Natural Language Processing **Tel Aviv University**

Final Project: Instructions & Ideas

Lecturer: Dr. Mor Geva

Important dates
Project proposal submission deadline: **June 10, 2025**
In-class project presentations (optional): **June 24, 2025**
Project submission deadline: **September 30, 2025**

1 Goals

The course project provides an exciting opportunity to:

- Conduct independent research in NLP. You will choose a paper or (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 choose a topic you are interested in and expand the scientific knowledge about it.
- Potentially contribute to science. While it would be great if your project yields new results, findings, or resources (in which case we will consider pursuing publication in an upcoming NLP conference), to get full credit it is sufficient 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.
- Experience research and the joy of new discoveries!

2 Project Instructions

2.1 Proposal submission and in-class presentations

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

(b) Choose one to three (but not more than three!) papers that are related to the course topics and were published in NLP/ML conferences (e.g., EMNLP, *ACL, ICLR, NeurIPS) ideally after 2018. Other papers/venues can be considered but need my approval.

✓ (c) Submit a proposal as described below for your project by **June 10, 2025** and get a binary approval from me. **Submitting a project proposal is mandatory** and subject to the same late-penalties as the homework assignments (affecting your final project grading).

✓ **Proposal** The project proposal should be up to a single page, and include the following information:

✓ 1. Project title

✓ 2. Names, IDs, and emails of the contributing students.

✓ 3. A brief description of the proposed project: what problem or research question you will tackle, what is the motivation for studying this problem/question (why is it interesting?), 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 proposed project is sufficient, 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 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 mentored project (see §4), you can simply submit that you will work on that, or extend the description if you have any additional thoughts about it.

✓ 7. If you would like to present your project in the last class of the semester (see details below), please indicate it in your proposal.

✗ **Project presentations in class (optional)** We will use most of the last class for project presentations. Groups are welcome to give 5-10 minute presentations of their project proposals and preliminary results (if relevant) in class. Presentations will be in English. Groups presenting their projects will receive up to 10 bonus points to their projects' grades.

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 could run jobs on the GPUs there. Please read the tips I will provide 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 under `/home/joberant/NLP_2425b/<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_2425b/<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.

In addition to the resources above, you may find that you need some special resources, such as OpenAI API credits or GPUs beyond what is available to you via Slurm. If you do, please email me ASAP, explaining your needs with the following details:

1. Group names, IDs and emails
2. Project title
3. As exact as possible estimates of your needs. For example, if you need API access, include how much credits you need and to which models (you can use OpenAI online tokenizer and pricing 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 required parts (see §3). If needed, after submission, I will schedule a short meeting with some groups to discuss their report.

3 Grading

Projects will be graded according to the guidelines below. While not all points are applicable to all projects, please read them carefully to get a sense of what should be considered when writing your project report.

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. Choosing a research question is part of the 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 background on prior research on the topic and contextualizing 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. 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 hyperparameters. Also, always compare with relevant baselines. For example, a naive majority vote classifier and/or the most common approach that is currently used in the literature. 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 a small sample of the data – otherwise you are doing something wrong and probably have a bug. Show me that the sanity experiment worked.
- Make sure that your results are reproducible (as much as possible). This is usually done by specifying the exact settings in which you ran the experiment and providing access to the code and data used. When not possible, clearly state the reason.

Results and discussion (25pt) The paper must present the findings accurately and coherently. 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. Points for consideration:

- When you present results, you should discuss what conclusions are derived 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 is often helpful for the reader to understand how the data was built, its size, and if annotated, what is the label distribution. When the samples' description is not trivial, also consider including examples. If you created data, make sure to describe the process through which you created it.
- Figures: it is often useful 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 discussing 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 have not 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 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). Points for consideration:

- Structure: The introduction should set the stage to what problem you are tackling, what high-level ideas you used to solve the problems, and what your main findings are. A thorough background is usually best placed in its own section.
- Make sure the paper is self-contained. 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 and properly cited.
- When using 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 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.
- 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 certain details.
- Figures should use large enough fonts and added as pdf files (not jpeg or png) to ensure high quality. Feel free to use LLMs to help you create beautiful plots easily!

Bonus of up to 10pt will be given to projects presented in class (see details in §2.1). The number of bonus points given will be determined based on how coherent, clear, and engaging the presentation is.

4 Project ideas

4.1 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, identifying problems that LLMs struggle with, 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.

4.2 Default projects

Existing model suites, such as OLMo and Pythia, include pre-trained/fine-tuned models and their training data, often with sequence-level metadata information (i.e. which sequences the model saw during each step of training). At TAU, we have been building such a suite that is focused on knowledge, called Knowledge Analysis Suite (KAS). KAS consists of models and training data, as well as fine-grained, entity-level metadata for every chunk of data seen by the models during pre-training. Specifically, the suite offers the following features:

- Each chunk is annotated with metadata listing the entities described (both explicitly or implicitly), the source Wikipedia page, etc.
- The models are of different sizes and were trained using standard tokenization, shuffling, and chunking techniques. These models perform reasonably well on simple factual queries.
- An API that allows querying entity statistics and chunk information, and retraining.

KAS (and other suites) provides an easy-to-use framework for studying various problems in LLMs and NLP. You can think about it as a playground where we can train, intervene, and analyze small-scale language models with full visibility into their training data. Keep in mind that these models are not as performant as leading LLMs (they are pre-trained, but not fine-tuned or aligned in any way, and were trained on little data). Still, as leading LLMs rely on similar architectures and their pre-training process is (likely) similar, conclusions from KAS could potentially transfer to them. Interesting problems one could explore include:

- Understanding training dynamics – how and when different capabilities are obtained by models.
- Knowledge and hallucinations – how and where in the model knowledge is being encoded during training, when do models start hallucinating, and more.
- Better selection of training data instances – we may be able to make training more efficient by removing data points that do not contribute to improving the model's performance.
- Memorization and generalization
- Adversarial settings (data privacy, malicious content injection into training data, biases, etc) + data contamination
- Exploring better architectures (mostly for knowledge-related capabilities)
- Many more, be creative!

Before choosing a project idea, please make sure you have a good understating of how you will approach the subject, and that you have access to any required resources to complete what you plan.

4.3 Mentored projects

The projects listed below were suggested by 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.

Standardized Web Agent Evaluation (mentored by Ph.D. student Ori Yoran) There has been a recent influx of new agents and benchmarks for autonomous web browsing. However, there are still gaps in our understanding regarding which tasks are challenging for current agents (e.g., do improvements generalize between benchmarks, are there similarities between tasks that models still struggle on) and how we should evaluate these agents in the future (which tasks should be used for evaluation, should we use LLM-as-a-judge or rule-based methods).

As part of this project, we will aim to answer these questions by performing a standardized evaluation of current agents and suggest improved methodologies for future evaluation.

Plan:

- Setting: Evaluate different SOTA web-agents (OpenAI CUA, DeepResearch, Manus, Claude computer use) on these benchmarks (BrowserComp, AssistantBench, Online-Mind2Web, GAIA, WebArena, WorkArena) to make sure we can directly compare between the models.
- Results: compare rule-based methods to LLM-as-a-judge - what are the pros and cons for each one? Which one should we use in the future?
- Goal: based on these results, derive a new benchmark for standardized evaluation of web tasks that are still challenging for current models.

Resources:

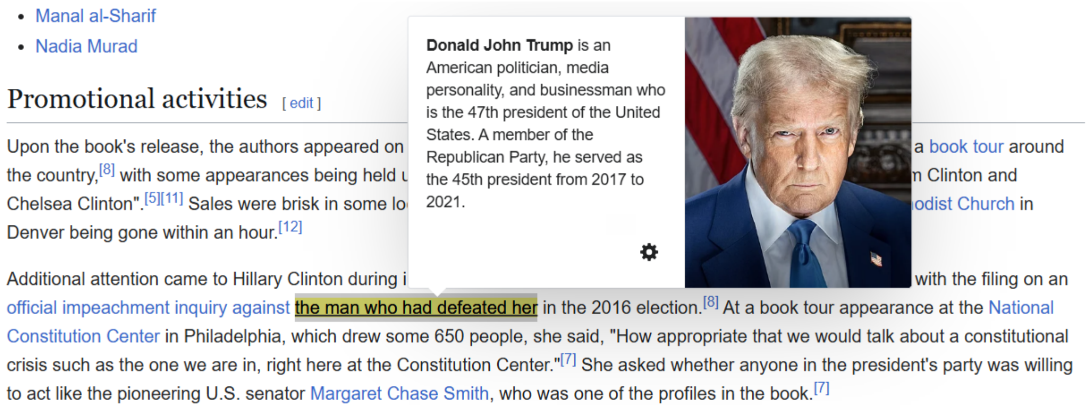
- BrowserComp
- CUA
- DeepResearch
- Manus
- Claude computer use
- Online-mind2web
- GAIA
- AgentRewardBench
- BrowserGym
- AssistantBench

Interpreting Steering Vectors (mentored by Ph.D. student Matan Ben-Tov) TL;DR: try to interpret and trace the origin of steering vectors, which are used to effectively control the behaviour of LLMs.

- What are these Steering Vectors? These are vectors that once added to (or ablated from) the LLM's residual stream, modify its "behavior", and control properties in its response. These are often extracted from the residual stream using diff-in-means (Arditi et al. 24', Panickssery et al. 23', inter alia), or PCA (Lee et al. 25', inter alia).
- As an example, Arditi et al. 24', extract the so-called refusal direction, using diff-in-means. They show that adding it to the residual stream encourages the model to refuse, while projecting it out of the residual stream drastically diminishes refusal in model response. Other works explore different concepts, using similar methodology.
- This phenomenon raises at least two intriguing research questions:

1. Can these directions be interpreted? We can try to interpret them in the input/output embedding space (Dar et al. 22') Do they promote a specific token(s) (w.r.t. the output embedding matrix)? (nostalgebraist 2020) Can we express them with a subset of output token embeddings (e.g., through a linear combination of them)?
2. Can we trace the origin of these directions? These directions are extracted from the residual stream, and specifically these directions are stronger for one example sample set, while being weaker in another. As the residual stream is merely a summation of the models' components, we could ask what component contributes most in this Steering Vector direction. Is this direction contributed exclusively by MLPs or attention sub-layers? Are there specific components (e.g., a specific head) that contribute in this direction?

Investigating how LLMs process implicit entity mentions (mentored by Ph.D. student Daniela Gottesman) Current large language models are trained on massive datasets containing both explicit and implicit entity references. However, it remains unclear to what extent they resolve more implicit references that require contextual understanding. This project aims to investigate how pre-trained language models resolve implicit entity mentions — references to entities not directly named but implied through context (e.g., "the man who had defeated her" referring to Donald Trump after a co-mention with Hillary Clinton, see Figure 1). We seek to understand whether and how language models make these associations, and whether specific data ordering strategies during pre-training — such as progressing from explicit to implicit mentions — can improve entity resolution capabilities. We are also interested in understanding how implicit knowledge is encoded in the model compared to explicit knowledge.



- Manal al-Sharif
- Nadia Murad

Promotional activities [edit]

Upon the book's release, the authors appeared on
the country,[8] with some appearances being held u
Chelsea Clinton".[5][11] Sales were brisk in some lo
Denver being gone within an hour.[12]

a book tour around
n Clinton and
odist Church in

Donald John Trump is an American politician, media personality, and businessman who is the 47th president of the United States. A member of the Republican Party, he served as the 45th president from 2017 to 2021.

Additional attention came to Hillary Clinton during i
with the filing on an
official impeachment inquiry against the man who had defeated her in the 2016 election.[8] At a book tour appearance at the National Constitution Center in Philadelphia, which drew some 650 people, she said, "How appropriate that we would talk about a constitutional crisis such as the one we are in, right here at the Constitution Center."[7] She asked whether anyone in the president's party was willing to act like the pioneering U.S. senator Margaret Chase Smith, who was one of the profiles in the book.[7]

Figure 1: Example of an implicit mention of Donald Trump.