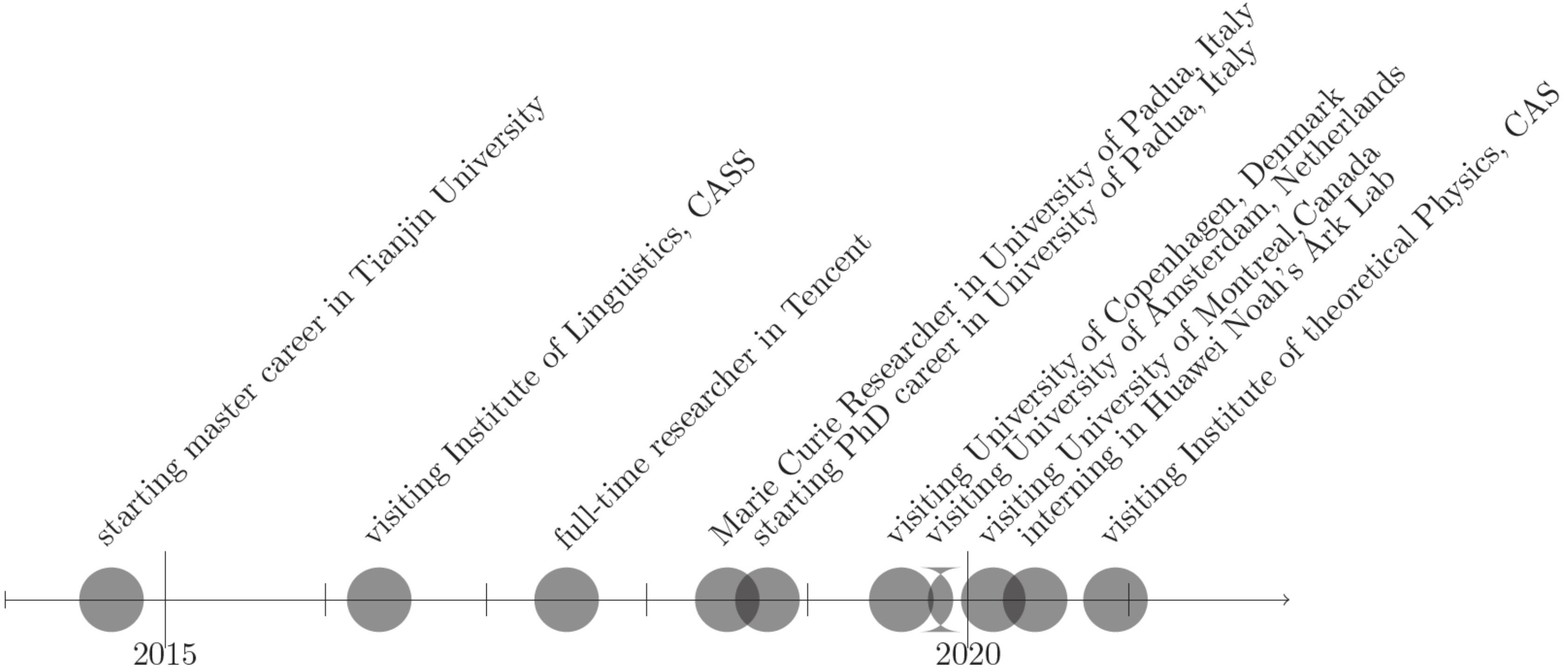
香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

CSC6052/5051/4100/DDA6307/ MDS5110
Natural Language Processing

Spring 2025
Benyou Wang
School of Data Science

About me



starting master career in Tianjin University

visiting Institute of Linguistics, CASS

full-time researcher in Tencent

Marie Curie Researcher in University of Padua, Italy

starting PhD career in University of Padua, Italy

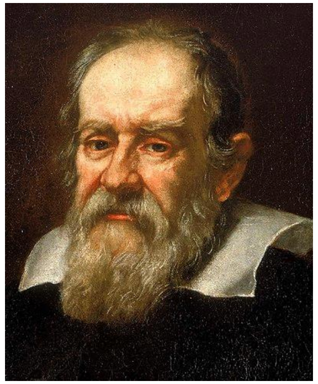
visiting University of Copenhagen, Denmark

visiting University of Amsterdam, Netherlands

visiting University of Montreal,Canada

interning in Huawei Noah's Ark Lab

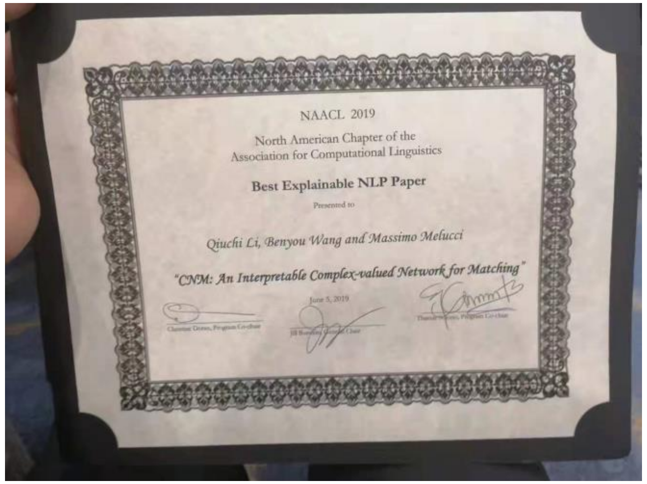
visiting Institute of theoretical Physics, CAS


Galileo Galilei

the "father of **modern physics**"
the "father of the scientific method"
the "father of modern science"

Alumni of University of Padua

Awards and honour



NAACL 2019

North American Chapter of the
Association for Computational Linguistics

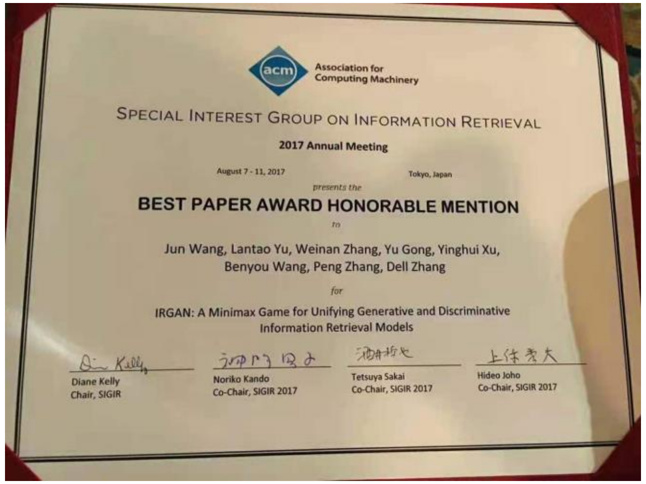
Best Explainable NLP Paper

Presented to

Qiuchi Li, Benyou Wang and Massimo Melucci

"CNM: An Interpretable Complex-valued Network for Matching"

June 5, 2019



Association for
Computing Machinery

SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL

2017 Annual Meeting

August 7 - 11, 2017 Tokyo, Japan

presents the

BEST PAPER AWARD HONORABLE MENTION

to

Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu,
Benyou Wang, Peng Zhang, Dell Zhang

for

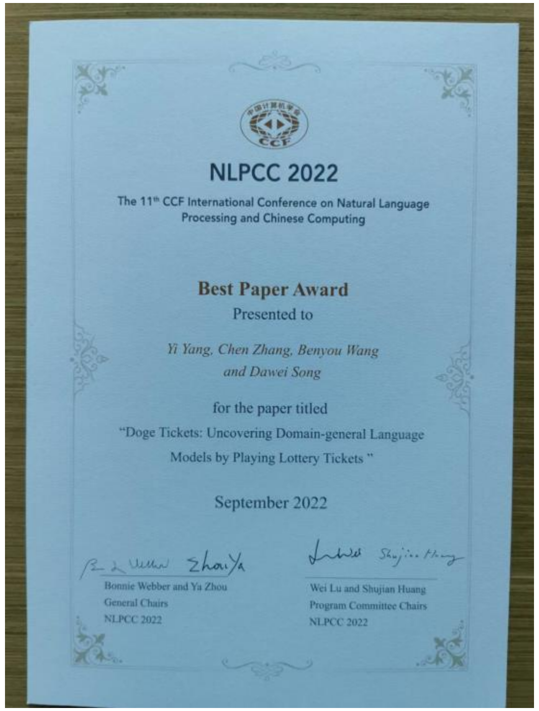
IRGAN: A Minimax Game for Unifying Generative and Discriminative
Information Retrieval Models

Diane Kelly
Chair, SIGIR

Noriko Kando
Co-Chair, SIGIR 2017

Tetsuya Sakai
Co-Chair, SIGIR 2017

Hideo Joho
Co-Chair, SIGIR 2017



NLPCC 2022

The 11th CCF International Conference on Natural Language
Processing and Chinese Computing

Best Paper Award

Presented to

*Yi Yang, Chen Zhang, Benyou Wang
and Dawei Song*

for the paper titled
"Doge Tickets: Uncovering Domain-general Language
Models by Playing Lottery Tickets "

September 2022

Bonnie Webber and Ya Zhou
General Chairs
NLPCC 2022

Wei Lu and Shujian Huang
Program Committee Chairs
NLPCC 2022

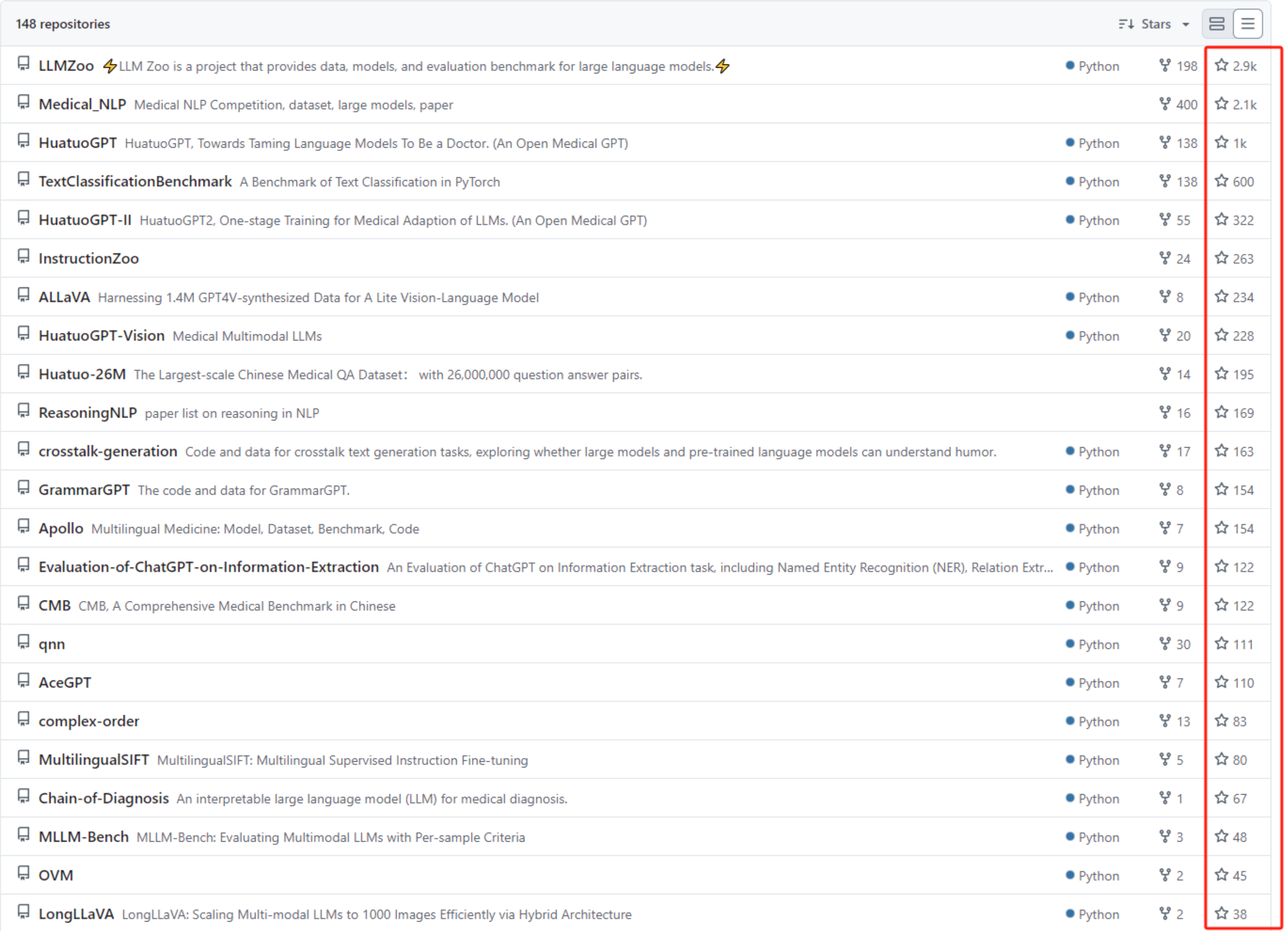
- **NLPCC 2022** Best Paper
- **ACM SIGIR** 2017 Best paper honourable mention. https://sigir.org/awards/best-paper-awards/
- **NAACL** 2019 best explainable NLP paper. https://naacl2019.org/blog/best-papers/
- EU Marie Curry researcher fellowship
- Huawei Spark award (华为火花奖)

Overall of our research (update every half an year)

- Medical Large language models
- LLMs for Math
  - Automatic theorem proof
  - AI Mathematical Olympiad
- Multi-modal LLMs
  - Speech
  - Image/video understanding
  - Video generation (SORA)
- Multilingual LLMs
  - AceGPT
- Agent and Human-machine interaction
- Efficient LLMs

- Applications
  - Math
  - Medical/financial domains
  - Multilingual extension
- LLMs
  - Text-only LLM
  - Vision language models
  - Speech language models
  - Video Generation (SORA)
- The foundations
  - Agent/HCI
  - LLM Evaluation
  - Efficiency
  - RLHF
  - Search

Our team - GitHub Homepage

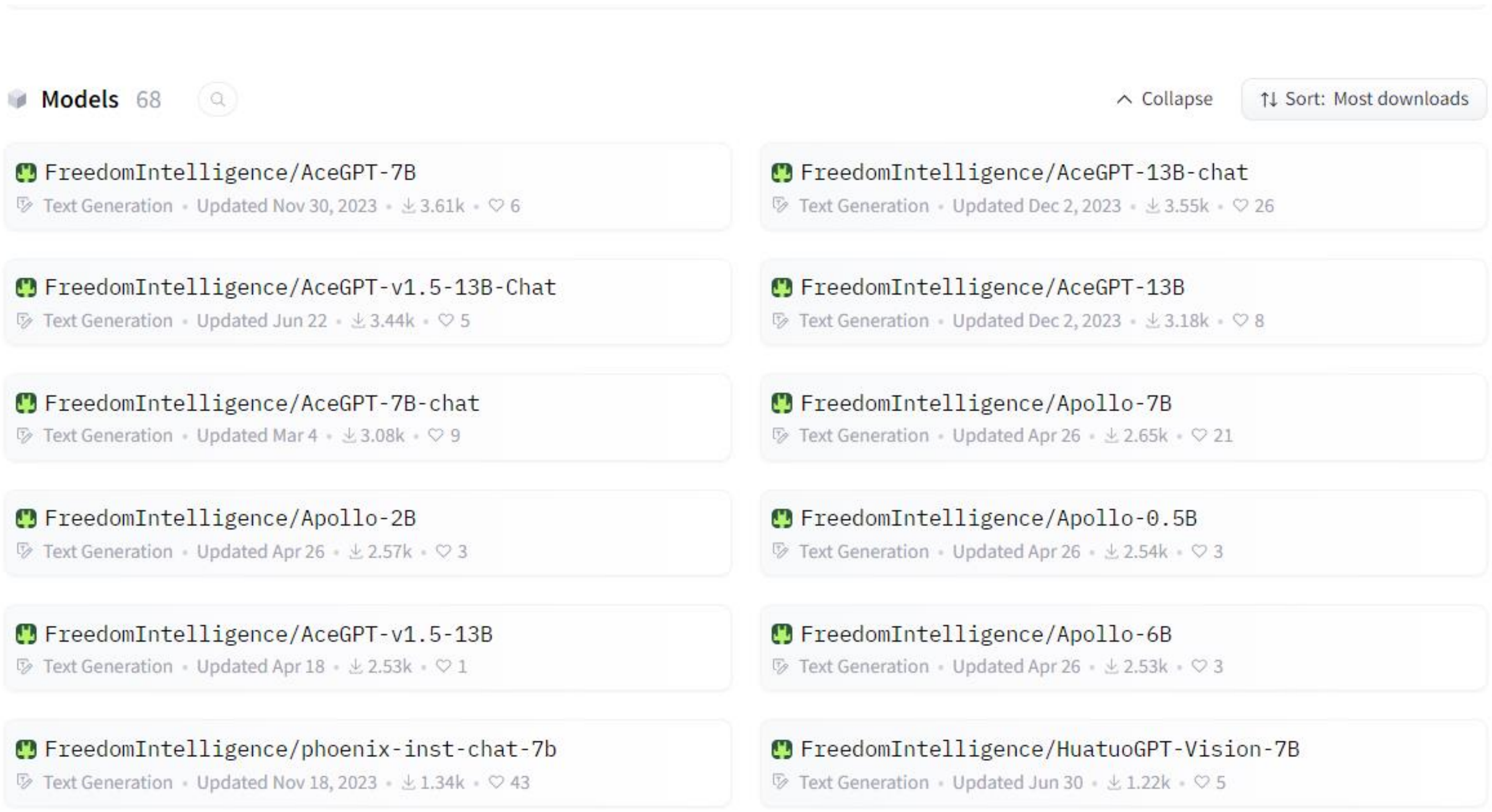


148 repositories

LLMZoo	⚡LLM Zoo is a project that provides data, models, and evaluation benchmark for large language models.⚡	Python	198	2.9k
Medical_NLP	Medical NLP Competition, dataset, large models, paper		400	2.1k
HuatuoGPT	HuatuoGPT, Towards Taming Language Models To Be a Doctor. (An Open Medical GPT)	Python	138	1k
TextClassificationBenchmark	A Benchmark of Text Classification in PyTorch	Python	138	600
HuatuoGPT-II	HuatuoGPT2, One-stage Training for Medical Adaption of LLMs. (An Open Medical GPT)	Python	55	322
InstructionZoo			24	263
ALLaVA	Harnessing 1.4M GPT4V-synthesized Data for A Lite Vision-Language Model	Python	8	234
HuatuoGPT-Vision	Medical Multimodal LLMs	Python	20	228
Huatuo-26M	The Largest-scale Chinese Medical QA Dataset： with 26,000,000 question answer pairs.		14	195
ReasoningNLP	paper list on reasoning in NLP		16	169
crosstalk-generation	Code and data for crosstalk text generation tasks, exploring whether large models and pre-trained language models can understand humor.	Python	17	163
GrammarGPT	The code and data for GrammarGPT.	Python	8	154
Apollo	Multilingual Medicine: Model, Dataset, Benchmark, Code	Python	7	154
Evaluation-of-ChatGPT-on-Information-Extraction	An Evaluation of ChatGPT on Information Extraction task, including Named Entity Recognition (NER), Relation Extr...	Python	9	122
CMB	CMB, A Comprehensive Medical Benchmark in Chinese	Python	9	122
qnn		Python	30	111
AceGPT		Python	7	110
complex-order		Python	13	83
MultilingualSIFT	MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning	Python	5	80
Chain-of-Diagnosis	An interpretable large language model (LLM) for medical diagnosis.	Python	1	67
MLLM-Bench	MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria	Python	3	48
OVM		Python	2	45
LongLLaVA	LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture	Python	2	38

https://github.com/orgs/FreedomIntelligence

Our team - HuggingFace Homepage



FreedomIntelligence/AceGPT-7B
Text Generation · Updated Nov 30, 2023 · ↓ 3.61k · ♡ 6

FreedomIntelligence/AceGPT-13B-chat
Text Generation · Updated Dec 2, 2023 · ↓ 3.55k · ♡ 26

FreedomIntelligence/AceGPT-v1.5-13B-Chat
Text Generation · Updated Jun 22 · ↓ 3.44k · ♡ 5

FreedomIntelligence/AceGPT-13B
Text Generation · Updated Dec 2, 2023 · ↓ 3.18k · ♡ 8

FreedomIntelligence/AceGPT-7B-chat
Text Generation · Updated Mar 4 · ↓ 3.08k · ♡ 9

FreedomIntelligence/Apollo-7B
Text Generation · Updated Apr 26 · ↓ 2.65k · ♡ 21

FreedomIntelligence/Apollo-2B
Text Generation · Updated Apr 26 · ↓ 2.57k · ♡ 3

FreedomIntelligence/Apollo-0.5B
Text Generation · Updated Apr 26 · ↓ 2.54k · ♡ 3

FreedomIntelligence/AceGPT-v1.5-13B
Text Generation · Updated Apr 18 · ↓ 2.53k · ♡ 1

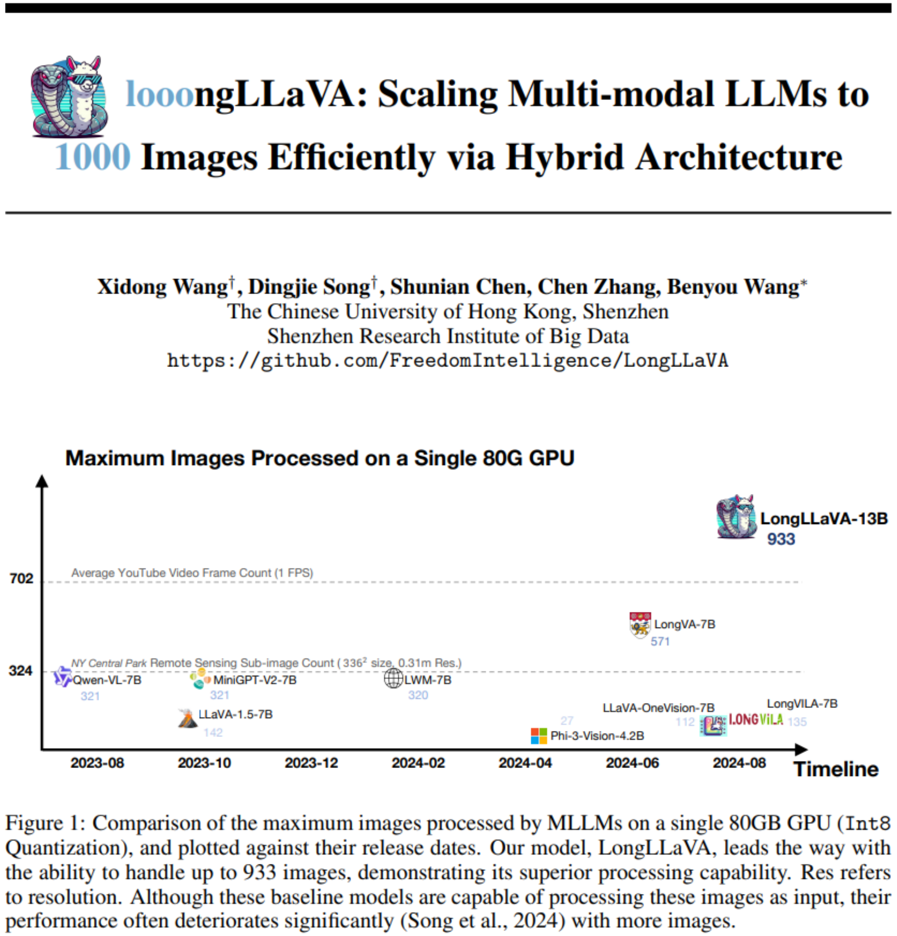
FreedomIntelligence/Apollo-6B
Text Generation · Updated Apr 26 · ↓ 2.53k · ♡ 3

FreedomIntelligence/phoenix-inst-chat-7b
Text Generation · Updated Nov 18, 2023 · ↓ 1.34k · ♡ 43

FreedomIntelligence/HuatuoGPT-Vision-7B
Text Generation · Updated Jun 30 · ↓ 1.22k · ♡ 5

https://huggingface.co/FreedomIntelligence

A Recent Study: LongLLaVA



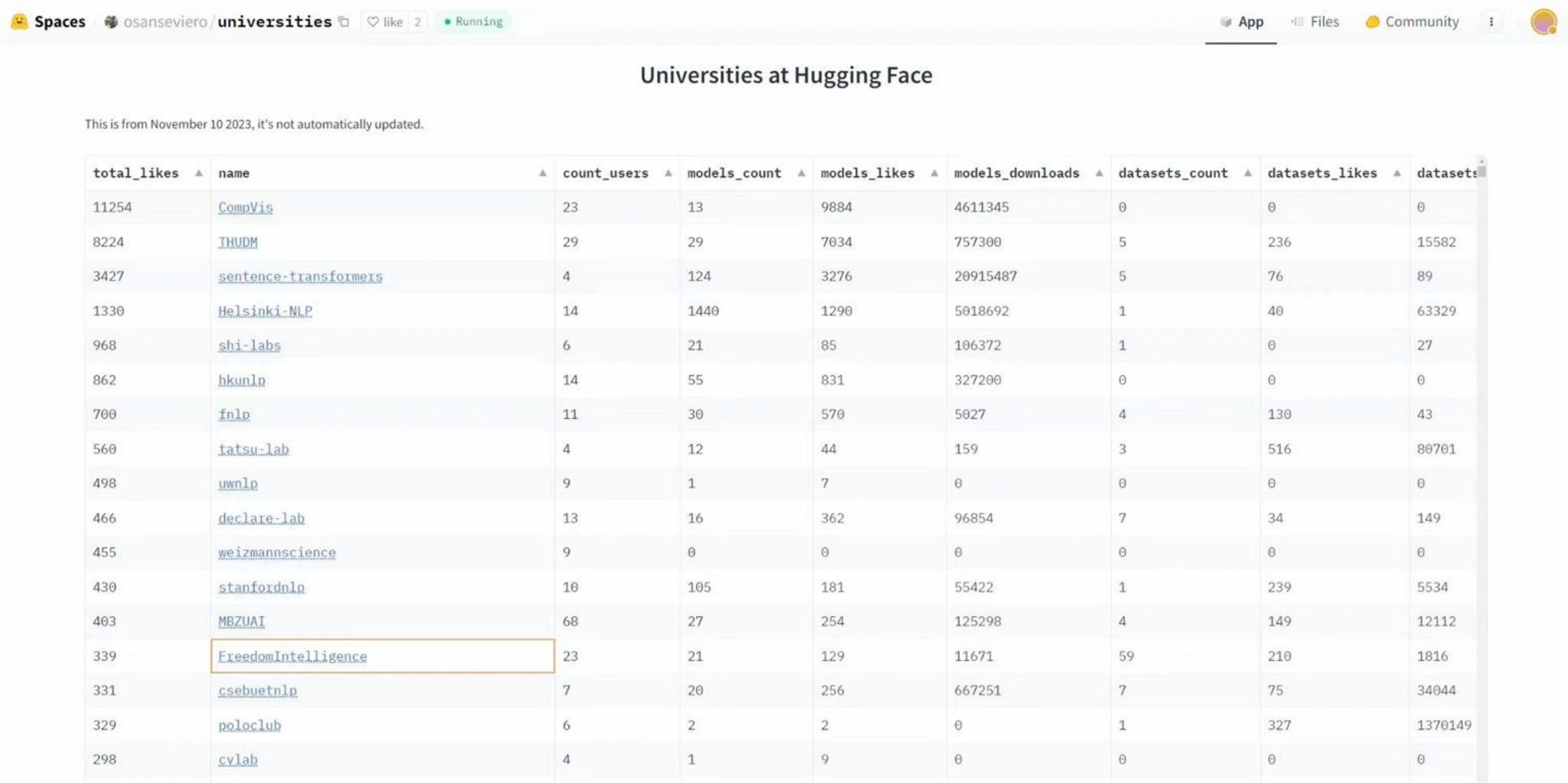
looongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture

Xidong Wang[†], Dingjie Song[†], Shunian Chen, Chen Zhang, Benyou Wang[*]
The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Institute of Big Data
`https://github.com/FreedomIntelligence/LongLLaVA`

Figure 1: Comparison of the maximum images processed by MLLMs on a single 80GB GPU (`Int8` Quantization), and plotted against their release dates. Our model, LongLLaVA, leads the way with the ability to handle up to 933 images, demonstrating its superior processing capability. Res refers to resolution. Although these baseline models are capable of processing these images as input, their performance often deteriorates significantly (Song et al., 2024) with more images.

https://arxiv.org/pdf/2409.02889

We rank 4th in Greater China



Universities at Hugging Face

This is from November 10 2023, it's not automatically updated.

total_likes	name	count_users	models_count	models_likes	models_downloads	datasets_count	datasets_likes	datasets
11254	CompVis	23	13	9884	4611345	0	0	0
8224	THUDM	29	29	7034	757300	5	236	15582
3427	sentence-transformers	4	124	3276	20915487	5	76	89
1330	Helsinki-NLP	14	1440	1290	5018692	1	40	63329
968	shi-labs	6	21	85	106372	1	0	27
862	hkunlp	14	55	831	327200	0	0	0
700	fnlp	11	30	570	5027	4	130	43
560	tatsu-lab	4	12	44	159	3	516	80701
498	uwnlp	9	1	7	0	0	0	0
466	declare-lab	13	16	362	96854	7	34	149
455	weizmannscience	9	0	0	0	0	0	0
430	stanfordnlp	10	105	181	55422	1	239	5534
403	MBZUAI	68	27	254	125298	4	149	12112
339	FreedomIntelligence	23	21	129	11671	59	210	1816
331	csebuetnlp	7	20	256	667251	7	75	34044
329	poloclub	6	2	2	0	1	327	1370149
298	cvlab	4	1	9	0	0	0	0

https://huggingface.co/spaces/osanseviero/universities
https://twitter.com/osanseviero/status/1723229014100255011

Our team - Join us as a Research Assistant!

What You Will Receive:

- Access to GPU **computing resources** and utilization of abundant **GPT APIs**.
- More frequent **communication** within our research team (also more pressure).
- Potential for research **publications** (sometimes coauthorship).
- The possibility of a part-time **contract** with a salary (also with office space).
- **Internship** recommendation (Microsoft, Amazon, BAT, Huawei, Bytedance)

What We Expect From You:

- A commitment of at least 15 hours per week for full-time engagement.
- satisfied programming skills.

If you find this opportunity intriguing, please reach out to Xidong (223040239@link.cuhk.edu.cn) for further details.

Contents

- Philosophy of this course
- Large language models?
- Introduction to ChatGPT

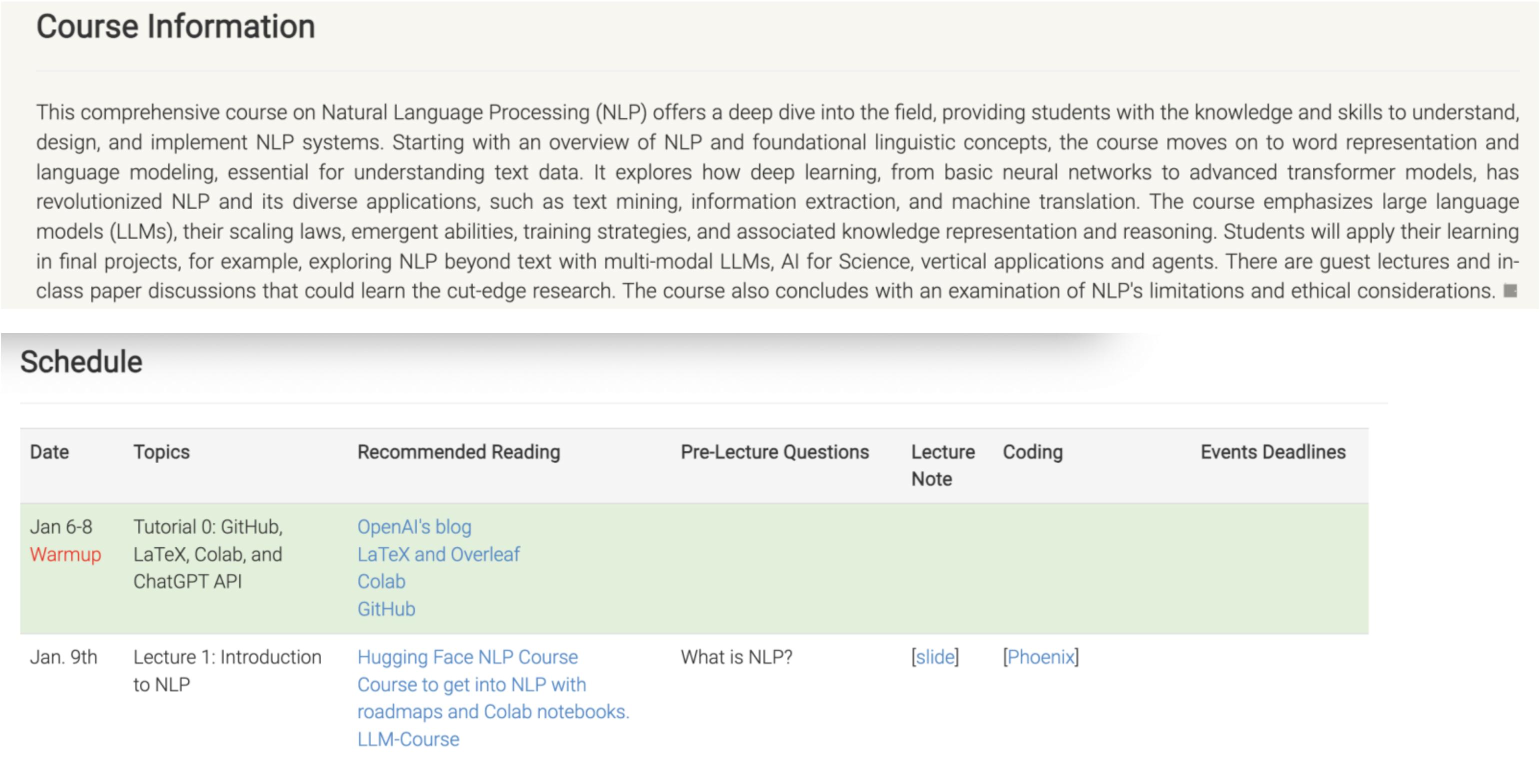
Logistics

- Instructor: Benyou Wang
- Teaching assistant:
  Shunian Chen (Leading TA)
  Xidong Wang, Juhao Liang, Ke Ji ,Rui Huang, Yuqi Fei

- Location: Teaching B 202
- Meetings: Thursday/Friday 10:30AM - 11:50AM
- Office hours:
  - Benyou Wang: Thursday 4:00-5:00 PM at Daoyuan Building 504A. (Email: wangbenyou@cuhk.edu.cn)
  - Shunian Chen: Friday 2:00-3:00 PM at Teaching Building D 412.(Email: shunianchen@link.cuhk.edu.cn)
  - Xidong Wang: Wednesday 7:30-8:30 PM. Teaching Building D 412. (Email: 223040239@link.cuhk.edu.cn)
  - Rui Huang: Tuesday 1:00-2:00 PM. Teaching Building A 521.(Email: rayhuang@cuhk.edu.cn)

Logistics

- [Official Website Link (https://nlp-course-cuhksz.github.io/)](https://nlp-course-cuhksz.github.io/)



Course Information

This comprehensive course on Natural Language Processing (NLP) offers a deep dive into the field, providing students with the knowledge and skills to understand, design, and implement NLP systems. Starting with an overview of NLP and foundational linguistic concepts, the course moves on to word representation and language modeling, essential for understanding text data. It explores how deep learning, from basic neural networks to advanced transformer models, has revolutionized NLP and its diverse applications, such as text mining, information extraction, and machine translation. The course emphasizes large language models (LLMs), their scaling laws, emergent abilities, training strategies, and associated knowledge representation and reasoning. Students will apply their learning in final projects, for example, exploring NLP beyond text with multi-modal LLMs, AI for Science, vertical applications and agents. There are guest lectures and in-class paper discussions that could learn the cut-edge research. The course also concludes with an examination of NLP's limitations and ethical considerations. ■

Schedule

Date	Topics	Recommended Reading	Pre-Lecture Questions	Lecture Note	Coding	Events Deadlines
Jan 6-8 Warmup	Tutorial 0: GitHub, LaTeX, Colab, and ChatGPT API	OpenAI's blog LaTeX and Overleaf Colab GitHub				
Jan. 9th	Lecture 1: Introduction to NLP	Hugging Face NLP Course Course to get into NLP with roadmaps and Colab notebooks. LLM-Course	What is NLP?	[slide]	[Phoenix]	

- Official Wechat Group

群聊: NLP 2025 Spring

该二维码7天内(1月15日前)有效，重新进入将更新

Course Structure

- This is **an advanced graduate course** and we will be offering a deep dive into the field and equip students with the comprehensive the knowledge and skills to understand, design, and implement NLP systems.
- All the students are expected to come to the class regularly and participate in discussion
- Prerequisites:
  - Proficiency in LaTex: All the reports need to be written by using LaTex. A template will be provided. If you are not familiar with LaTex, please learn from the tutorial in advance.
  - Proficiency in GitHub: All the source codes need to be submitted in GitHub.
  - Proficiency in Python: All the assignments will be in Python (using Numpy and PyTorch).
  - Basic machine learning knowledge: It is possible to take this course without any machine learning knowledge, however, the course will be easier if you have foundations of machine learning.

Course Structure (tentative)

- Introduction to NLP
- Linguistics and Word Embeddings
- Language Models
- Deep Learning in NLP
- Large Language Models (LLMs)
- Prompt Engineering
- LLM Agents
- Training Large Language Models
- Final Project Introduction and Research Sharing
- Multimodal Learning
- LLM Reasoning and Guest Lecture

Components and Grading (Undergraduate)

- Assignments (50%)
  - Assignment 1 (15%): Training word vector.
  - Assignment 2 (15%): Using API for testing prompt engineering and LLM agents.
  - Assignment 3 (20%): Training NLP model with SFT and RLHF.

  All assignments need a report and code attachment if it has coding. See the relevant evaluation criterion as the final project.

- Final project (40%)

  The project could be done by a group but each individual is separately evaluated.You need to write a project report (max 6 pages) for the final project. You are also expected to make a project poster presentation. After the final project deadline, feel free to make your project open source; we appreciate if you acknowledge this course.

- Participation (10%)

Components and Grading (Post-Graduate)

- Assignments (40%)
  - Assignment 1 (15%): Training word vector.
  - Assignment 2 (15%): Using API for testing prompt engineering and LLM agents.
  - Assignment 3 (20%): Training NLP model with SFT and RLHF.

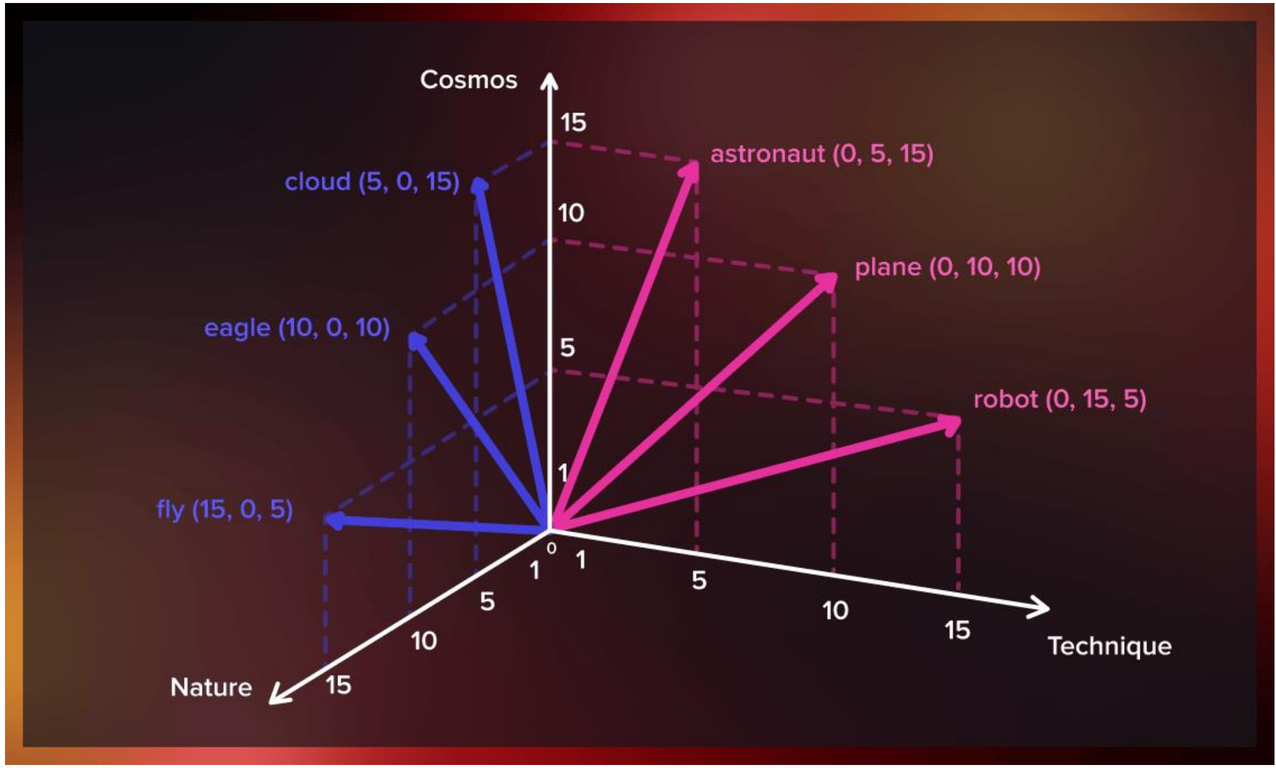
  All assignments need a report and code attachment if it has coding. See the relevant evaluation criterion as the final project.

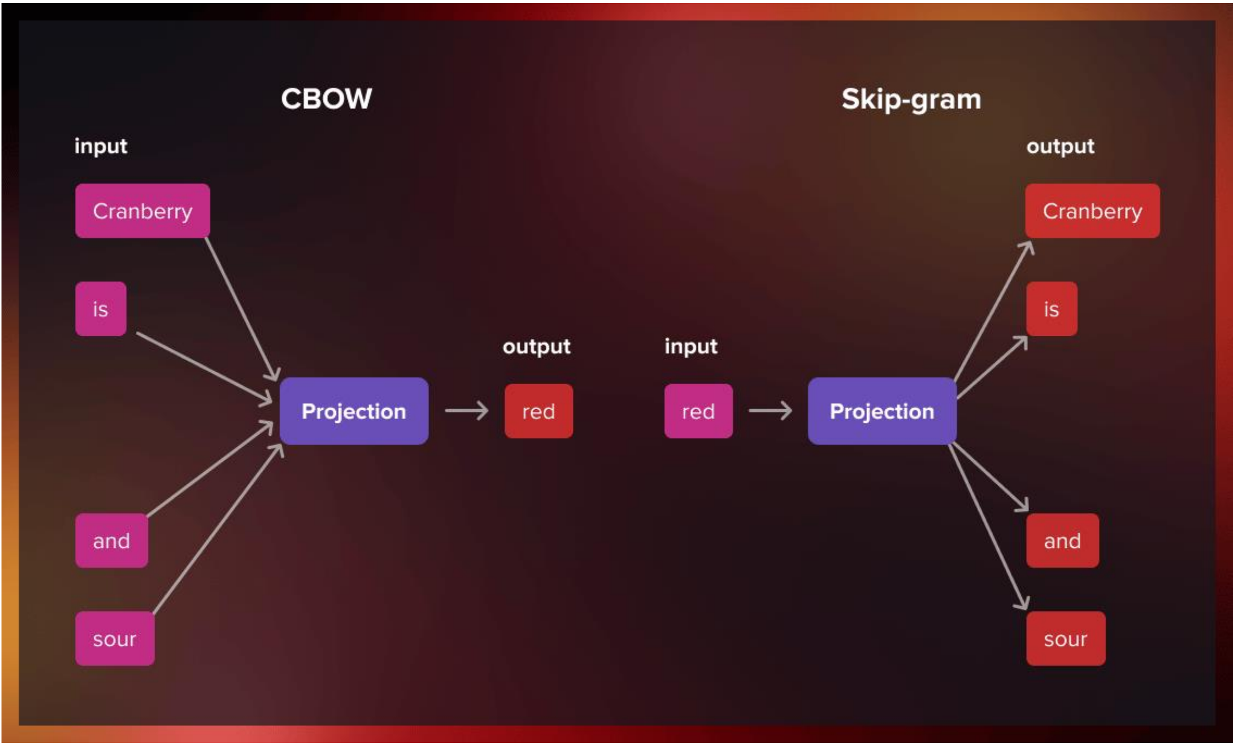
- Final project (55%)

  You need to write a project report (max 6 pages) for the final project. You are also expected to make a project poster presentation. After the final project deadline, feel free to make your project open source; we appreciate if you acknowledge this course.

- Participation (5%)

Assignments 1: Training word vector.

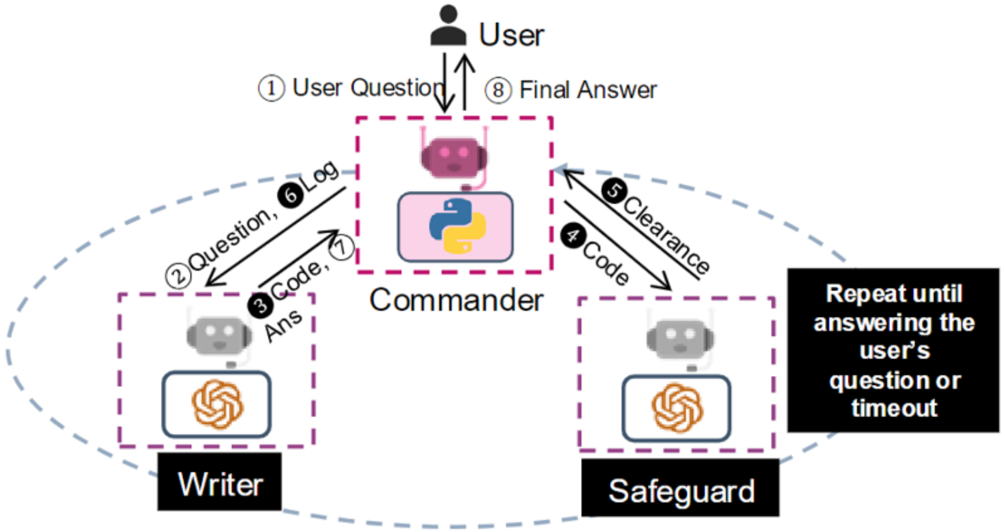




Task description: This task aiming at deepening your practical understanding of word embedding.

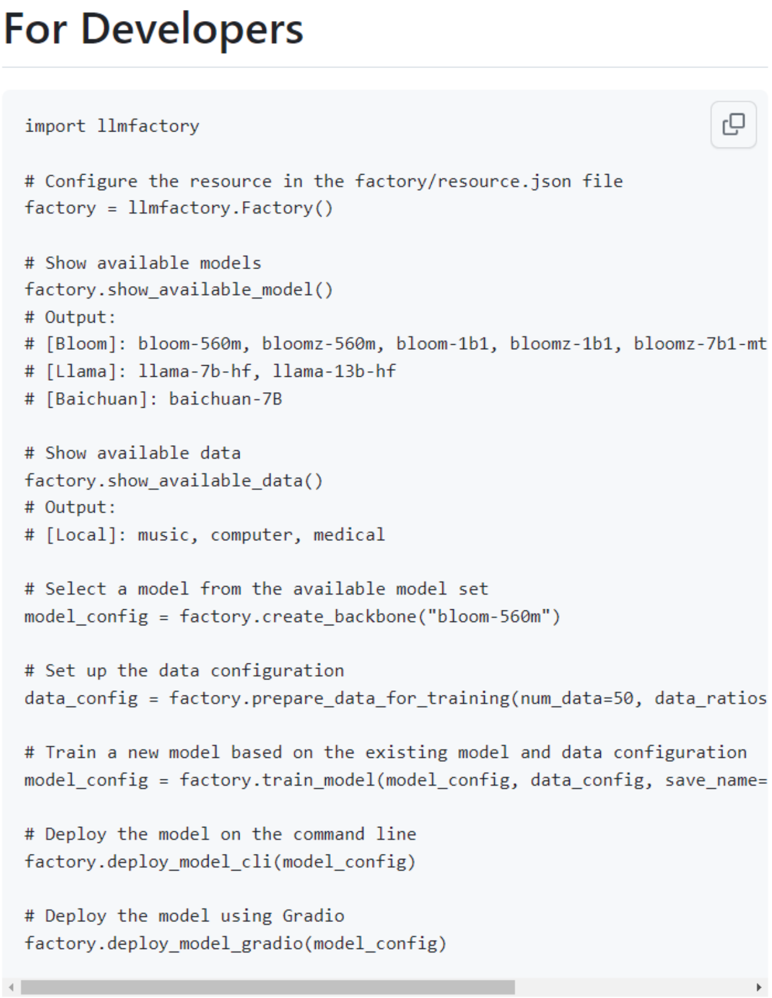
Source: https://serokell.io/blog/word2vec

Assignment 2: Using API for Prompt Engineering and LLM Agents



- Task Description: Explore prompt engineering using API. Utilizing existing LLM Agent frameworks to achieve practical goals.
- Requirement: Submit a report with code showcasing your practical skills in using LLM agents. More details released later.

Assignments 3: Training NLP model with SFT and RLHF



For Developers

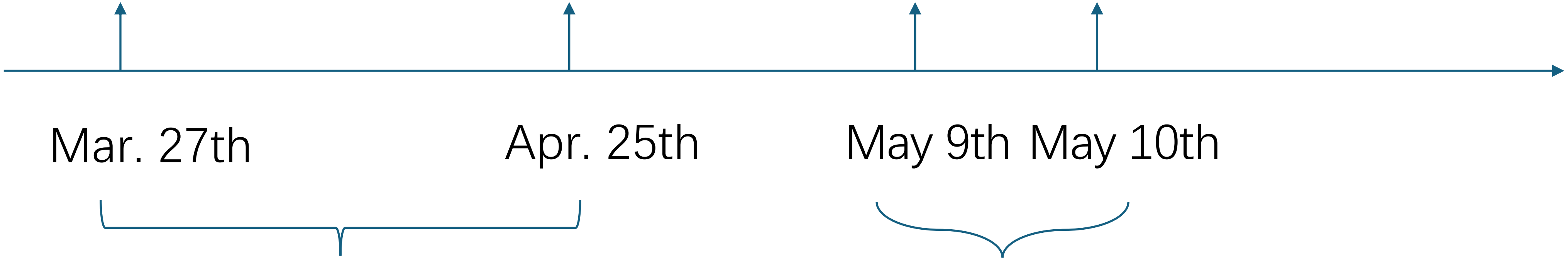
```
import llmfactory

# Configure the resource in the factory/resource.json file
factory = llmfactory.Factory()

# Show available models
factory.show_available_model()
# Output:
# [Bloom]: bloom-560m, bloomz-560m, bloom-1b1, bloomz-1b1, bloomz-7b1-mt
# [Llama]: llama-7b-hf, llama-13b-hf
# [Baichuan]: baichuan-7B

# Show available data
factory.show_available_data()
# Output:
# [Local]: music, computer, medical

# Select a model from the available model set
model_config = factory.create_backbone("bloom-560m")

# Set up the data configuration
data_config = factory.prepare_data_for_training(num_data=50, data_ratios

# Train a new model based on the existing model and data configuration
model_config = factory.train_model(model_config, data_config, save_name=

# Deploy the model on the command line
factory.deploy_model_cli(model_config)

# Deploy the model using Gradio
factory.deploy_model_gradio(model_config)
```

https://github.com/FreedomIntelligence/LLMFactory

Final project

- Students complete a research project in teams of 1-3
- Draft proposal deadline: Apr. 17th 11:59pm (simply explain your idea in one page)
- Final proposal deadline: Apr. 25th11:59pm (TAs will provide suggestions for the revision!)

- In-class presentation: Apr. 25th and 27th. and May. 9th -10nd
- Final poster deadline: Apr. 24th 17:00pm
- Final paper deadline: May. 2nd(final date)



Mar. 27th

Apr. 25th

May 9th

May 10th

Better finish your project within

Busy

What can we do for the final project?

Philosophy :

- select one of given proposals (you could do nothing but wait for the release of proposals)**.**
- **OR** submit a customed project with a proposal that needs to be approved

- **Call for Project Proposals (optional):** (Deadline tentatively Apr. 17th 11:59pm .).
    - Choose an NLP research topic of interest or select from provided options.
    - The team members should be decided
- **Who could submit the proposal**
    - The students in this course (You could decide whether the submitted proposal could be used by other teams if approved)
    - Research scientists or engineers in a company
    - RAs and Phd students in our research team
    - Falcuty members in our university (your supervisors for example)
    - Actually, everyone is welcome, we reserve the rights to reject a proposal.

Final project

Typical projects (we will release a detailed list later):

Hugging Face https://huggingface.co/models

1. Train or fine-tune a medium-sized language model (e.g., T5, Bloom, TinyLLaMA, Baichuan, LLaMA) yourself for any problem of your interest. Check out HuggingFace's model hub!
2. Evaluate one of the largest language models (e.g., ChatGPT/GPT4) and understand their capabilities, limitations and risks;
3. An Agent System (math/financial/medical/legal)
4. An APP that is for the campus, hospital, etc.
5. An survey that involving LLMs
6. A research paper involving LLMs (please acknowledge this course if possible)

Note: You might get computing resources to train 10B+ model if Tas/instructor like your proposal

More project ideas will be presented during the lectures and all creative ideas or research topics are encouraged for further discussion.

https://openai.com/api/ https://opt.alpa.ai

Final project

Report of final project should be publicly released, otherwise please specify reasons.

- released in our website
- released in ArXiv if quality is high; do not preprint low-quality paper there. If you cannot access the quality, you could consult your supervisor or the teaching team.

Consider Github to maintain your code (link Github repo in the report) if coding needed

A objective is could be **earning 100+ GitHub stars for your repo.**

Consider uploading your model in HuggingFace, serving it there is possible.

Note: You might get computing resources to train 10B+ model if instructor like your proposal

https://openai.com/api/ https://opt.alpa.ai

Team work

Team size: Students may do final projects solo, or in teams of up to 3 people. Please specify reasons to do project with more than 3 people, we reserve the rights to reject it

Team work is encouraged: We strongly recommend you do the final project in a team. Larger teams are expected to do correspondingly larger projects, and you should only form a 3-person team if you are planning to do an ambitious project where every team member will have a significant contribution.

Contribution: In the final report we ask for a statement of what each team member contributed to the project. Team members will typically get the same grade, but we may differentiate in extreme cases of unequal contribution. **You can contact us in confidence in the event of unequal contribution**.

External collaborators: You can work on a project that has external (non course student) collaborators, but you must make it clear in your final report which parts of the project you are extractly contributing.

Mentors: Every custom project team has a mentor, who gives feedback and advice during the project. A project may have an external (i.e., not in our teaching team) mentor; otherwise, we will assign a TA or any research staff in our research team to custom project teams after project proposals.

Computing resources: We will use Kaggle computing resources, later we will share the detail.

Final project assessment

Final report assessment （45%）

  TA review

Poster presentation （10%）

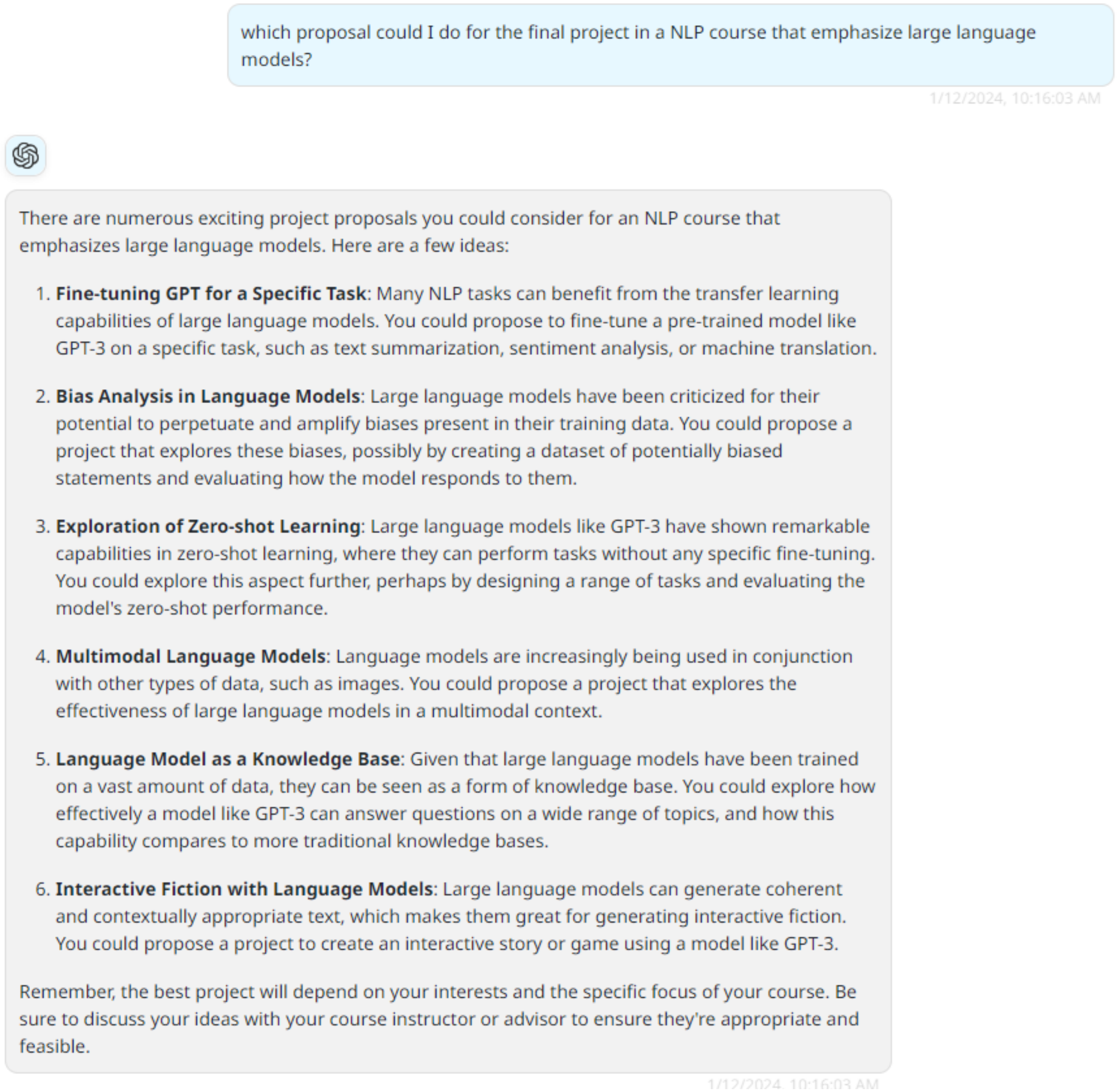
  Cross review between students + TA/instructor review

Submit your **Final project** to a **conference**

- Students are encouraged to submit papers to ACL ARR;
  - https://openreview.net/group?id=aclweb.org/ACL
- Submission Deadlines (every two months)
  - 2024 Oct 15th
  - 2024 Dec 15th
  - **2025 Feb 15th**

- TAs might help upon their interest. They might also refer you to other students in this group

How to better prepare a final project in NLP?



which proposal could I do for the final project in a NLP course that emphasize large language models?

1/12/2024, 10:16:03 AM

There are numerous exciting project proposals you could consider for an NLP course that emphasizes large language models. Here are a few ideas:

1. **Fine-tuning GPT for a Specific Task**: Many NLP tasks can benefit from the transfer learning capabilities of large language models. You could propose to fine-tune a pre-trained model like GPT-3 on a specific task, such as text summarization, sentiment analysis, or machine translation.
2. **Bias Analysis in Language Models**: Large language models have been criticized for their potential to perpetuate and amplify biases present in their training data. You could propose a project that explores these biases, possibly by creating a dataset of potentially biased statements and evaluating how the model responds to them.
3. **Exploration of Zero-shot Learning**: Large language models like GPT-3 have shown remarkable capabilities in zero-shot learning, where they can perform tasks without any specific fine-tuning. You could explore this aspect further, perhaps by designing a range of tasks and evaluating the model's zero-shot performance.
4. **Multimodal Language Models**: Language models are increasingly being used in conjunction with other types of data, such as images. You could propose a project that explores the effectiveness of large language models in a multimodal context.
5. **Language Model as a Knowledge Base**: Given that large language models have been trained on a vast amount of data, they can be seen as a form of knowledge base. You could explore how effectively a model like GPT-3 can answer questions on a wide range of topics, and how this capability compares to more traditional knowledge bases.
6. **Interactive Fiction with Language Models**: Large language models can generate coherent and contextually appropriate text, which makes them great for generating interactive fiction. You could propose a project to create an interactive story or game using a model like GPT-3.

Remember, the best project will depend on your interests and the specific focus of your course. Be sure to discuss your ideas with your course instructor or advisor to ensure they're appropriate and feasible.

1/12/2024, 10:16:03 AM

Define the learning objectives:

- **Knowledge**: a) Students will understand basic concepts and principles of NLP; b) Students could effectively use NLP for daily study, work and research; and c) Students will know which tasks NLP are suitable to solve and which are not.
- **Skills**: a) Students could train NLP models following a complete pipeline and b) Students could call ChatGPT API for daily usage in study, work and research.
- **Valued/Attitude**: a) Students will appreciate the importance of data; b) Students will tend to use data-driven paradigm to solve problems; and c) Students will be aware of the limitations and risks of using ChatGPT.

Select appropriate resources:

- **Recent ArXiv papers**
  - (People share daily ArXiv papers in Twitter)
- **GitHub**
  - (popular GitHub means a lot)
- **HuggingFace**
  - (New models and datasets)
  - https://huggingface.co/papers (hot papers here)
- **Blogs**
  - (from Open AI and famous guys, Lilian Weng, Yao Fu, Jianlin Su)

Design engaging lectures:

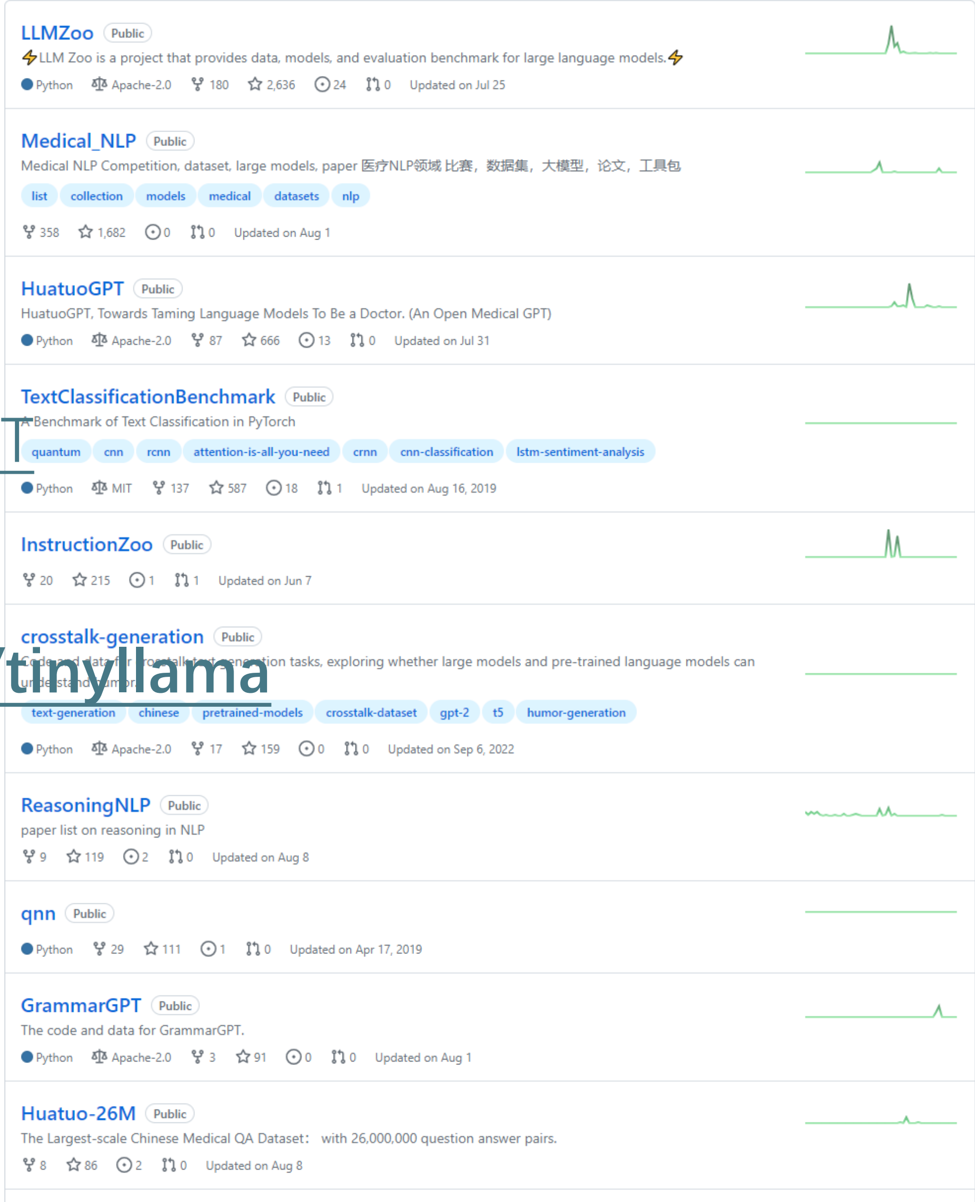
- Discussions in the end of each lecture
- In-class presentation
- Interrupting me whenever needed
- **Make friends with the instructor and TAs**

Provide hands-on practice:

Github Repositories

- **nanoGPT** https://github.com/karpathy/nanoGPT
- **minGPT** https://github.com/karpathy/minGPT
- **Llama2.c** https://github.com/karpathy/llama2.c
- **TinyLLaMA** **https://github.com/eivindbohler/tinyllama**

- HautuoGPT
- GPT review
- GPT API
- LLMZoo
- LLMFactory



LLMZoo Public
⚡LLM Zoo is a project that provides data, models, and evaluation benchmark for large language models.⚡
Python Apache-2.0 180 ☆ 2,636 24 0 Updated on Jul 25

Medical_NLP Public
Medical NLP Competition, dataset, large models, paper 医疗NLP领域 比赛，数据集，大模型，论文，工具包
list collection models medical datasets nlp
358 ☆ 1,682 0 0 Updated on Aug 1

HuatuoGPT Public
HuatuoGPT, Towards Taming Language Models To Be a Doctor. (An Open Medical GPT)
Python Apache-2.0 87 ☆ 666 13 0 Updated on Jul 31

TextClassificationBenchmark Public
A Benchmark of Text Classification in PyTorch
quantum cnn rcnn attention-is-all-you-need crnn cnn-classification lstm-sentiment-analysis
Python MIT 137 ☆ 587 18 1 Updated on Aug 16, 2019

InstructionZoo Public
20 ☆ 215 1 1 Updated on Jun 7

crosstalk-generation Public
Code and data for crosstalk text generation tasks, exploring whether large models and pre-trained language models can understand humor.
text-generation chinese pretrained-models crosstalk-dataset gpt-2 t5 humor-generation
Python Apache-2.0 17 ☆ 159 0 0 Updated on Sep 6, 2022

ReasoningNLP Public
paper list on reasoning in NLP
9 ☆ 119 2 0 Updated on Aug 8

qnn Public
Python 29 ☆ 111 1 0 Updated on Apr 17, 2019

GrammarGPT Public
The code and data for GrammarGPT.
Python Apache-2.0 3 ☆ 91 0 0 Updated on Aug 1

Huatuo-26M Public
The Largest-scale Chinese Medical QA Dataset： with 26,000,000 question answer pairs.
8 ☆ 86 2 0 Updated on Aug 8

https://github.com/orgs/FreedomIntellig

Foster collaboration and discussion:

- You own the copyright of your own project if our teaching team do not have a substantial contribution. Otherwise please acknowledge us.
- You are welcome to have discussions with our teaching team.
- Students are encouraged for collaboration and discussions.

Seek feedback and iterate:

- Tell us if you have any suggestions about this course
- We will continue polishing this course.

what is in our course

- Very basics of NLP (most old NLP techniques are not that practical now)
- Large Langauge models (training and beyond)
- Data engineering
- Prompt engineering
- NLP applications
- Future tendency NLP

what is not in our course

- How to do basic coding
- Machine learning
- How to understand the mathematical machanmisn of NLP models

- *SORA might not introduced*

Use ChatGPT easily

Check https://gpt.cuhk.edu.cn

A break!

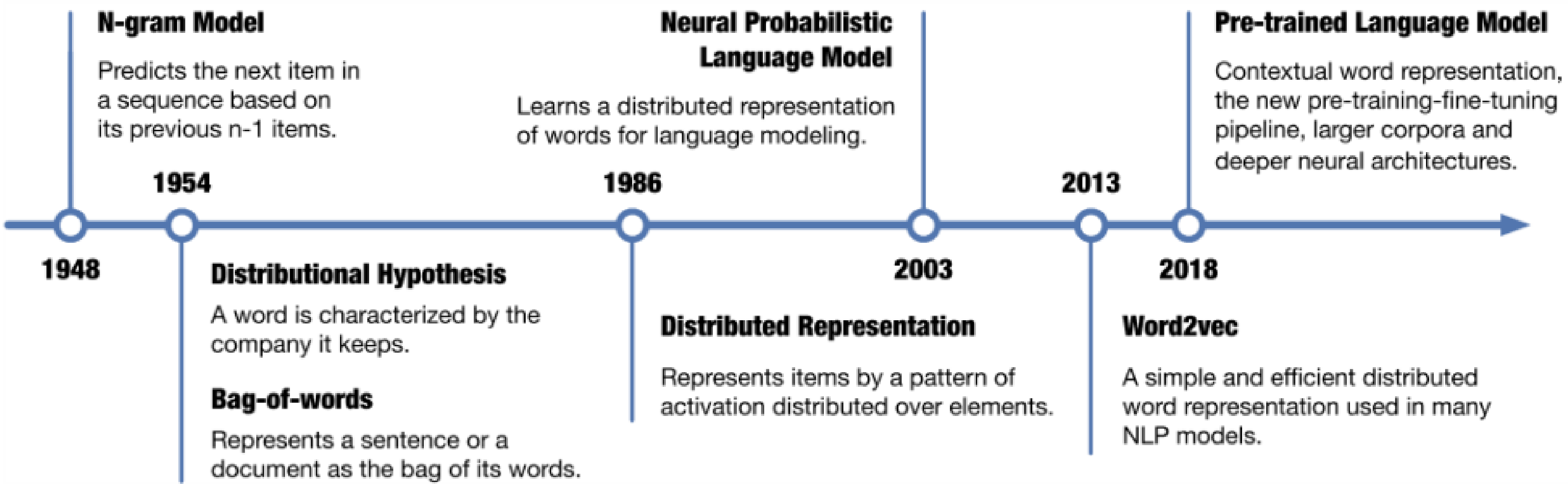
Contents

- Philosophy of this course
- **Large language models**
- Introduction to ChatGPT

What are Large Language models (LLMs)?

Background

- language model



1948

N-gram Model

Predicts the next item in a sequence based on its previous n-1 items.

1954

Distributional Hypothesis

A word is characterized by the company it keeps.

Bag-of-words

Represents a sentence or a document as the bag of its words.

1986

Distributed Representation

Represents items by a pattern of activation distributed over elements.

2003

Neural Probabilistic Language Model

Learns a distributed representation of words for language modeling.

2013

Word2vec

A simple and efficient distributed word representation used in many NLP models.

2018

Pre-trained Language Model

Contextual word representation, the new pre-training-fine-tuning pipeline, larger corpora and deeper neural architectures.

Liu et al., Representation Learning for Natural Language Processing, Springer, 2020

What is language modeling?

A **language model** assigns a probability to a N-gram

$$f : V^n \rightarrow R^+$$

What is language modeling?

A **language model** assigns a probability to a N-gram

$$f : V^n \rightarrow R^+$$

Sfklkljf fskjhfkjsh kjfs fs kjhkjhs fsjhfkshkjfh

Low probability

ChatGPT is all you need

high probability

What is language modeling?

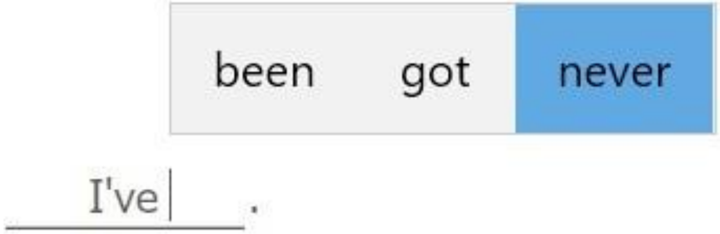
A **language model** assigns a probability to a N-gram

$$f : V^n \rightarrow R^+$$

A **conditional language model** assigns a probability of a word given some conditioning context

$$g : (V^{n-1}, V) \rightarrow R^+$$

And $p(w_n|w_1 \cdots$ $= \frac{f(w_1 \cdots w_n)}{f(w_1 \cdots w_{n-1})}$



I've ______.

What is language modeling?

A **language model** assigns a probability to a N-gram

$$f: V^n \rightarrow R^+$$

A **conditional language model** assigns a probability of a word given some conditioning context

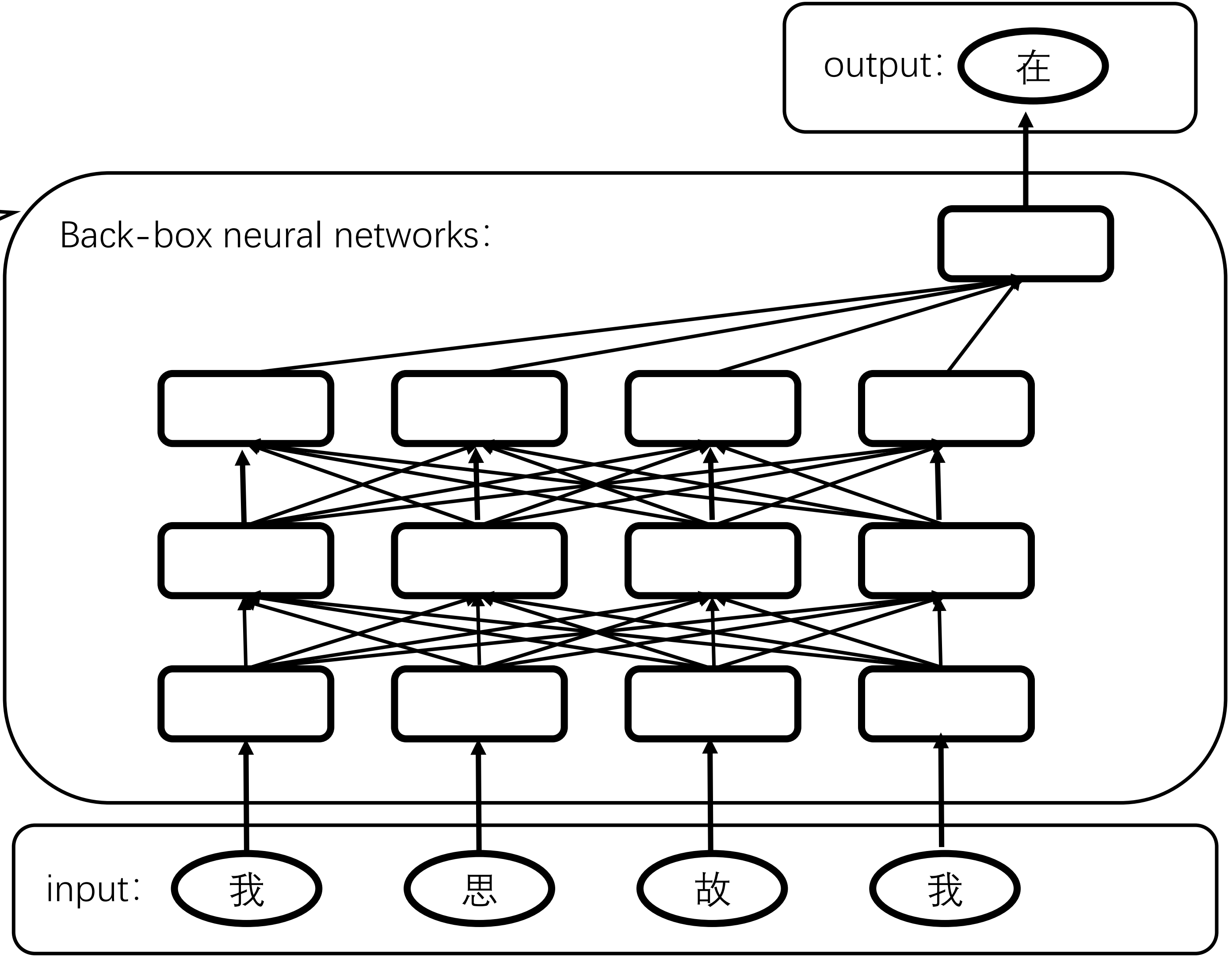
$$g: (V^{n-1}, V) \rightarrow R^+$$

And $\boldsymbol{p(w_n|w_1 \cdots w_{n-1})} = g(w_1 \cdots w_{n-1}, w) = \frac{f(w_1 \cdots w_n)}{f(w_1 \cdots w_{n-1})}$

$\boldsymbol{p(w_n|w_1 \cdots w_{n-1})}$ is the foundation of **modern large language models** (GPT, ChatGPT, etc.)

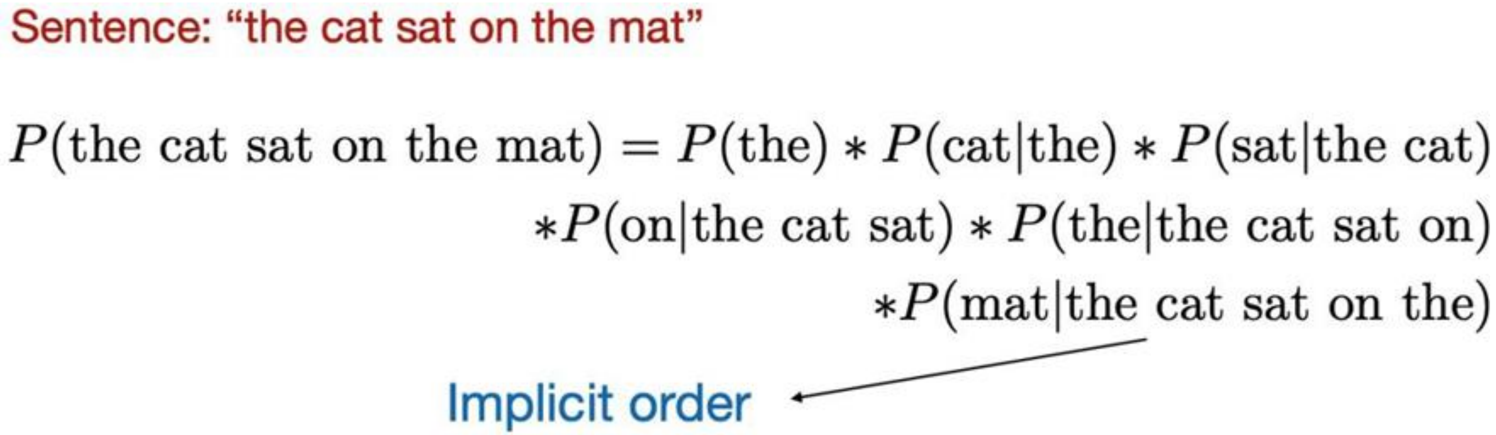
Language model using neural networks

GPT-3/ChatGPT/GPT4 have 175B+ parameters
Humans have 100B+ neurons



Language models: Narrow Sense

A probabilistic model that assigns a probability to every finite sequence (grammatical or not)



Sentence: "the cat sat on the mat"

$$\begin{aligned} P(\text{the cat sat on the mat}) = {} & P(\text{the}) * P(\text{cat}|\text{the}) * P(\text{sat}|\text{the cat}) \\ & *P(\text{on}|\text{the cat sat}) * P(\text{the}|\text{the cat sat on}) \\ & *P(\text{mat}|\text{the cat sat on the}) \end{aligned}$$

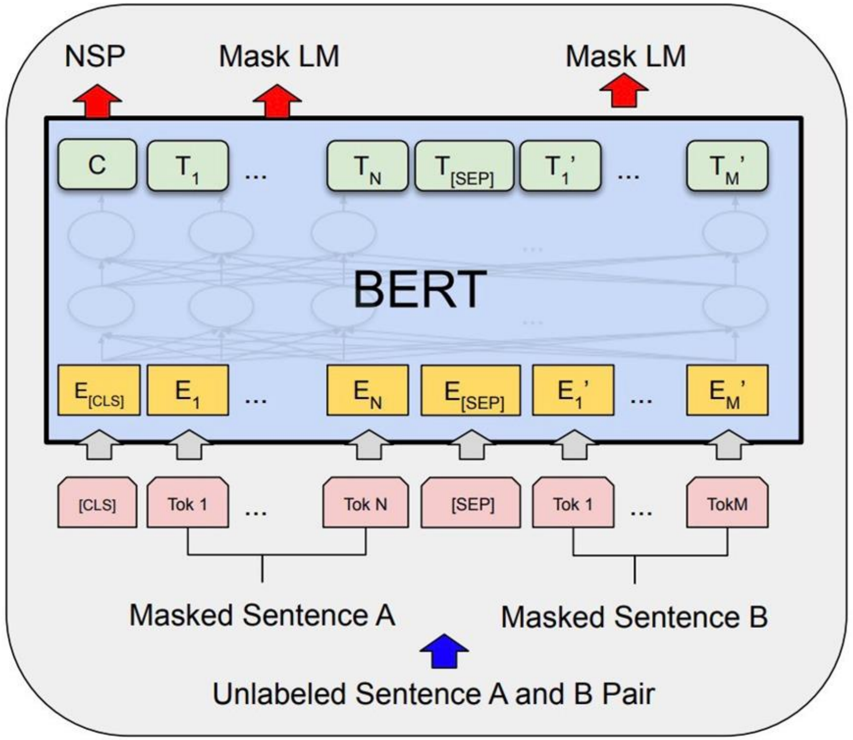
Implicit order

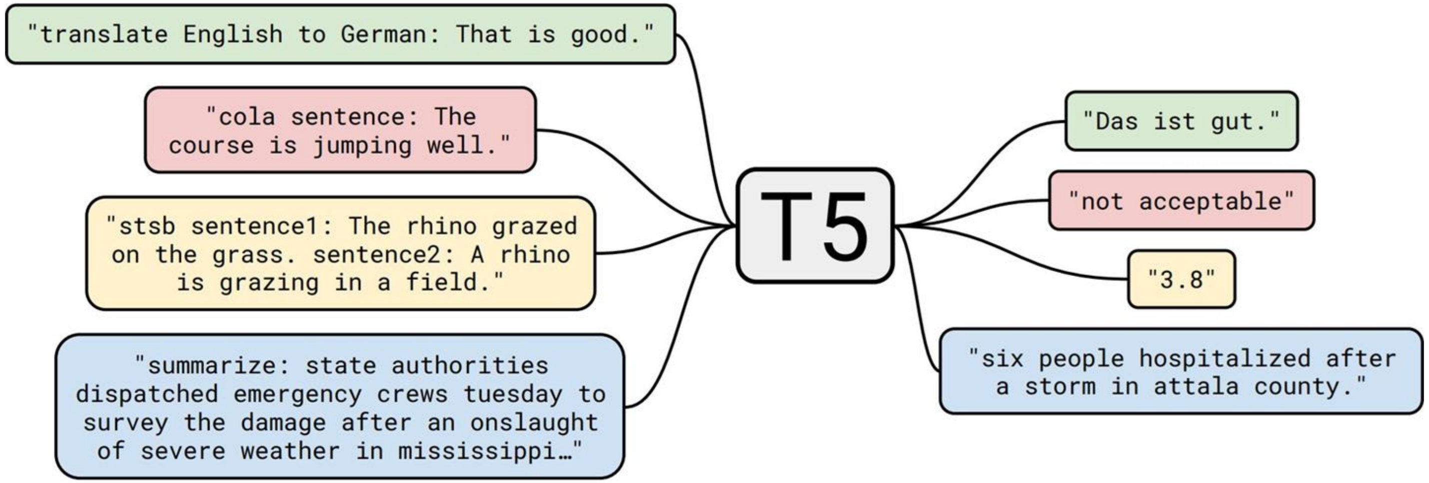
GPT-3 still acts in this way but the model is implemented as a very large neural network of 175-billion parameters!

Language models:Broad Sense

- Decoder-only models (GPT-x models)
- Encoder-only models (BERT, RoBERTa, ELECTRA)
- Encoder-decoder models (T5, BART)

The latter two usually involve a different **pre-training** objective.





"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

"Das ist gut."

"not acceptable"

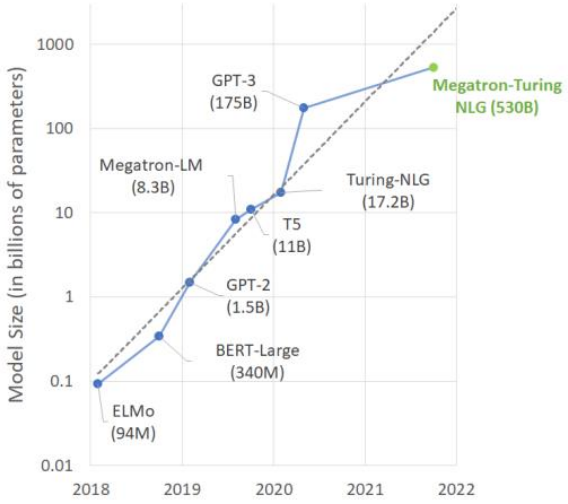
"3.8"

"six people hospitalized after a storm in attala county."

PLM vs. LLM

- Pre-trained language model
- **Large** ~~pre-trained~~ **L**anguage **M**odel (LLM)

We do not explicitly mention pre-training because pre-training and training use the same language models objective (e.g., autoregressive generation)



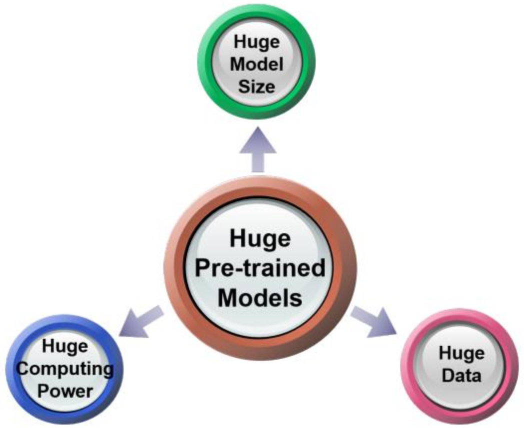
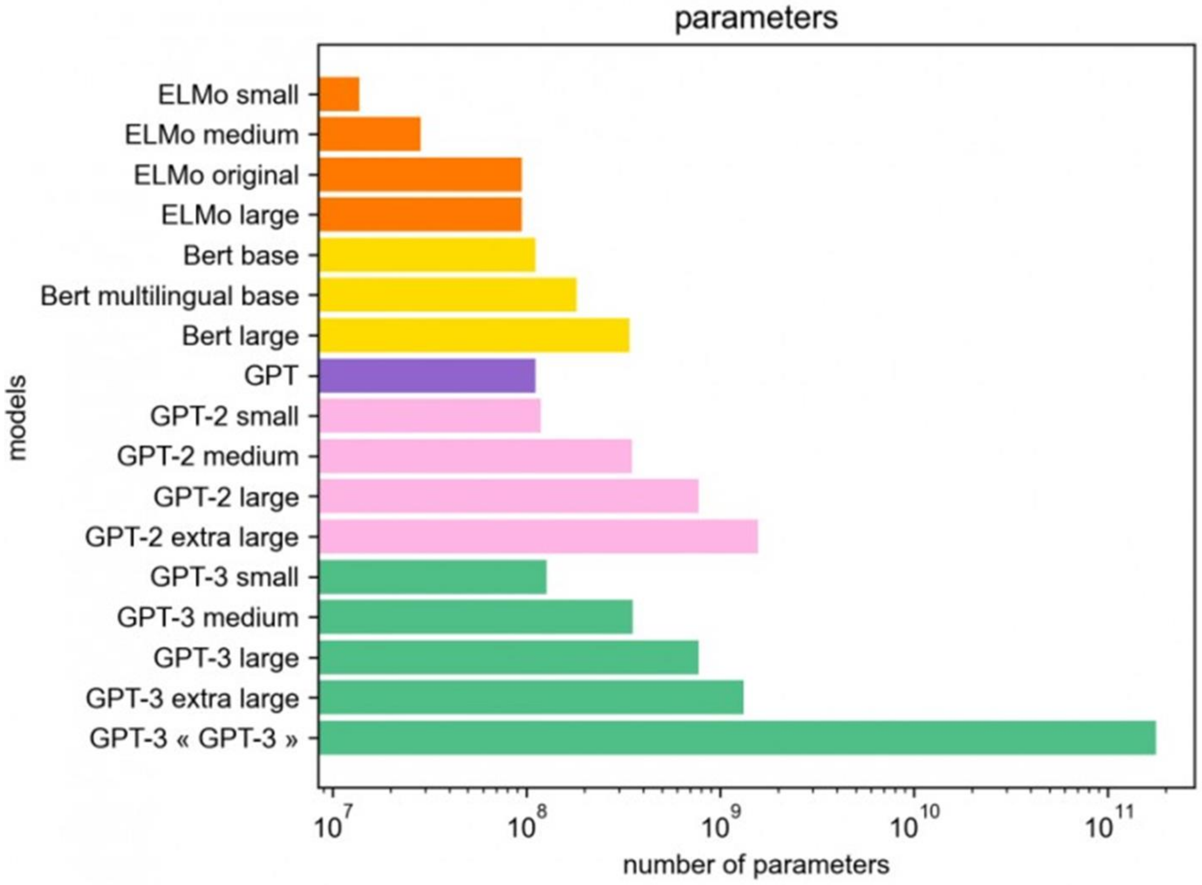
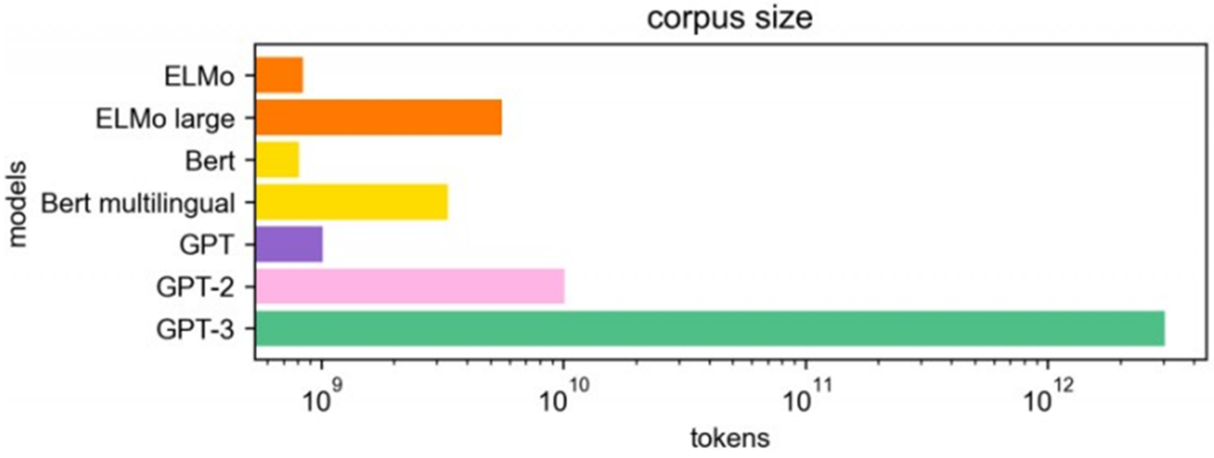


Image source: https://liuquncn.github.io/talks/20220228-Thai-AI-engineer-group/Huge-Pre-trained-Language-Models.public.pdf

How Large are “Large” LMs?



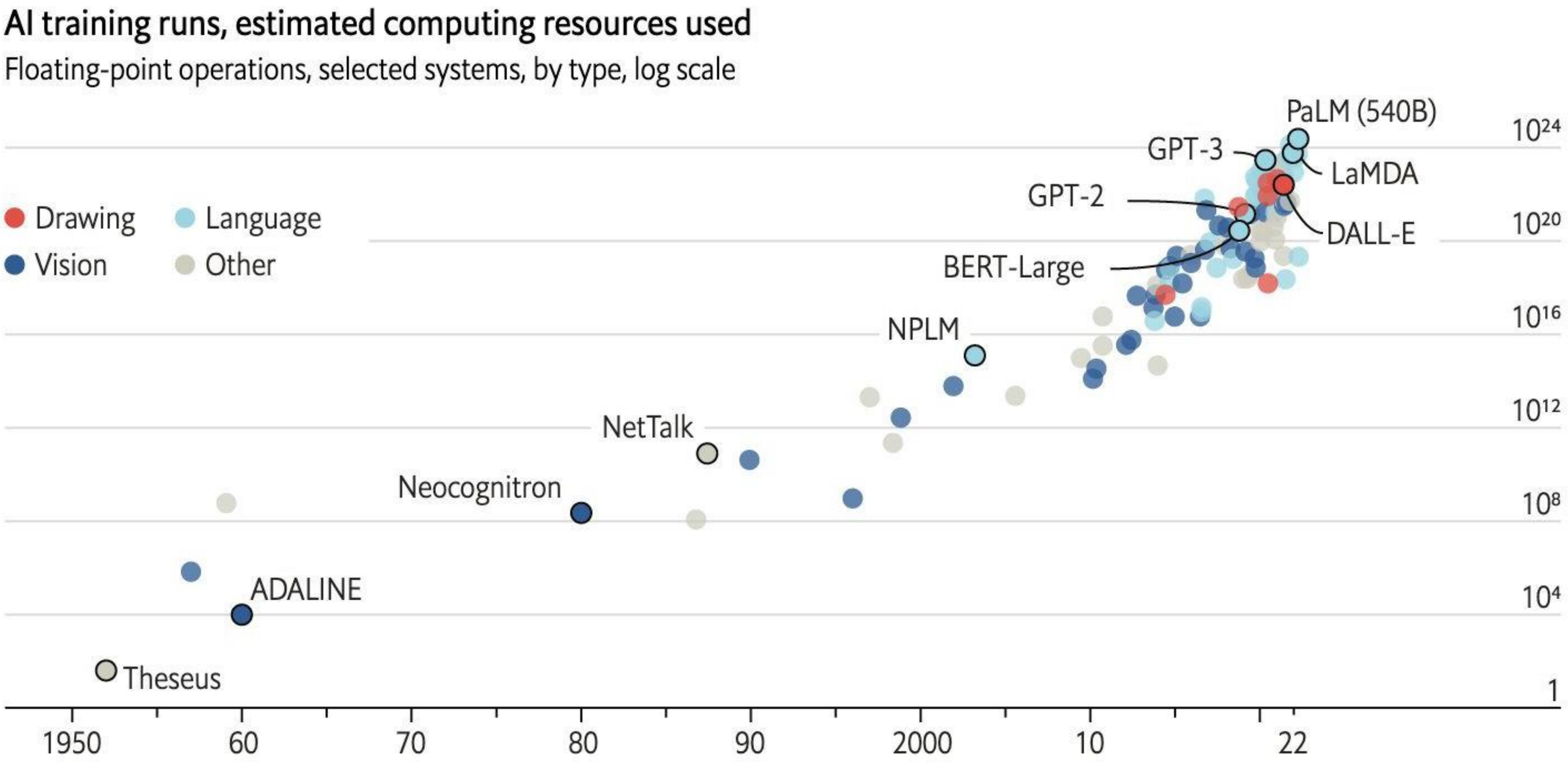


More recent models: PaLM (540B), OPT (175B), BLOOM (176B)…

https://huggingface.co/blog/large-language-models

Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

Large Language Models

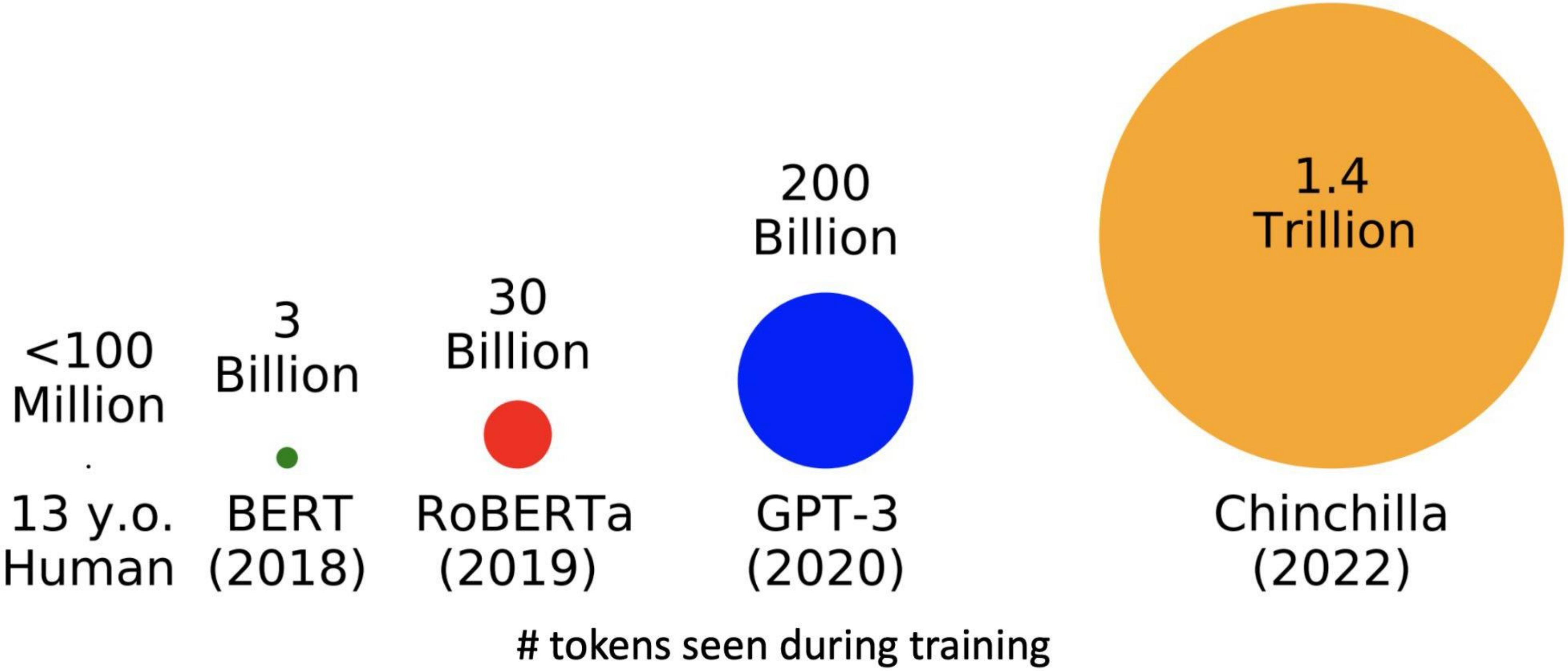


AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale

https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf 1 yotta = 10^{24} **FLOPs: floating point operations**

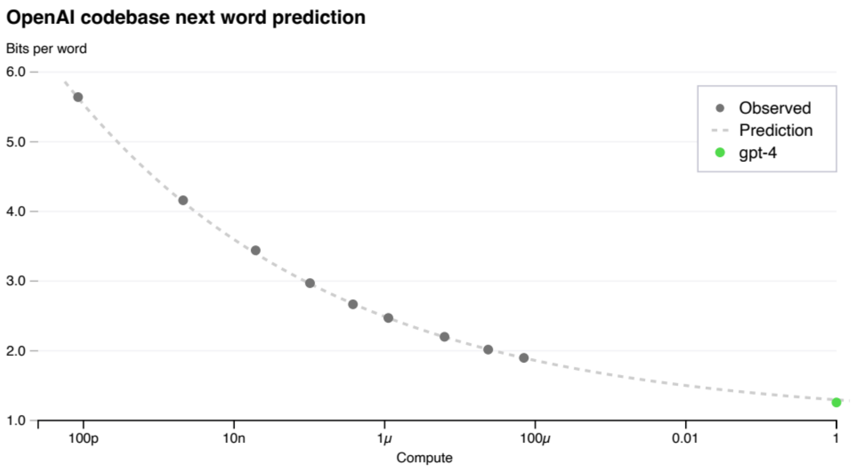
Large Language Models - Hundreds of Billions of Tokens



https://babylm.github.io/

Some basics for large language models

- Scalable network **architecture** (Transformer vs. CNN/RNN)
- Scalable **objective** (**conditional/**auto-regressive LM vs. Masked LM)



- Scalable **data** (plain texts are everywhere vs. supervised data)
    - https://github.com/esbatmop/MNBVC

OpenAI, GPT-4 Technical Report, https://cdn.openai.com/papers/gpt-4.pdf

How Large are “Large” LMs?

- Today, we mostly talk about two camps of models:
  - Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 770M, 3B)
  - “Very” large LMs: models of 100+ billion parameters

- Larger model sizes larger compute, more expensive during inference
- Different sizes of LMs have different ways to adapt and use them
  - Fine-tuning, zero-shot/few-shot prompting, in-context learning…
- Emergent properties arise from model scale
- Trade-off between model size and corpus size

Any Question?