority-vote classifier and/or most common approach used in the literature at the moment. This is the only way to frame your results in any meaningful way. Similarly, always account for randomness (e.g. seeds, prompts affect, randomness from the decoding scheme).
- If you train a model and can't get any meaningful results, at least make sure you are able to overfit on at least a small sample of the data - otherwise you're doing something wrong and probably have a bug. Show me that sanity experiment worked.
- When performing experiments, it is generally a best practice to support reproducibility of your work. This is usually done by specifying the exact settings in which you ran the experiment, and provide access to the code and data you used. When not possible, clearly state what the reason is.

Results and Discussion (25pt) The paper must present findings with precision and coherence. The discussion should contextualize these results within the research question and existing literature. Both positive and negative results are valued, provided they stem from sound methodology, are thoroughly analyzed, and meaningfully contribute to the discourse.

Some points for consideration

- When you present results, you should discuss what conclusions are stemmed from them.
- Always make sure your results are given in full and in an easy-to-understand format (whether it is a table, a scatter plot, a bar plot or simply inline text)
- Dataset statistics: when working with datasets, it often helpful for the reader to understand how the data is built, what is its size, and if annotated, what is the label distribution. When the samples' description is not trivial, also consider including an example.
- Figures: it is often very helpful to have a figure/algorithm box outlining your method (when appropriate) which both helps the reader understand what you are describing, and lets you refer to it while outlining your method. Additionally, it is often a nice touch to add figure in the first or second page with some demonstration of your main results, to which you can refer from your introduction when discussion your contributions.

Citations and Bibliography (5pt) The paper should properly cite all sources of information used in the project. The citation style should be consistent and the bibliography should be formatted correctly (if you haven't used the `\citet` and `\citep` macros, you probably used the wrong format). Note that while literature review is aimed to give an understating of where your work is placed in the current state of your field, this criterion verifies that when mentioning work or claims from prior work you give the correct attribution.

Presentation and Communication (25pt) This assesses the clarity of writing, narrative coherence, grammatical accuracy, and the visual clarity of figures and explanations. It evaluates the overall organization and aesthetic presentation of the work, highlighting the importance of conveying complex research in a clear, concise, and engaging manner. This is also where you are evaluated on your division of content to the proper sections and following the expected format (ACL template).

Some points for consideration

- Structure: The introduction should set the stage to what is the problem you are trying to solve, what are the high-level ideas you used to solve the problems and what are your main findings. A thorough background is usually best placed in its own section.
- Make sure the paper is self sufficient. Namely, Ensure that any uncommon models and metrics are clearly defined, and all the important background is there. If you rely on previous work, make sure it is accessible.
- When using ChatGPT/LLMs, make sure you stand behind what it writes. Beware of using it too much. While on the surface the generated content sounds good, very often the text becomes too boastful, and very not succinct, hiding the important details in too much text.
- Avoid compiling a detailed work log with every step and issue encountered. Instead, focus on creating a white paper. For example, rather than 'We tried X but it failed because of Y, leading us to try Z', write 'X was ineffective due to Y. Z proved successful. Implementation details are in...'. Be concise and omit exhaustive challenges.

- Use appendix with sparsity. While the appendix is a great place to put additional figures, proofs or examples, it should not contain any content that is critical to the flow of the paper. In fact, it should be assumed that the appendix is not read at all unless the reader is specifically interested in a more in-depth overview of a certain topic.
- Tip - figures should use large enough fonts and should be added as a pdf and not jpeg to ensure high quality. Do use ChatGPT to help you create beutiful plots easily!