50.040
Natural Language Processing

Lecture 9: Final Project & Practical Tips

Wenxuan Zhang

wxzhang@sutd.edu.sg

Outline

We'll talk about:

1. Details of the final project (which counts for 30% of your final grade!)

2. Practical tips for custom project / any NLP project

Outline

We'll talk about:

1. **Details of the final project (which counts for 30% of your final grade!)**

2. Practical tips for custom project / any NLP project

Project: two options

Default Project

- Tasks:
  - implement a GPT-2
  - pretrain on a toy dataset
  - fine-tune on English data
  - extend to multilingual settings
  - with other possible extensions
- It connects most of the important concepts in this course (with more practical details)
- Check the doc (with video walk-through)

Custom Project

- Propose your own project
- Has to be related to NLP!
  - For example, it can be a multimodal one, but at least one modality should be text
  - Example-1: Sign Language Translation ✅
  - Example-2: RNN for Music categorization ❌
- Proposal presentation is a must (to ensure what you proposed is doable)

Project: two options – how to choose?

Default Project

- If you:
  - have limited experience with research, don't have any clear idea of what you want to do, or want guidance and a goal…
- It starts with clear guidelines, but ends with open-ended extensions. You can still explore different topics of interest of your own (e.g., test different models, efficiency, multilingual etc).
- Strong default final projects do new things.

Custom Project

- If you:
  - Have some research project that you're excited about (and are possibly already working on)
  - You want to try to do something different on your own
  - You want to see more of the process of defining a research goal, finding data and tools, and working out something you could do that is interesting, and how to evaluate it

- There are great default final projects and great custom final projects … and there are weak default final projects and weak custom final projects.
- The evaluation is based on what you have done in the project

Project Grading

- Proposal & Presentation: 5%
- Final Submission: 25%
  - Presentation and QA: 5%
  - Final Report: 15%
  - Demonstration and Impact: 5%

Project Grading: Proposal & Presentation

- **Proposal & Presentation: 5%**
  - A proposal is required for ***all*** teams (due: Nov 9)
    - It should cover what you plan to work on and how you can innovate in your final project work (e.g., suggest some milestones)
      - For default project: this is more like a progress check to remind you to start working on it!
      - For custom project: you should clearly describe your plan (the research question, relevant existing literature, the models you will use/explore, the data you will use and how it is obtained etc)
    - Use the ACL paper template, 3-4 pages (excluding references)

Project Grading: Proposal & Presentation

- **Proposal & P**
  - A proposa
    - It shou
      - Fo
      - Fo
    - Use th

project work (e.g.,

working on it!

tion, relevant

how it is

Disentangling Language and Culture for Evaluating Multilingual Large Language Models

Jiahao Ying [1,2*], Wei Tang [3,5], Yiran Zhao [3,5], Yixin Cao [4], Yu Rong [5], Wenxuan Zhang [6†]
[1] Singapore Management University
[2] University of Science and Technology of China
[3] National University of Singapore
[4] Institute of Trustworthy Embodied AI, Fudan University
[5] DAMO Academy, Alibaba Group
[6] Singapore University of Technology and Design

Abstract

This paper introduces a Dual Evaluation Framework to comprehensively assess the multilingual capabilities of LLMs. By decomposing the evaluation along the dimensions of linguistic medium and cultural context, this framework enables a nuanced analysis of LLMs' ability to process questions within both native and cross-cultural contexts cross-lingually. Extensive evaluations are conducted on a wide range of models, revealing a notable "Cultural-Linguistic Synergy" phenomenon, where models exhibit better performance when questions are culturally aligned with the language. This phenomenon is further explored through interpretability probing, which shows that a higher proportion of specific neurons are activated in a language's cultural context. This activation proportion could serve as a potential indicator for evaluating multilingual performance during model training. Our findings challenge the prevailing notion that LLMs, primarily trained on English data, perform uniformly across languages and highlight the necessity of culturally and linguistically aware evaluations. Our code can be found at https://yingjiahao14.github.io/Dual-Evaluation/.

1 Introduction

With the rapid development of large language models (LLMs), increasing efforts have been made to make these models beneficial for people worldwide. To achieve this, non-English corpora are also incorporated into the training data, enabling LLMs to understand and generate text in various languages (i.e., multilingual capabilities) (Xue et al., 2021; Grattafiori et al., 2024; OpenAI et al., 2024; Nguyen et al., 2024; Zhang et al., 2024).

To evaluate the LLMs' multilingual capabilities, researchers primarily rely on translating English-centric benchmarks into target languages, such as translating MMLU (Hendrycks et al., 2021) into MMMLU (OpenAI, 2024). While this approach allows for efficient cross-lingual comparisons, it limits the evaluation to scenarios rooted in English-speaking cultural contexts, as the original data was predominantly collected from perspectives prevalent in English-speaking countries. In contrast, recent work has developed culture-specific benchmarks such as M3Exam (Zhang et al., 2023) and BLEnD (Myung et al., 2024), where evaluation data are sourced from authentic, real-world scenarios in native-speaking regions. While these better capture the majority of local usage, they also overlook that multilingual users frequently ask questions across cultural boundaries. For example, a Spanish speaker might inquire about Chinese tea usage in Spanish, while a user from China could seek details about Diwali celebrations in Chinese. These existing evaluations on multilingual capabilities, however, treat language and cultural context as inseparable dimensions, restricting analyses to single-language scenarios.

Figure 1: Dual Evaluation Framework for evaluating multilingual capabilities of LLMs. The figure is divided into four quadrants, each showing the model's performance on questions framed in different **languages** (horizontal-axis) and **cultural contexts** (vertical-axis). The score refers to the aggregated performance of the model Claude-3.5-Sonnet on these four question sets.

To comprehensively evaluate multilingual capability, especially considering the real-world usage, we propose a Dual Evaluation framework in this paper, which decomposes the multilingual capability evaluation along two critical dimensions: (1) **linguistic medium** (the language used for questioning) and (2) **cultural context** (the regional and cultural knowledge being tested). As illustrated in Figure 1 through a preschool enrollment example, this framework generates four distinct evaluation scenarios from a single question template. This structured decomposition enables multiple essential multilingual capability assessments, including native cultural-linguistic alignment (same language and culture), cross-lingual understanding (different language, same culture), and cross-cultural ability (same language, different culture).

With such a dual evaluation framework design, we construct a dataset by adopting and extending the BLEnD dataset (Myung et al., 2024), which contains every-day questions across different cultural contexts. We then evaluate a wide range of open-source and close-source models with this newly constructed benchmark. Our findings indicate that: 1) Models generally perform better on scenarios rooted in English-speaking culture, a pattern that persists cross-lingually (Section 3.2), and 2) LLMs perform better when questions are posed in the language that corresponds to the cultural context of the question, rather than in English (Section 3.3). The second finding, in particular, draws our attention because most existing models are primarily trained on English data and have demonstrated strong performance in other multilingual evaluations like MMMLU. However, when faced with real-world culturally relevant questions in the corresponding language, these models perform better in that language than in English. We refer to this phenomenon as "**Cultural-Linguistic Synergy**" (as shown in Figure 1, Claude-3.5-Sonnet has better performance on the Chinese test than the English test when asking about Chinese culture questions, vice versa).

To understand the underlying causes of this phenomenon, we conduct interpretability probing by analyzing the activation status of neurons when answering questions in different languages and cultural contexts, we find that: 1) The proportion of specific neurons tends to be higher when the question is in the corresponding language and cultural context, which could explain the observed "Cultural-Linguistic Synergy" (Section 4.3); 2) Additionally, this proportion of specific neurons could serve as a potential indicator for comparing multilingual capabilities during model training (Section 4.3.1); 3) The number of neurons activated in the model is strongly correlated with the model's performance in the corresponding language. Specifically, when the question is in the English-speaking cultural context, the model tends to activate more neurons, leading to better performance (Section 4.3.2).

Our main contributions can be summarized as:

- We propose a Dual Evaluation Framework, which decomposes the multilingual capability evaluation along two critical dimensions, linguistic medium and cultural context.
- Through extensive experiments, we find the Cultural-Linguistic Synergy phenomenon: the selected models perform better on native cultural scenario questions when asked in the corresponding language, compared to English.
- We demonstrate that the proportion of specific neurons activated for a given language can explain the observed Cultural-Linguistic Synergy, and that this proportion can serve as a potential indicator for comparing multilingual capabilities.

2 Dual Evaluation Framework

To comprehensively assess the multilingual capabilities of LLMs, we propose a Dual Evaluation framework that evaluates along two critical dimensions: (1) **linguistic medium** (the language used to pose questions) and (2) **cultural context** (the regional and cultural knowledge being tested). This dual-axis approach reflects three fundamental requirements for real-world applications: first, the ability to handle native language queries within their cultural context (e.g., answering "*What is a common children's snack in Spain?*" in Spanish), second, the capability for cross-lingual understanding, (e.g., answering questions about Spanish culture in Spanish and English); and third the capability to address cross-cultural inquiries through a single linguistic medium (e.g., answering "*What is a traditional festival in Japan?*" in English). By evaluating LLMs in both dimensions, we can measure how well models adapt to language-specific usage scenarios while maintaining cross-lingual and cross-cultural competence.

*This work was performed when Jiahao Ying, Wei Tang and Yiran Zhao were interns at Alibaba DAMO Academy.
†Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22230–22251
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Project Grading: Proposal & Presentation

- **Proposal & Presentation: 5%**
  - A proposal is required for ***all*** teams (due: Nov 9)
    - It should cover what you plan to work on and how you can innovate in your final project work (e.g., suggest some milestones)
      - For default project: this is more like a progress check to remind you to start working on it!
      - For custom project: you should clearly describe your plan (the research question, relevant existing literature, the models you will use/explore, the data you will use and how it is obtained etc)
    - Use the ACL paper template, 3-4 pages (excluding references)
  - A presentation is compulsory for custom project (to ensure what you proposed could be done within this term) during Nov 10 lecture.
    - Timetable will be decided and announced later (depending on how many custom project groups)
    - For default project: you are welcome to present as well

Project Grading – Final Presentation & QA

- **Final Presentation and QA: 5%**
  - Present your project on Week 13
  - Expected 7min presentation + 3min QA
  - Evaluation mainly on clarity and accuracy of what you have done

Project Grading – Final Report

- **Final Report: 15%**
  - Use the ACL paper template, max 6 pages (excluding references)
  - Due on Dec 7, 23:59pm (same late submission policy applies: 1 day late = -1)
  - Evaluations on:
    - How much work you have done
      - For default project: finish task 1 - 3
    - Innovation (Is there anything new?)
      - For default project: try some extension tasks
    - Writing quality will be considered as well

Project Grading – Demonstration and Impact

- **Demonstration and Impact: 5%**
  - Showing *what you have done is important and has real-world impact*
  - For all project: you need to submit the report + code (2%)
  - To earn more scores, consider:
    - Build a demo (to show the internal attention map?)
    - Conduct visualization to observe difference on different languages?

    Or even better if you:
    - Train and release a useful model (and get 10k downloads?)
    - "Sell" it to a stakeholder and he/she showed great interest?

Project-related Timeline

- Week 5: Finish team grouping ✅
- Week 8
  - Due Nov 3 Mon: Decide custom or default project
  - Due Nov 9 Sun: Project Proposal
- Week 9:
  - Nov 10: Compulsory proposal presentation for custom project
- Week 12:
  - Due Dec 7 Sun: Submit final Report (with code and any other material)
- Week 13:
  - Final project presentation for all teams

Project-related Timeline

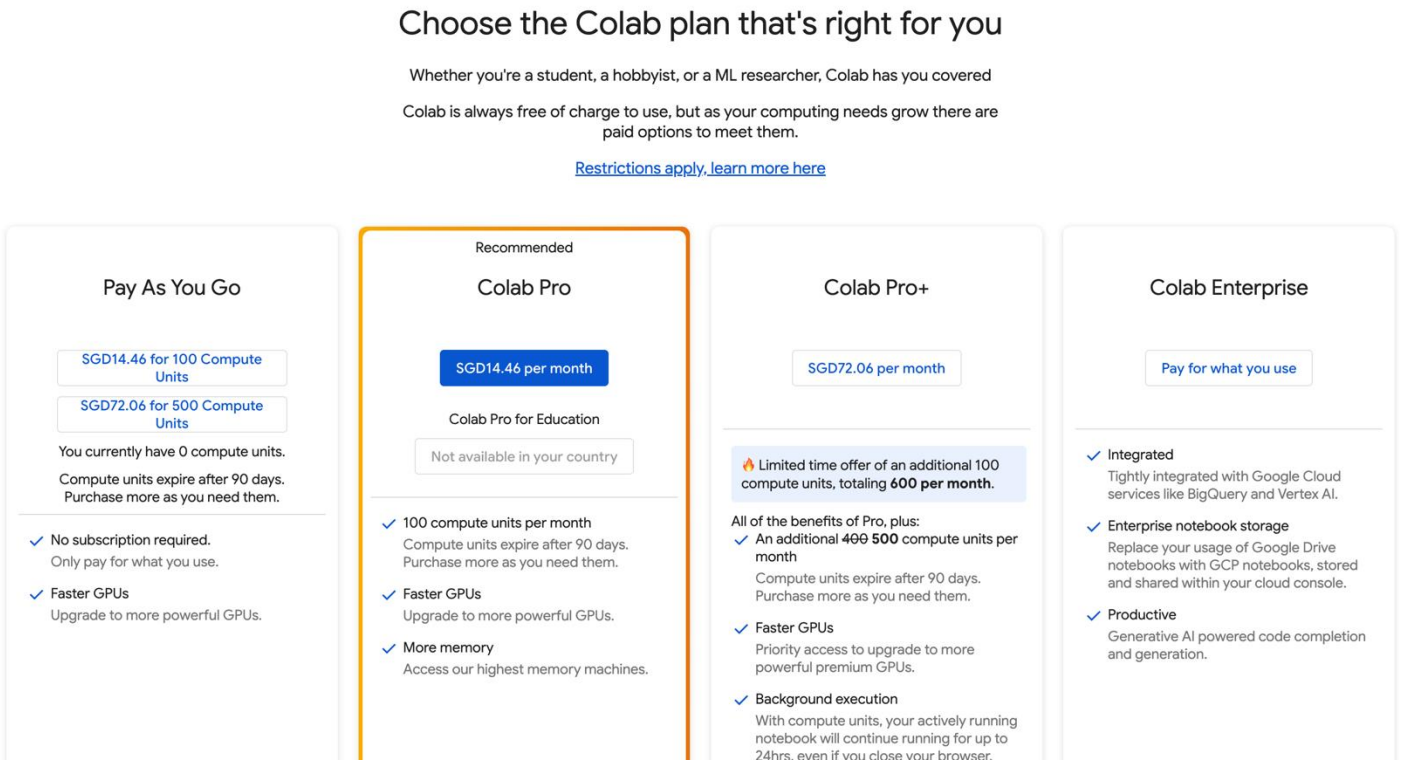
- Week 5: Finish team grouping ✅
- Week 8
  - **Due Nov 3 Mon: Decide custom or default project**
    - **Help fill in the project type in the group list form**

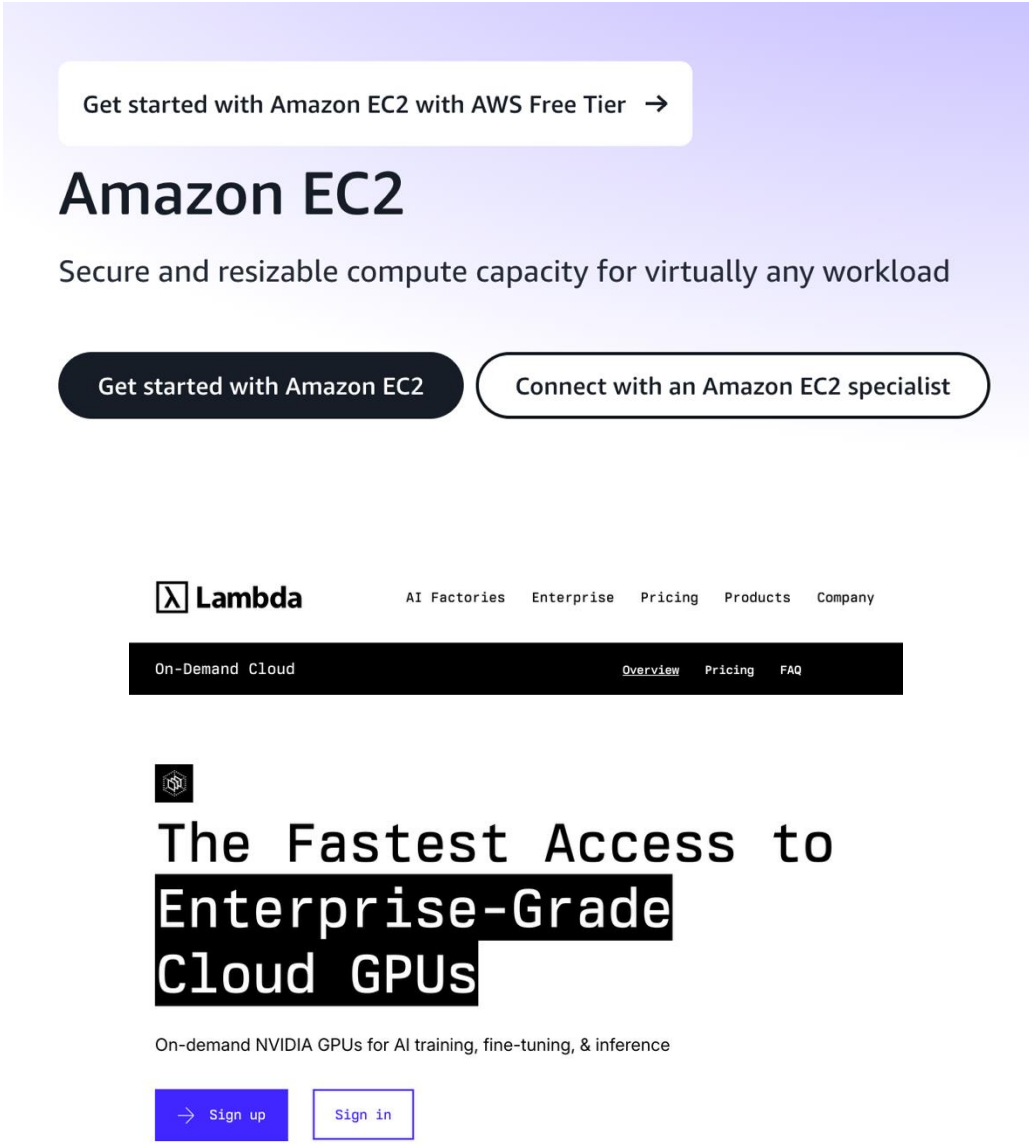
Team	Project Type	Name	Sid	Email
1		[redacted]	[redacted]	[redacted]
2		[redacted]	[redacted]	[redacted]

Resources & Support

- Computing resources
  - $80 budget per group (pending for final confirmation)
  - You can use the budget to
    - Cloud GPU platforms
    - API call (LLMs)
    - Data annotation
    - …

Resources & Support





Resources & Support

- Computing resources
  - $80 budget per group (pending for final confirmation)
  - You can use the budget to
    - Cloud GPU platforms
    - API call (LLMs)
    - Data annotation
    - …

- Support (particularly for custom projects)
  - Talk to TA/me/anyone you know that can be helpful
  - We are happy to help – but the problem is - it's really a big class 

Rules

- Again, respect the code of conduct

- It's okay to use existing code/resources/LLMs, but you must document it
  - You will be graded on your value-add

- Project scoring
  - Include a brief statement on the work/contribution of each team-mate in the final report (this won't be counted into the 6-page limit)
  - We'll give a score to a team first, then adjust according to the survey
  - **Survey**: we'll send out a survey near the end, asking you to comment on your teammates' contributions, then we'll adjust the scores of individuals from the group scores (adjust = add or minus)

Outline

We'll talk about:

1. Details of the final project (which counts for 30% of your final grade!)

2. **Practical tips for custom project / any NLP project**

1. Finding Research Topics

Two basic starting points (for all of science):

1. [Nails] Start with a (domain) problem of interest and try to find good/better ways to address it than are currently known/used
2. [Hammers] Start with a technical method/approach of interest, and work out good ways to extend it, improve it, understand it, or find new ways to apply it

In practice, you are often facing scenario #1 (e.g., your boss tells you what he/she wants you to solve/improve), but now you might be in scenario #2 (i.e., you want to apply what you have learnt in this course to some interesting problems)

Example NLP research projects

Improve a system

- Pang et al. (2002) propose a task of sentiment analysis, because "*labeling these articles with their sentiment would provide succinct summaries to readers*"
- Gehrmann et al. (2018) propose a method of bottom-up abstractive summarization because "*NN-based methods for abstractive summarization produce outputs that are fluent but perform poorly at content selection*."

Know more about language

- Cotterell et al. (2018) ask *"are all languages equally hard to language model?"*
- Tenney et al. (2019) quantify *where specific types of linguistic information are encoded in BERT*.

Sentiment Analysis (Pang et al. 2002)
Bottom-up Abstractive Summarization (Gehrmann et al. 2018)
Are All Languages Equally Hard to Language-Model? (Cotterell et al. 2018)
BERT Rediscovers the Classical NLP Pipeline (Tenney et al. 2019)

Example of custom course projects

- **Example-1: Fake News Detection in Thai**
  - Build a classifier to distinguish between real and fake news in Thai.
  - Explore small/large models that are capable of handling Thai
  - You can also analyze what linguistic features contribute to misleading content (e.g., exaggerated adjectives, sensational phrasing).

- **Example-2: SUTD Campus Q&A Chatbot**
  - Create a chatbot that answers questions about SUTD campus life, such as facilities, food options, academic schedules, or directions
  - Optionally use retrieval-augmented generation (RAG) to search campus documents or websites for accurate responses

- **Example-3: Bias & Toxicity in LLM Outputs in Singapore Context**
  - Examines whether large language models generate biased or toxic content when prompted with Singapore-specific topics, such as race, gender, or national service.
  - Design relevant prompts, analyze model outputs, and explore ways to reduce bias or toxicity

Suggestions for choosing a custom project

- Consider feasibility and timeline
  - Make sure the project can be completed within 1–2 months, including data collection, implementation, and evaluation.
- Leverage your own background and strengths
  - Think about your unique skills: do you speak other languages? Are you familiar with specific cultures (e.g., Singaporean context) that could inspire interesting datasets or problems?
- Start with a clear and focused research question
  - Avoid overly broad topics. Define a specific task, dataset, or hypothesis early on to stay on track.
- Incorporate novelty in a manageable way
  - You don’t need to fully invent something new — small improvements, applying NLP to a unique language/context, or combining two ideas is enough.

2. Literature Review

Question: Has anyone tried this before?

Answer: Do a literature review!

Research Survey Methods

1. Keyword search
2. Find older/newer papers
3. Read abstract/intro
4. Read details of most relevant papers

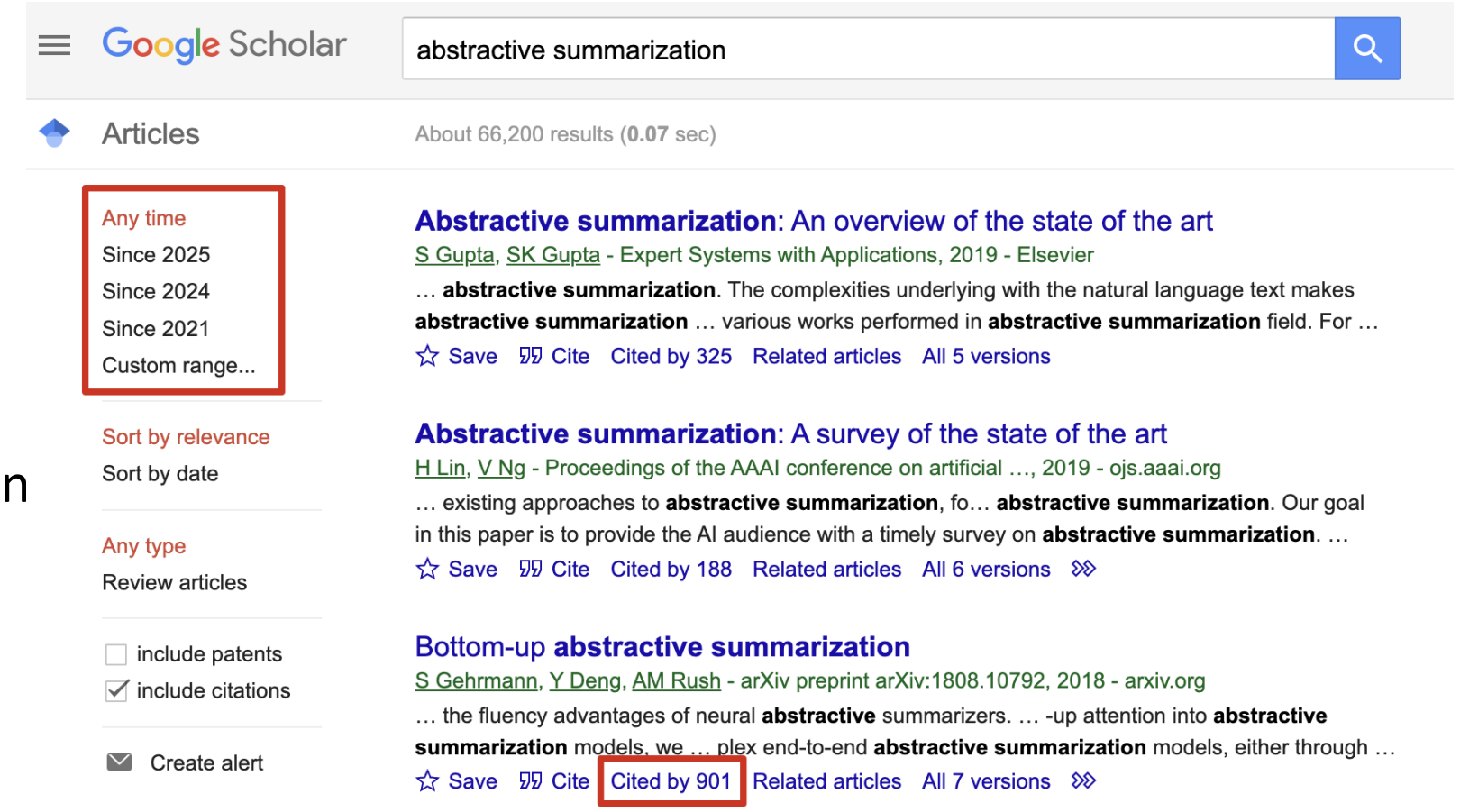
2. Literature Review

Question: Has anyone tried this before?

Answer: Do a literature review!

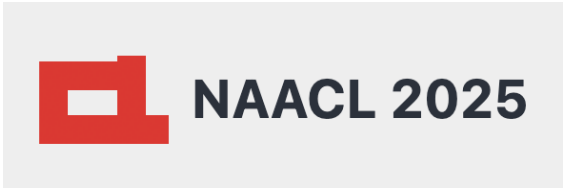
Research Survey Methods

1. Keyword search
2. Find older/newer papers
3. Read abstract/intro
4. Read details of most relevan



Sources of NLP papers

- Look at ACL anthology for NLP papers: https://aclanthology.org/
  - Start with past 3-5 years of several top venues (e.g. ACL, EMNLP, NAACL, TACL)



- Also look at the online proceedings of major ML conferences:
  - NeurIPS https://papers.nips.cc
  - ICML https://icml.cc/virtual/2025/papers.html
  - ICLR https://openreview.net/group?id=ICLR.cc

- Look at online preprint servers, especially: https://arxiv.org

2. *Extensive* Literature Review

Surveying extensively before doing research:

- Prevents you from duplicating work
- Increases your "toolbox" of methods
- [Cons] Constrains your thinking

Please still conduct an extensive survey before you start

- Your proposal should include a summary of previous methods

3. Test with Experiment

3.1 Find data for your project

3.2 Run experiments and calculate numbers

3.3 Analyze effects

3.1 Data

- **You can collect your own data for a project – it's great, but it's trick to do fast!**
  - You can annotate a small amount of data
  - You can find a website that effectively provides annotations, such as likes, stars, ratings, responses, etc (recall the sentiment classification example!).
    - This let's you learn about real word challenges / tips of applying ML/NLP!
  - But be careful on scoping things so that this doesn't take most of your time!!!

- Some people have existing data from a research project or company
  - Fine to use providing you can provide data samples for submission, report, etc.

- **Most people make use of an existing, curated dataset built by previous researchers**
  - You get a fast start and there is obvious prior work and baselines

3.1 Data

Your decision making process should be:

If building on previous work:

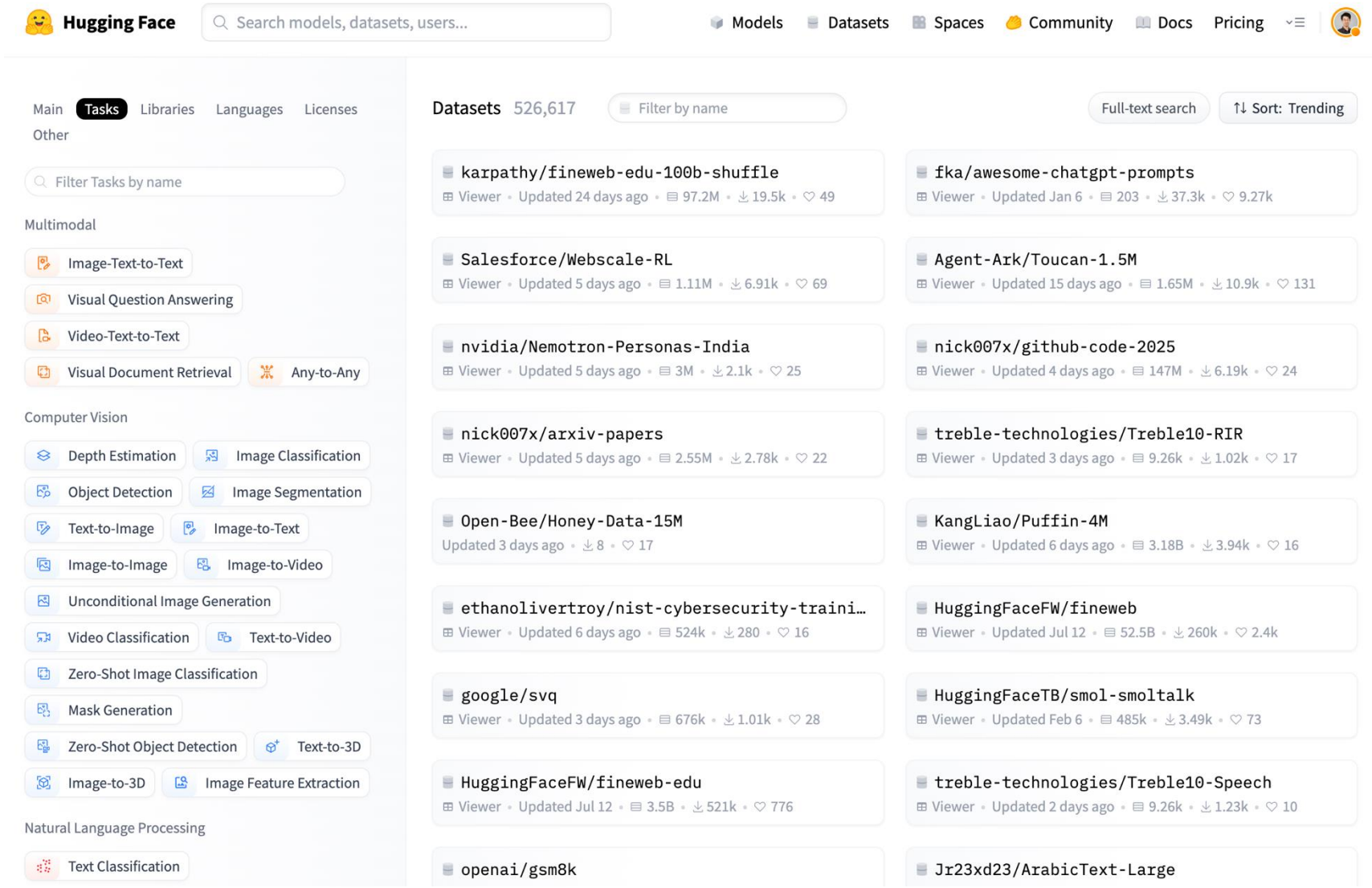
safest to start with same datasets ← most common cases

Elif answering a new question:

Can you repurpose other datasets to answer the question?

If not, you'll have to create your own

Example: Huggingface Datasets



https://huggingface.co/datasets

3.1 Data

Annotating data:

1. Decide how much to annotate
2. Sample appropriate data
3. Create annotation guidelines
4. Hire/supervise annotators
5. Evaluate quality

Note: this describes the general process, but it is not recommended to do data annotation for your final project (unless your whole custom project is about this)

Best Practices for Data Annotation (Tseng et al. 2020)

3.1 Data

- How much train/test/dev data do I need?
  - For dev/test: enough to observe effect between methods
  - For train: more is usually better

- How Should I Sample Data?
  - Coverage of the domains that you want to cover
  - Coverage of the language varieties, demographics of users
  - Documentation: Data Statements for NLP (Bender and Friedman 2018)

3.2 Running Experiments

- **Start with a positive attitude - Neural networks want to learn!**
  - If the network isn't learning, you're doing something to prevent it from learning successfully
  - There are lots of things that can cause neural nets to not learn at all or to not learn very well => finding and fixing them (with patience)

- **Work incrementally**
  - Start with a simple model and get it to work, it's hard to fix a complex but broken model
  - Initially run on a tiny amount of data. If everything is fine, train and run your model on a large dataset

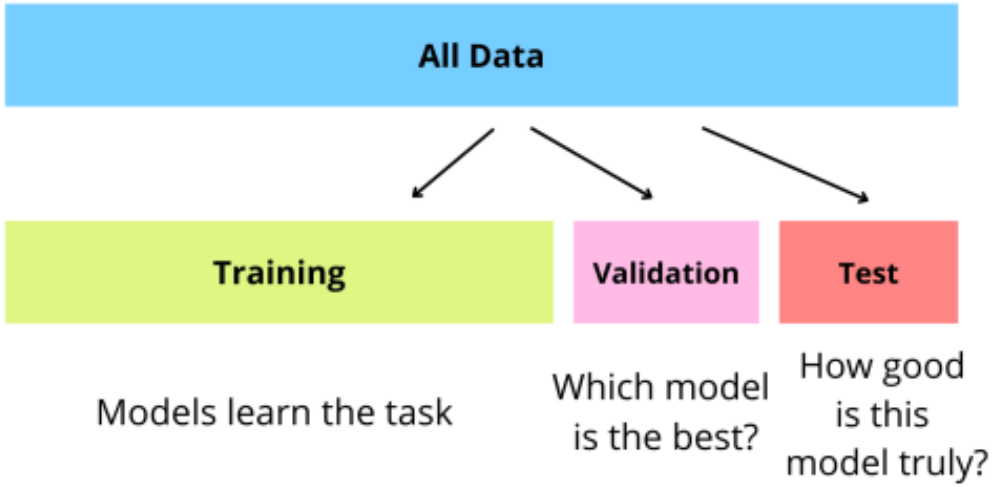
3.2 Running Experiments

- **Details matter!**
  - Modularize each step of experiment (e.g., into directory in -> directory out)
  - Name directories by parameters
    - E.g., `transformer-layer8-node512-dropout0.5-labelsmooth0.02`
  - Don't re-run directories that are already done
  - Tuning hyperparameters, learning rates, getting initialization right, etc. is often important to the successes of neural nets

3.3 Evaluation

Many publicly available datasets are released with a train/dev/test structure.

- You build (estimate/train) a model on a **training** set.
- You measure progress as you go on a **dev** set (development test set or validation set)
- Only at the end, you evaluate and present final numbers on a **test** set
  - Use the final test set extremely few times … ideally only once



Be careful about the distribution difference!

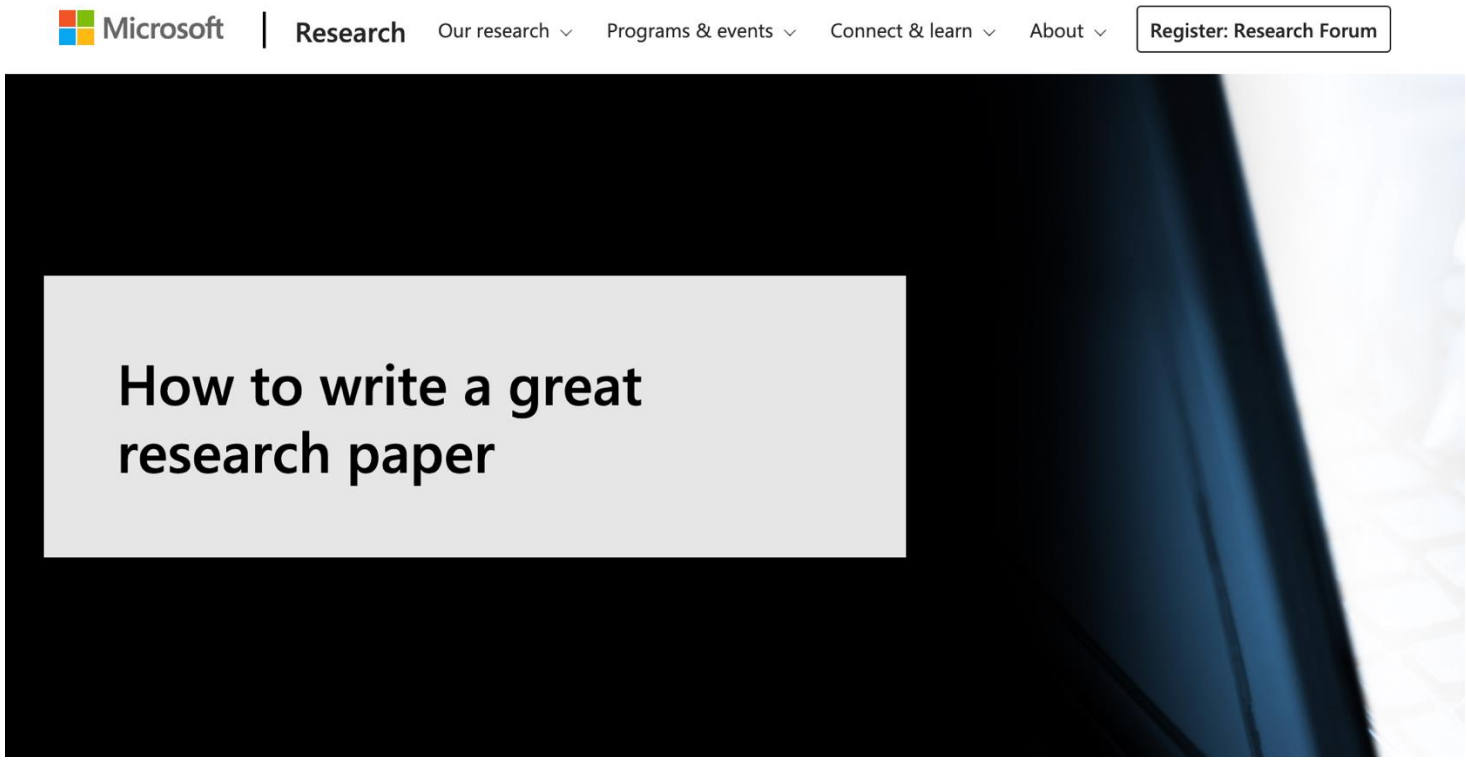
4. Report conclusions

The final project report should include:

1. Introduction & Problem Statement
   - What problem are you addressing? Why is it important (real-world motivation)?
2. Related work / background
   - Brief summary of existing work, models, and datasets; What is the gap?
3. Dataset Description
4. Methodology
5. Experiments and Results
6. Discussions and Analysis
   - Interpretation of results—what worked well, what didn't
   - Insights you learned from the experiments
7. Conclusion and Future Work
8. Reference and Appendix (e.g., contributions of each team member)

4. Report conclusions

- We are expecting a project report only, there is much more to be discussed for an academic paper writing
- Too much for a single class, but highly recommended (if you are interested in academia):



https://www.microsoft.com/en-us/research/academic-program/write-great-research-paper/