



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

CSC5051/MDS5110/CSC6052: Natural Language Processing

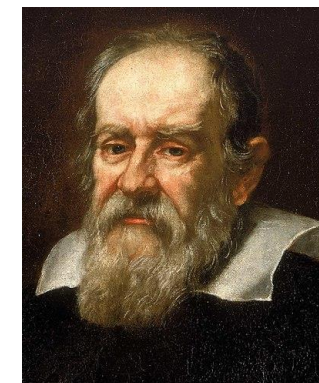
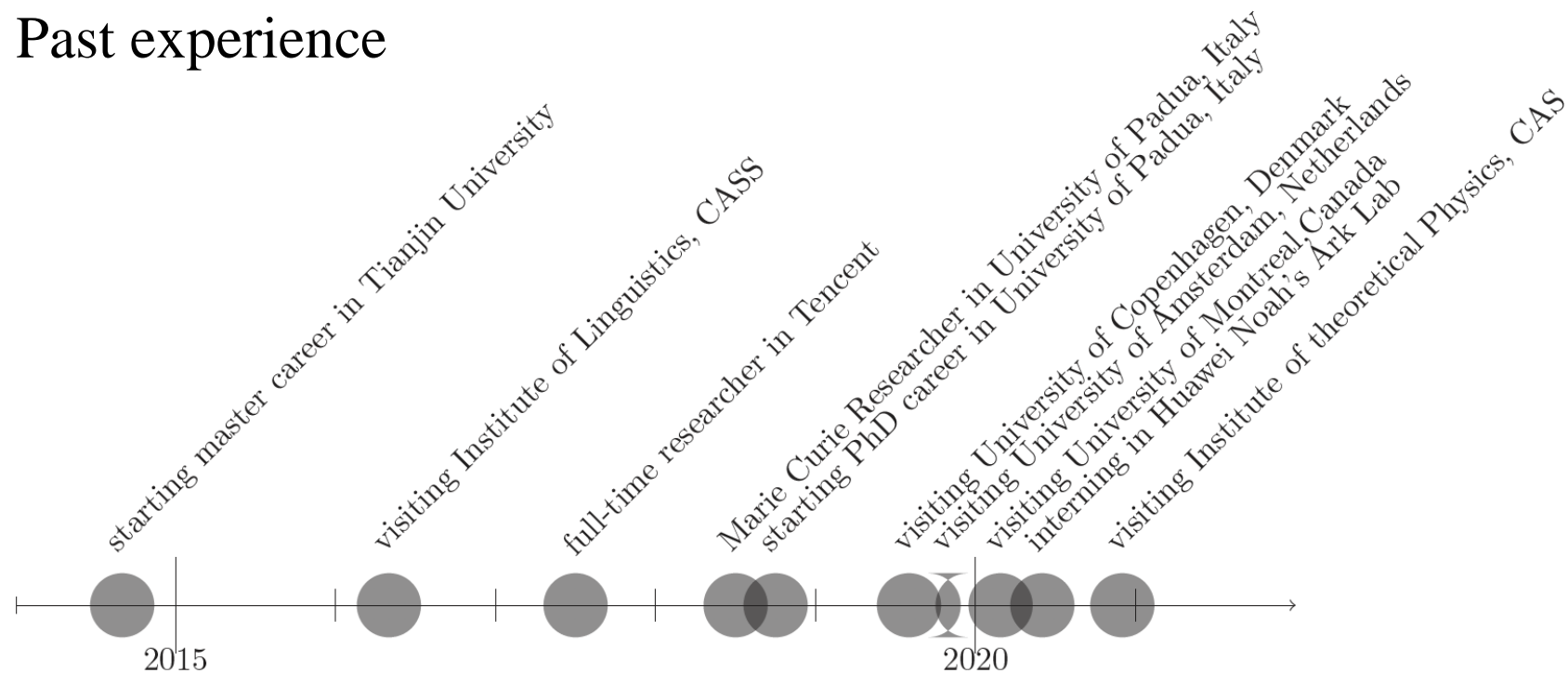
Spring 2026
Benyou Wang
School of Data Science

Contents

- **About me**
- **Philosophy of this course**
- **NLP and large language models (LLM)**
- **Introduction to our research**

About me 王本友 Benyou Wang.

Past experience



Galileo Galilei

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

Alumni of University of Padua



Academia or industrial career

- For money, **less** is **more**?
 - Work for **yourself** or **others**
 - **Free** to work in any topics
 - Benefit the **society** or just the **employer**?
 - easier to connect with diverse people.
 - **Long** vacation is much easier (but you may never take it)
-
- Weekend is **weekend**
 - **Less pressure** to be sustainable

Fitting yourself is the **best**



About me – 王本友 Benyou Wang.

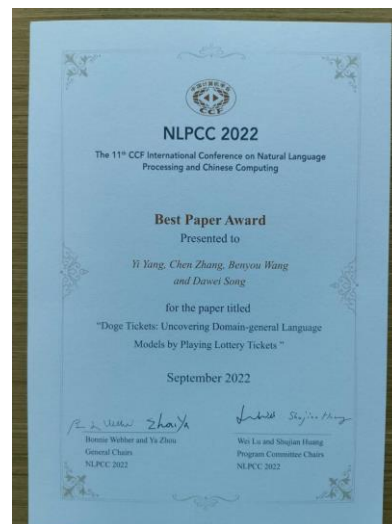
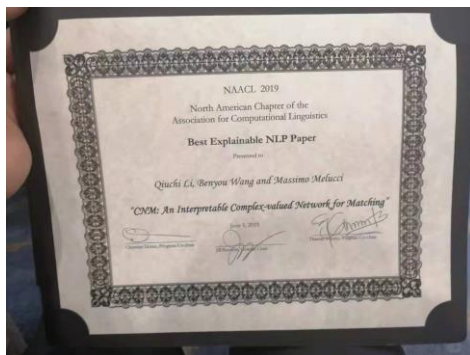
2022-~, The Chinese University of Hong Kong, Shenzhen

Assistant professor in **School of Data Science**, also affiliated in **School of Medicine in CUHKSZ**, as well as **the Shenzhen Loop Area Institute** (深圳河套學院) .

Research:

- **Planning** of Agent (complex reasoning/math modeling/tool calling/automatic theorem proof)
 - AI4math/modeling
- **Perception** of Agent (audio, vision, time-series, etc.)
 - Multimodal LLMs
- **Human-agent interaction** (HCI, agent simulation, brain-computer interface etc.)
- **Applications** of Agent (**Medical**/financial/legal/education)

Awards and honour



- **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/>
- **ICLR 2025** Financial AI Best Paper
- **Outstanding Paper Paper in NeurIPS 2025** ResponsibleFM Workshop
- EU Marie Curry researcher fellowship

- Huawei Spark award (华为火花奖)
- 腾讯犀牛鸟项目
- 滴滴盖亚学者项目
- 华为AI百校计划

We rank 14th 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_downloads ▲
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>

Be Open for AI



Hugging Face

Find **dataset/models/papers!**



Find **code!**

HuggingFace: <https://huggingface.co/FreedomIntelligence> (**1,000,000+** models downloading)

GitHub: <https://github.com/FreedomIntelligence> (**100,000** + stars)

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** (coauthorship).
- ❑ The possibility of a part-time **contract** with a salary (also with office space).
- ❑ **Internship** recommendation (Microsoft, Qwen, BAT, Huawei, Bytedance)

What We Expect From You:

- ❑ A commitment of at least 30 hours per week for full-time engagement.
- ❑ satisfied programming skills.

If you find this opportunity intriguing, please reach out to me
(wangbenyou@cuhk.edu.cn和 lixiang2@cuhk.edu.cn) for further details.

Contents

- **About me**
- **Philosophy of this course**
- **NLP and large language models (LLM)**
- **Introduction to our research**

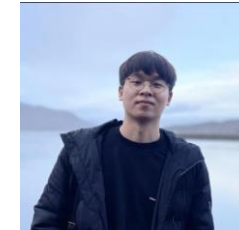
Logistics

- ❖ Instructor: Benyou Wang
- ❖ Teaching assistant: Shunian Chen (Leading TA)
Shunian Chen, Minghao Wu, Zhenyang Tang, Ke Ji



- ❖ Location: **Administration Bldg E101**

- ❖ Time: Friday 2:00PM - 4:50PM



- ❖ Office hours:

- Benyou Wang: Friday 6:30-7:30 PM. Daoyuan Building 504A. (Email: wangbenyou@cuhk.edu.cn)
- Shunian Chen: Friday 2:00-3:00 PM at Zhixin Building 431. (Email: shunianchen@link.cuhk.edu.cn)
- Juhao Liang: Wednesday 2:00-3:00 PM. Zhixin Building 431. (Email: 223010147@link.cuhk.edu.cn)
- Minghao Wu: Wednesday 6:30-7:30 PM. . Zhixin Building 431. (Email: 225040492@link.cuhk.edu.cn)

Logistics

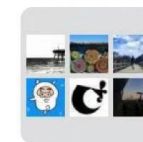
- ❖ [Official Website Link \(https://nlp-course-cuhksz.github.io/\)](https://nlp-course-cuhksz.github.io/)

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. ■

Date	Topics	Recommended Reading	Pre-Lecture Questions	Lecture Note	Coding	Events Deadlines
Sept. 1-4 Warmup	Tutorial 0: GitHub, LaTeX, Colab, and ChatGPT API	OpenAI's blog LaTeX and Overleaf Colab GitHub				
Sept. 5th	Lecture 1: Introduction to NLP	Hugging Face NLP Course Course to get into NLP with roadmaps and Colab notebooks. LLM-Course On the Opportunities and Risks of Foundation Models Sparks of Artificial General Intelligence: Early experiments with GPT-4	What is NLP?		[Phoenix]	

- ❖ Official Wechat Group



群聊: CUHKSZ NLP 2026
Spring



该二维码7天内(1月16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 are suggested 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.

Survey: How many of you have you ever used **LaTeX, GitHub, Python (PyTorch)?**

Course Structure (tentative)

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

Course Structure (tentative)

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

Too Old!

Course Structure (tentative)

~~•Introduction to NLP~~

~~•Linguistics and Word Embeddings~~

Too Old!

~~•Language Models~~

~~•Deep Learning in NLP~~

•Large Language Models (LLMs)

•Training Large Language Models

Practical

•Prompt Engineering

•LLM Agents

•Final Project Introduction and Research Sharing

•Multimodal Learning

•LLM Reasoning and Guest Lecture

•LLM Applications and Guest Lecture

Course Structure (tentative)

~~•Introduction to NLP~~

~~•Linguistics and Word Embeddings~~

Too Old!

~~•Language Models~~

~~•Deep Learning in NLP~~

•Large Language Models (LLMs)

•Training Large Language Models

Practical

•Prompt Engineering

•LLM Agents

•Final Project Introduction and Research Sharing

Final projects are all you need!

•Multimodal Learning

•LLM Reasoning and Guest Lecture

•LLM Applications and Guest Lecture

Course Structure (tentative)

~~•Introduction to NLP~~

~~•Linguistics and Word Embeddings~~

~~•Language Models~~

~~•Deep Learning in NLP~~

•Large Language Models (LLMs)

•Training Large Language Models

•Prompt Engineering

•LLM Agents

•Final Project Introduction and Research Sharing

•Multimodal Learning

•LLM Reasoning and Guest Lecture

•LLM Applications and Guest Lecture

Too Old!

Practical

Final projects are all you need!

Cutting-edge research

Components and Grading (Post-Graduate)

❖ Assignments (40%)

- Assignment 1 (10%): Training word vector.
- Assignment 2 (15%): Training NLP model with SFT and RLHF
- Assignment 3 (15%): Using API for testing prompt engineering and LLM agents.

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

❖ Final project (55%)

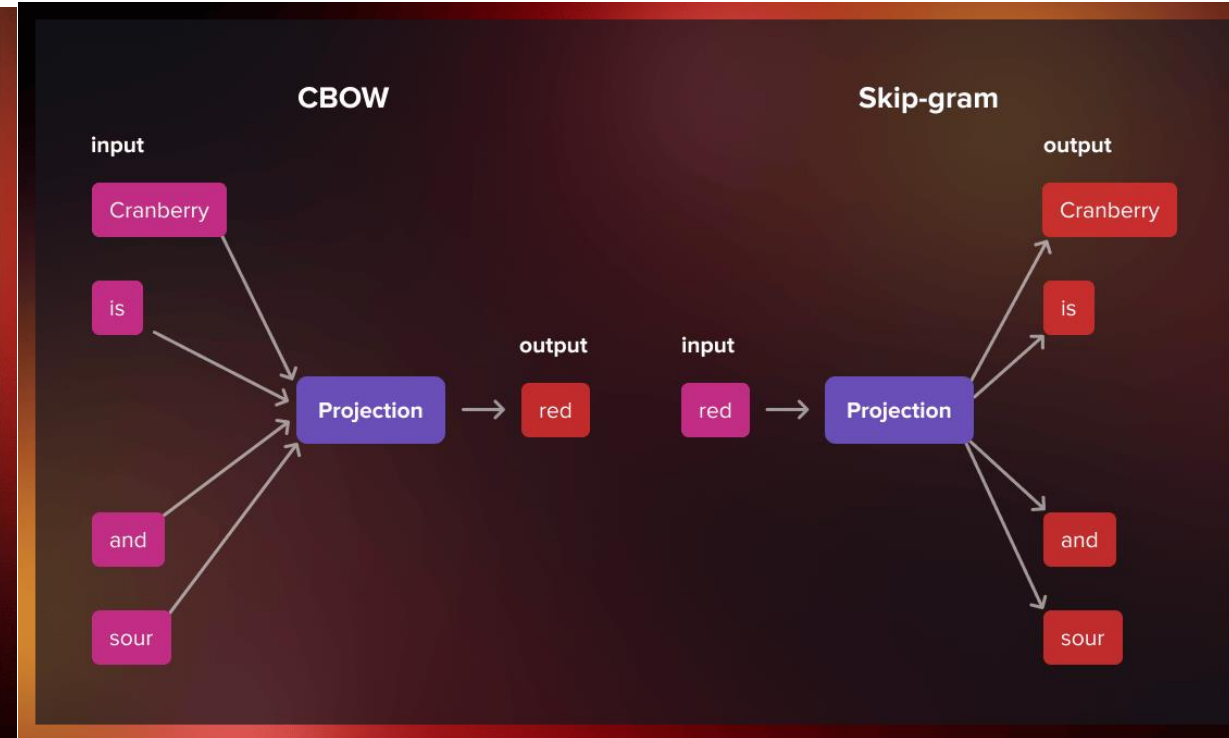
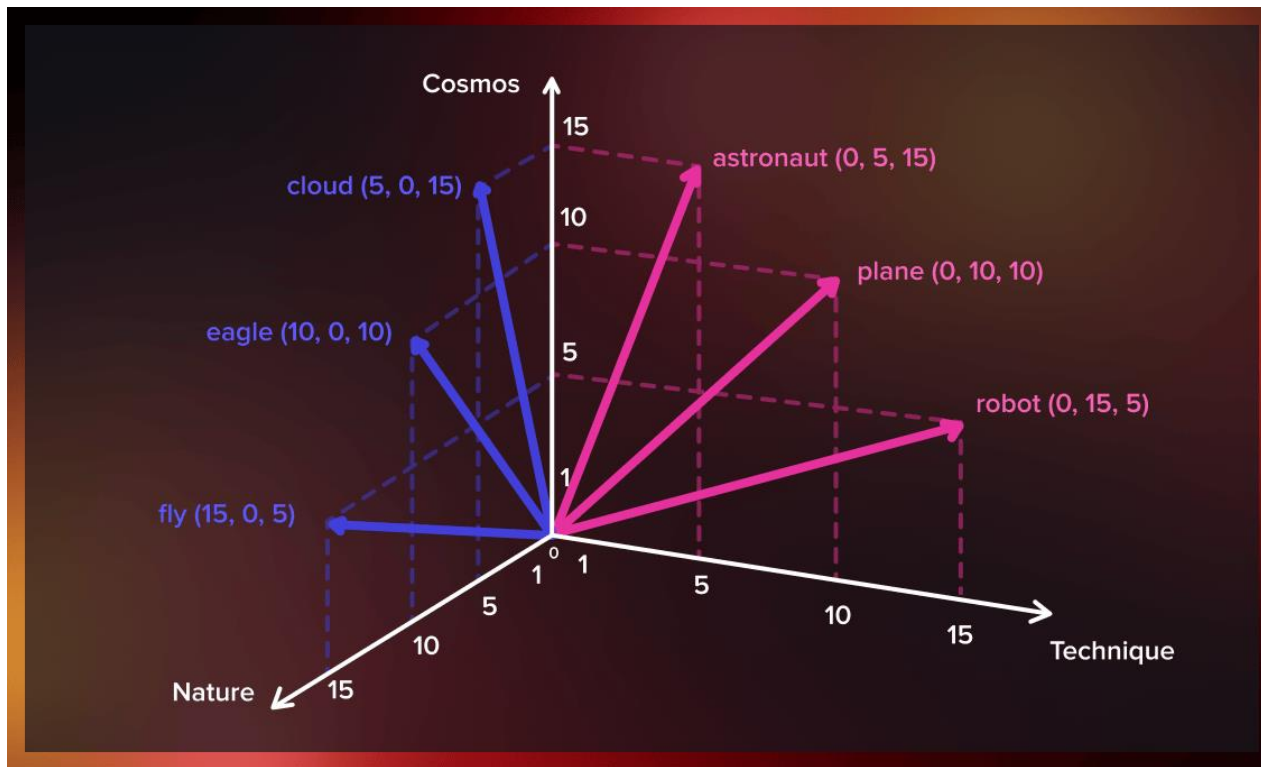
You need to write a project report (max 6 pages for the main body) for the final project [*Be longer is not necessary, not recommended*].

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%)

Quiz, guest lecture and CTE

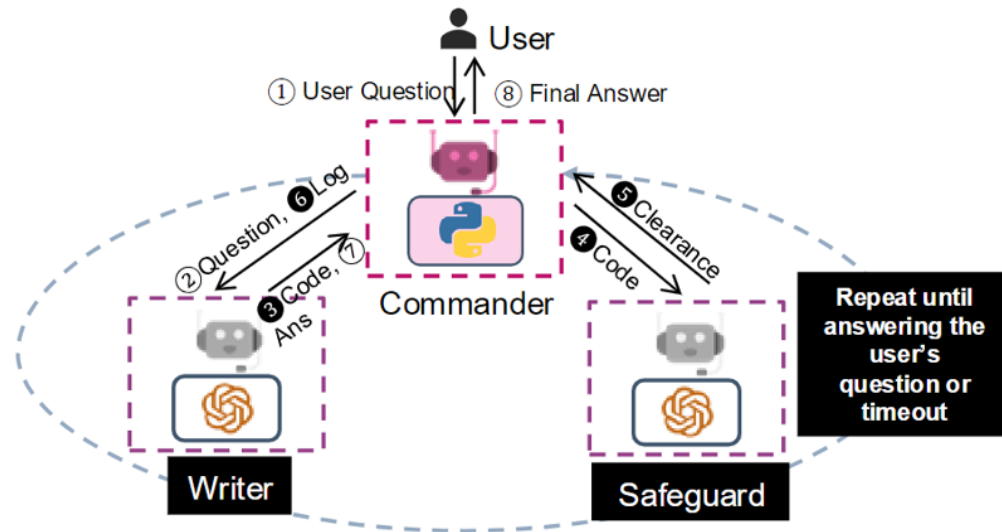
Assignments 1: Training word vector.



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.

API: 硅基流动: <https://siliconflow.cn/> or <https://openrouter.ai/>

Agent: Camel.AI, Dify, LangChain, Coze

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)
```

Finetuning: <https://github.com/FreedomIntelligence/LLMFactory>

RL: <https://github.com/volcengine/verl>

VSTASA: Verifying Subtask-level Training Supervisions in Autonomous Skill Acquisition

Group Number: 73

Authors: Guo Shuqi, Wang Chujiao

Student IDs: 120090167, 224045007

Email: 120090167@link.cuhk.edu.cn, 224045007@link.cuhk.edu.cn

Introduction

Recent advances in Embodied AI have positioned autonomous skill acquisition as a critical research frontier, driven by the need for systems that can facilitate continuous learning, adapt to dynamic environments, and make scalable decisions. While Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities across various domains, they still struggle with long-horizon, multi-step reasoning tasks essential for embodied skill learning.

The VSTASA framework addresses a fundamental question: How can we optimize autonomous skill acquisition by integrating verification models into embodied tasks? Unlike mathematical problems with definitive answers, embodied tasks present unique challenges: they are less structured and their success evaluation demands external observation and assessment.

Our approach bridges these gaps through two key components: a dynamic exemplar task pool that maintains structured representations of successfully executed tasks, and a scalable mechanism for training subtask-level verifiers using Monte Carlo Tree Search (MCTS) to simulate subtask sequences without requiring costly human annotations.

The VSTASA framework produces refined subtask training supervisions that significantly enhance autonomous skill acquisition. Our experiments demonstrate that providing relevant examples substantially improves the quality of LLM-generated supervisions, while our Process Reward Model verification system increases success rates by filtering and selecting quality supervisions. This task-specialized approach outperforms general-purpose verification systems, offering an effective solution for enhancing embodied AI learning capabilities.

Motivation

I. The Challenge: Enabling Autonomous Skill Learning in Embodied AI

Goal: Robots learn by interacting with the physical world (Embodied AI).

Why Autonomous Learning?

- General-purpose robots
- Adaptability in dynamic environments
- Scalable decision-making

Bottleneck: Effective supervision for complex, long-horizon tasks remains difficult.

II. Inspiration: Verification Models (e.g., PRMs) in Math Reasoning

- Assess LLM step-by-step correctness
- Provide feedback → Improve performance

III. Research Question: Can verification be adapted to optimize embodied skill acquisition?

- *Input: "Solve the equation: $2x + 8 = 34$."
- *Output: x
- *Prompt, abstract & from hints: $2(x + 4) = 34 - 8 \Rightarrow x + 4 = 13 \Rightarrow x = 9$.
- *Task, divide both sides by 2: $(2x + 8) / 2 = (34 - 8) / 2 \Rightarrow x + 4 = 13$.
- *Check the solution: $2(9) + 8 = 18 + 8 = 26 \neq 34$.
- *Model: $x = 9, 13, 14$.

Methodology

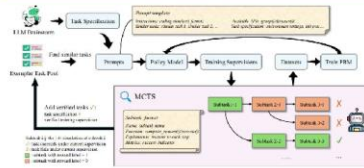


Figure 2: The VSTASA framework.

I. Exemplar Task Pool - Providing Structured Guidance

- Scores Successful past tasks (decompositions & supervisions).
- New Task → Encode (Sentence - BERT) → Retrieve Top-K similar via cosine similarity.
- Inject similar examples into LLM prompt.
- Dynamic Pool: Add newly successful tasks.

II. Automated Reward Labeling via MCTS

- Use MCTS to simulate/explore possible future subtask sequences to task completion.
- Execute these simulated sequences (in sim).
- Assign reward ($1=Success, 0=Fail$) based on final task outcome.

III. The Process Reward Model (PRM)

- Training Data: MCTS - generated dataset: <Subtask Sequence, Reward Label (0/1)>.
- Input: Task spec + Sequence of subtasks & their supervisions.
- Output: Predicted probability (0 - 1) of overall task success for that sequence.

Results

Method	ATW (%)	ASW (%)	Avg. Num. of Hints
BaseModel	0.74	0.60	9.0
BaseModel w/ ETP	0.75	0.55	17.8
CPT w/	0.76	0.63	8.2
CPT w/ w/ ETP	0.81	0.64	15.2

Table 3. Performance comparison for BaseModel and CPTs with and without the guidance of the Exemplar Task Pool.

Method (w/ or w/o ETP)	ATW (%)	ASW (%)	Avg. Num. of Hints	Model (w/ or w/o ETP)	ATW (%)	ASW (%)	Avg. Num. of Hints
BaseModel	0.74	0.60	9.0	BaseModel	0.74	0.60	9.0
BaseModel w/ ETP	0.75	0.55	17.8	BaseModel w/ ETP	0.75	0.55	17.8
CPT w/	0.76	0.63	8.2	CPT w/	0.76	0.63	8.2
CPT w/ w/ ETP	0.81	0.64	15.2	CPT w/ w/ ETP	0.81	0.64	15.2

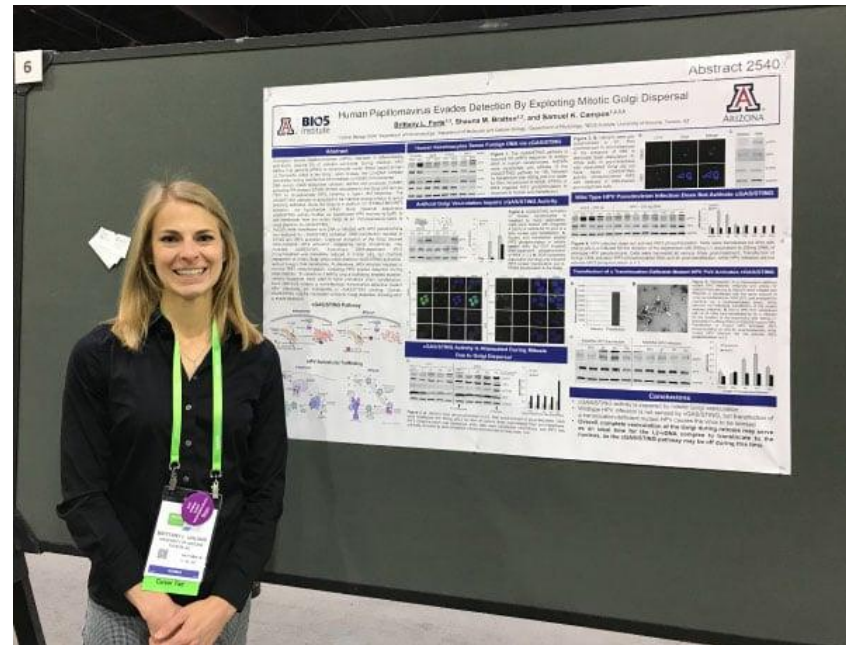
Table 4. Performance of LLM as an oracle in the verification model.

Method (w/ or w/o ETP)	ATW (%)	ASW (%)	Avg. Num. of Hints	Model (w/ or w/o ETP)	ATW (%)	ASW (%)	Avg. Num. of Hints
BaseModel	0.74	0.60	9.0	BaseModel	0.74	0.60	9.0
BaseModel w/ ETP	0.75	0.55	17.8	BaseModel w/ ETP	0.75	0.55	17.8
CPT w/	0.76	0.63	8.2	CPT w/	0.76	0.63	8.2
CPT w/ w/ ETP	0.81	0.64	15.2	CPT w/ w/ ETP	0.81	0.64	15.2

Our research shows that relevant examples significantly enhance the quality and efficiency of LLM-generated subtask supervisions. PRM verification substantially increases success rates by filtering and selecting quality supervisions. This task-specialized approach outperforms general-purpose verification systems, offering a more effective solution for autonomous skill acquisition.



Dennis Schroeder/NREL via flickr.com, Public domain (US government agency)



Example of Poster

Report for the Final project

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

It is like a paper

Final project

- Students complete a research project in teams of 1-3
- [**Optional**] Draft proposal deadline: **March. 27th** 1:59pm (simply explain your idea in one page)
- Poster presentation (poster sessions): **May. 8th**
- (TAs and the Instructor may provide suggestions for the revision during poster session!)
- Final poster deadline: **May. 7th** 17:00pm
- Final paper deadline: **May. 16th**

Prepare your project earlier to avoid the rush during exam weeks

Date	Topics	Recommended Reading	Pre-Lecture Questions	Lecture Note	Coding	Events Deadlines
Jan. 5-8 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 On the Opportunities and Risks of Foundation Models Sparks of Artificial General Intelligence: Early experiments with GPT-4	What is NLP?		[Phoenix]	
Jan. 16th	Lecture 2: Basics of Linguistics and Word Representation	Universal Stanford Dependencies: A cross-linguistic typology Insights between NLP and Linguistics End-to-end Neural Coreference Resolution Efficient Estimation of Word Representations in Vector Space (original word2vec paper) Evaluation methods for unsupervised word embeddings	What is structure of language (string of words)? How to model language and the inside words?		[Linguistics repo]	Assignment 1 out
Jan. 16th	Tutorial 1: Introduction to Overleaf, GitHub, Python, Pytorch, and Word Embeddings				[Colab]	
Jan. 23rd	Lecture 3: Language Modeling	A Neural Probabilistic Language Model BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding A Neural Probabilistic Language Model BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	What is a language model and how does it work in natural language processing?		[BERT]	
Jan. 30th	Lecture 4: Deep Learning in NLP	Attention Is All You Need HuggingFace's course on Transformers Scaling Laws for Neural Language Models The Transformer Family Version 2.0 On Position Embeddings in BERT	How to better compose words semantically as language?		[Transformer]	
Feb. 6th	Lecture 5: Large Language Models (LLMs)	Training language models to follow instructions with human feedback LLaMA: Open and Efficient Foundation Language Models Llama 2: Open Foundation and Fine-Tuned Chat Models OpenAI's blog Qwen2.5 Technical Report DeepSeek-V3 Technical Report	what are LARGE language models and why LARGE?		[Fine-tune Llama 2]	Assignment 1 due (11:59pm) Assignment 2 out

Mar. 13th	Lecture 7: Prompt Engineering		What are the current research topics in NLP?		[Prompt_engineer]	Final Project out
Mar. 20th	Lecture 8: LLMs as agents	Best practices for prompt engineering with OpenAI API prompt engineering	How to better prompt LLMs?			Assignment 2 due (11:59pm) Assignment 3 out
Mar. 20th	Tutorial 3: Prompt Engineering				[Colab]	
Mar. 27th	Lecture 9: Final Projects and Research Sharing	ToolBench AgentBench Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks LLM Powered Autonomous Agents Mobile Agent Cline Roo-Cline	How to make LLMs more useful?			
Mar. 27th	Tutorial 4: LLM Agents				[Colab]	Optional: Project Proposal due Team Formation due
Apr. 3rd	Lecture 10: NLP and Beyond NLP	Blog post: Generalized Visual Language Models	Can large models speak, see and perform actions ?		[NEXT-GPT]	
Apr. 10th	Lecture 11: LLM Reasoning	OpenAI's O1 DeepSeek-R1-Lite-Preview Qwen-32B-Preview HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs	How to improve LLMs' reasoning?			Assignment 3 due (11:59pm)
Apr. 17th	Lecture 12: Guest Lecture					
Apr. 24th	Lecture 13: LLM Applications and Future Cont.	Large Language Models Encode Clinical Knowledge Survey of Hallucination in Natural Language Generation Superalignment GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models	What are the future of LLMs?			
May. 8th	Final Project Presentation		How to solve real-world problems using LLMs			Final Project Poster due (17:00pm May 7th)

What can we do for the final project?

Workload of final project is **very high!**

Be careful if you do not have enough time!

Prepare as earlier as possible

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 Nov. 1st 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
 - Faculty 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. Fine-tune (or RL) a language model (e.g., Qwen) yourself for any problems of your interest.

“Do we really need to tune the base model?”

2. Evaluate existing LLMs (e.g., GPT-5.2) and understand their capabilities, limitations and risks;

3. An Agent System (math/financial/medical/legal domains) **Note: You might get computing resources to train 10B+ model if TAs/instructor like your proposal**

4. An APP that is for something.

5. A research paper involving NLP (please acknowledge this course if possible)

- ~~6. An survey that involving NLP~~

7. New!: A competition with a technical report.

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

Final project (suggested to be publicly released)

Report of final project might 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 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

Example: ArXiv+HF+Github

MedGen: Unlocking Medical Video Generation by Scaling Granularly-annotated Medical Videos

Rongsheng Wang^{1*}, Junying Chen^{1*}, Ke Ji¹,
Zhenyang Cai¹, Shunian Chen¹, Yunjin Yang¹, Benyou Wang^{1†}

¹The Chinese University of Hong Kong, Shenzhen
wangbenyou@cuhk.edu.cn

<https://github.com/FreedomIntelligence/MedGen>

Link your code here

Abstract

Recent advances in video generation have shown remarkable progress in open-domain settings, yet medical video generation remains largely underexplored. Medical videos are critical for applications such as clinical training, education, and simulation, requiring not only high visual fidelity but also strict medical accuracy. However, current models often produce unrealistic or erroneous content when applied to medical prompts, largely due to the lack of large-scale, high-quality datasets tailored to the medical domain. To address this gap, we introduce **MedVideoCap-55K**, the first large-scale, diverse, and caption-rich dataset for medical video generation. It comprises over 55,000 curated clips spanning real-world medical scenarios, providing a strong foundation for training generalist medical video generation models. Built upon this dataset, we develop **MedGen**, which achieves leading performance among open-source models and rivals commercial systems across multiple benchmarks in both visual quality and medical accuracy. We hope our dataset and model can serve as a valuable resource and help catalyze further research in medical video generation.

⚠ Disclaimer: This paper contains clinical content that may be disturbing to some readers.

1 Introduction

Recent advances in video generation have led to impressive breakthroughs, with models now capable of producing high-quality, cinematic visuals that align closely with user prompts (Blattmann et al. 2023). In particular, latent video diffusion models (LVDMs), such as Sora (OpenAI 2025) and Veo (Sharma et al. 2024), have achieved state-of-the-art performance by operating efficiently in latent space and delivering diverse, coherent video outputs from textual descriptions.

Despite this progress, **medical video generation** remains a largely underexplored yet crucial domain. Medical videos are indispensable in numerous real-world applications, including clinical training, surgical simulation, and patient education (Li et al. 2024b). Unlike everyday video content, medical videos demand precise rendering of anatomical structures, accurate surgical steps, and realistic physiologi-

demands on visual fidelity, semantic correctness, and temporal coherence.



Figure 1: Failure cases of Sora, Pika, and Hailuo on medical video generation. (a): Medical common sense errors. (b): Distortion errors.

However, current video generation models are trained almost exclusively on general-purpose datasets that focus on natural scenes and everyday activities (Blattmann et al. 2023). As a result, when applied to medical prompts, they often generate outputs with critical errors—such as anatomical inconsistencies, tool misuse, and implausible clinical scenarios. As shown in Figure 1, even leading models like Sora (OpenAI 2025), Pika (Team 2025a), and Hailuo (MiniMax 2025) fail to maintain basic medical realism, revealing a clear mismatch between training data and medical domain requirements. A key bottleneck lies in the lack of large-scale, high-quality datasets tailored for medical video generation (Sun et al. 2024). Existing medical datasets are limited in size, narrow in scope (e.g., only endoscopic or surgical videos), and mostly provide categorical labels instead of detailed descriptions—making them unsuitable for training or

arXiv:2507.05675v1 [cs.CV] 8 Jul 2025

Team work

Team size: Students may do final projects sololy, 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%)

TA/instructor review

Cross review between students (not contributed in assessment)

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

- Submission Deadlines
 - ACL ARR (five times per year)
 - March 16 , May 25
 - Feb. ACL, SIGIR, ICML
 - March, **COLM 2026** (<https://colmweb.org/>)
 - **May, NeurIPS 2026**
- TAs might help upon their interest. They might also refer you to other students in this group.
- You could also work with external members, please adhere to academic ethics

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

Ideas from where?

- Papers from **NLP** conferences
 - <https://aclanthology.org/>
 - <https://aclanthology.org/events/acl-2025/>
- Papers from **ML** conference
 - <https://openreview.net/group?id=ICLR.cc/2026/Conference#tab-active-submissions>
- Papers from **LLM** conference
 - <https://colmweb.org/AcceptedPapers.html>
- Papers from **HCI** conferences
 - <https://hci-deadlines.github.io/?sub=HCI,DES,CSCW,AI,HRI,XR,HAP,VIS,SP>
 - <https://dl.acm.org/doi/proceedings/10.1145/3706598>

Natural language processing (**NLP**) + Large language models (**LLMs**) + human-computer interactions (**HCI**)

Ideas from where?

- Previous **major**
 - **Finance, management, social science?**
- Personal Interest
- Existing resources from your friends
 - Data ?
- Existing use cases
 - ways to connect users

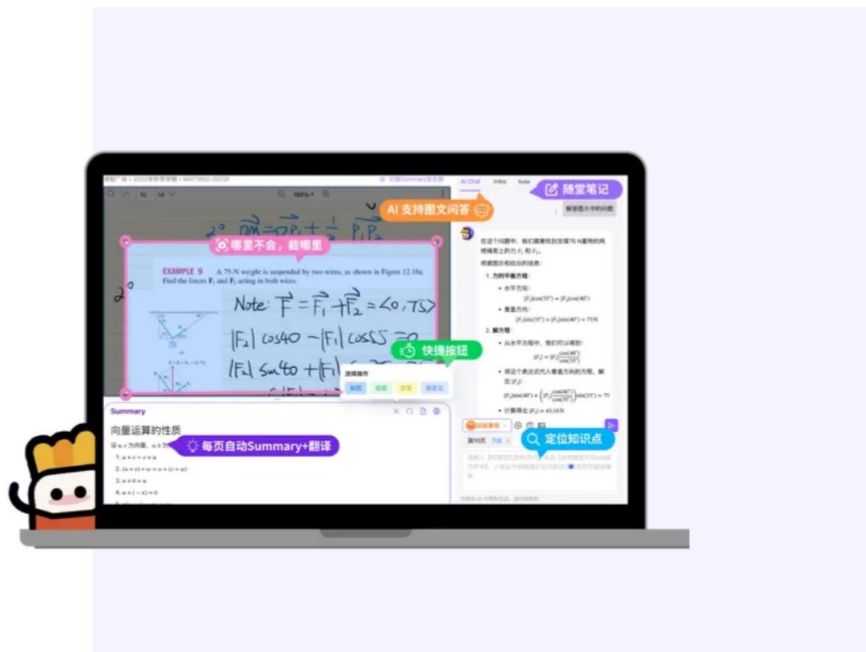
For startups: fired AI for classroom learning

薯塔AI

你的赛博「课件学习」搭子

集「课件共享社区、翻译、AI问答、跨文档知识检索、随堂笔记」为一体的AI课件学习助手，高效助力你的课件学习、预习、复习。

立即开始



薯塔AI是一款专为中外合作院校、国际学校及更广泛教育场景设计的智能助学平台。依托先进的AI技术（如大语言模型、智能问答引擎等），薯塔AI为学生提供高效、便捷的学习支持；通过课件共享社区、AI问答、跨文档知识检索等核心功能，薯塔AI为学生提供个性化的学习支持，通过智能化工具增强学习互动和参与感，全面提升学习效率。

本课题组支持的学生创业团队已获得了超过2000每月活跃用户，服务大部分港中深课程，正在拓展到其他中外合办高校和国外高校

薯塔

我的课程

CSC1001-2022F Introduction to Computer Science... [课件预览 >](#)

MAT1002-2022F Calculus II [课件预览 >](#)

MAT2041-2022F Linear Algebra and Applications [课件预览 >](#)

CSC1001-WU,Zhizheng Introduction to Computer Science... [课件预览 >](#)

CSC1001-WU,Zhizheng Introduction to Computer Science... [课件预览 >](#)

课程助手

管理课程

成为共建者




AI辅助PPT学习，轻松掌握课程要点

< 2024年秋季学期 >


搜索课程 (请注意区分大小写)

☆ 收藏



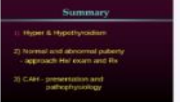
CHI1000-Yu,Peng
Chinese

☆ 收藏




MED3003-Alice SIU
Chemical Pathology

☆ 收藏




MED4012-Nelson,Tony
Paediatrics Medicine

★ 已收藏




CSC1001-WU,Zhizheng
Introduction to Computer Science: Programming Methodology

☆ 收藏




CSC1001-Han,Xiao
Guang
Introduction to Computer Science: Programming Methodology

☆ 收藏




CSC1003-LI,Wenye
Introduction to Computer Science and Java Programming

☆ 收藏



CSC3100-YU,Tiansu
Data Structures

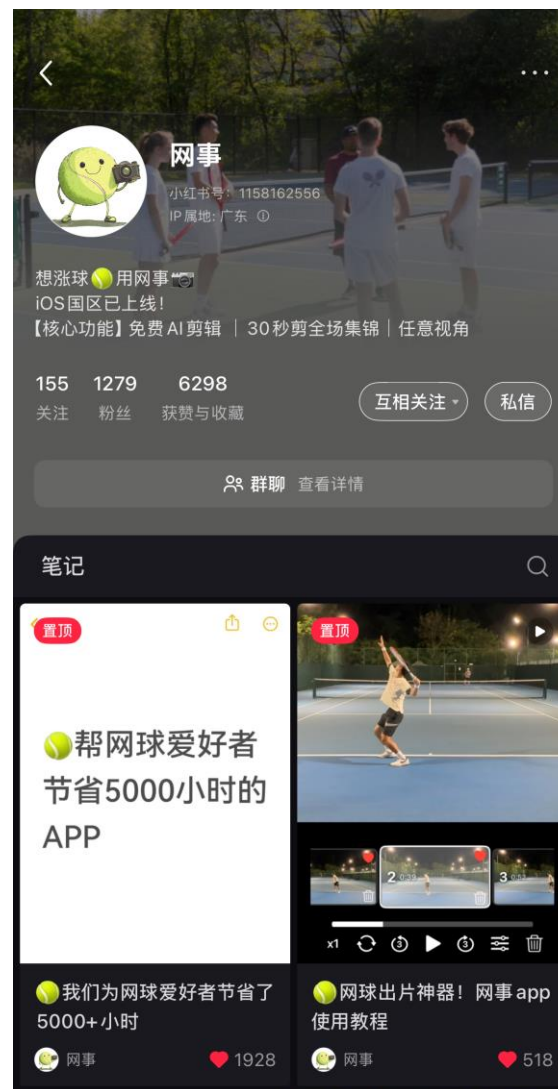
☆ 收藏



DDA3010-CHEN,Tianshi
Applied Machine Learning

截至2025年4月12日13点30分，薯塔AI已覆盖港中深**60%**的本科生，**港中深注册人数达3579人**，**总注册人数达4680人**。其中日活跃用户为1239人，平均每日AI问答数2898条。薯塔AI目前正在逐步推广至其他七所中外合办院校，进一步提升平台的用户覆盖率和影响力。随着平台影响力的扩大，我们将为更多院校的学生享受更高效、更智能的学习体验。

For startups



朱宇轩同学把网球和AI相结合，他开发的网事APP已服务一千多个用户，为网球爱好者节省超过5000个小时

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:

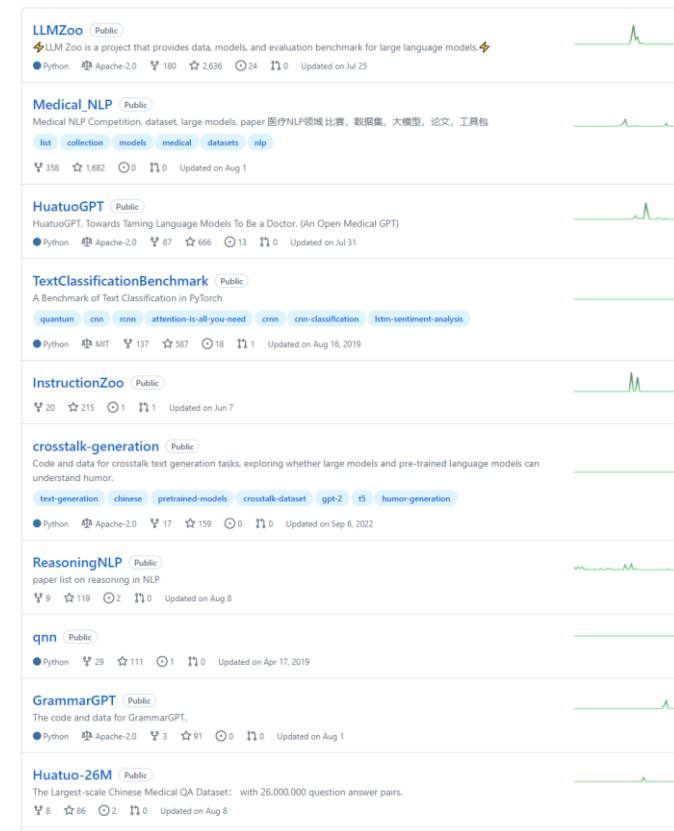
- Discussions in the end of each lecture
- In-class presentation
- Interrupting me whenever needed
- **Make friends with the instructor and TAs**

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 API
- LLMZoo
- LLMFactory



<https://github.com/orgs/FreedomIntelligence>

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 between yourself.

Seek feedback and iterate this course:

- 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 Language 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 mechanism of NLP models
- *SORA might not be introduced*

How to not fail in this course

- Finish all assignments, the poster and final report!
- Be honest even if you get negative results
- Be nice to your teammates, classmates, and TAs

How to success in this course

- Start final projects as earlier as possible.
- Actively talk to the instructor and TAs frequently
 - You could talk to us after the lecturing time, e.g. after 5:00 pm
- Bring your personal expertise/interest/user/data and combine it with AI
- Aim Higher
 - A **high-quality paper** that help you get a nice PhD offer or internship
 - Run a **startup**; find (future) co-funders

You might obtain from a great project

- Recommendation for internship from Qwen, Tencent, Huawei, DeepSeek, Kimi
 - TAs interned or are interning in MicroSoft, Qwen, Tencent, Kimi, MiniMax, Huawei etc.



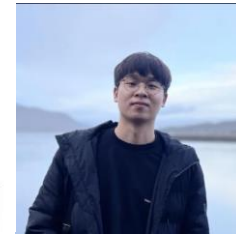
MicroSoft, Qwen, Tencent, Kimi



Tencent, MiniMax, Kimi



Huawei



Top VC

- Recommendation Letter for PhD

- Possibility to join our group as a PhD

- 蔡镇阳 (本组转博)
- **Zhenyang Cai**, Junying Chen, Rongsheng Wang, Weihong Wang, Yonglin Deng, Dingjie Song, Yize Chen, Zixu Zhang, Benyou Wang. Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging. ACL main conference 2025.
- 曾子懿 (本组转博)
- **Ziyi Zeng**, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Sigi Cai, Benyou Wang., Zhiguo Zhang, Haizhou Li. WaveMind: a Conversational EEG Foundation Model Aligned to Textual and Visual Modalities. Submitted to AAAI 2025
- 洪畅 (潜在转博)
- **Chang Hong**, Minghao Wu, Qingying Xiao, Yuchi Wang, Xiang Wan, Guangjun Yu, Benyou Wang, Yan Hu. Towards Assessing Medical Ethics from Knowledge to Practice. submitted to KDD. <https://arxiv.org/pdf/2508.05132>

A break!

Try Doubao and DeepSeek

And see the difference?



Contents

- **About me**
- **Philosophy of this course**
- **NLP and large language models (LLM)**
- **Introduction to our research**

Natural Language Processing

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core Technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- ...

NLP lies at the intersection of computational linguistics and machine learning.

What is NLP?

Wikipedia: Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) language.



Why **natural language** instead of **language**

What is the difference
between a

Programming
language

Natural
language

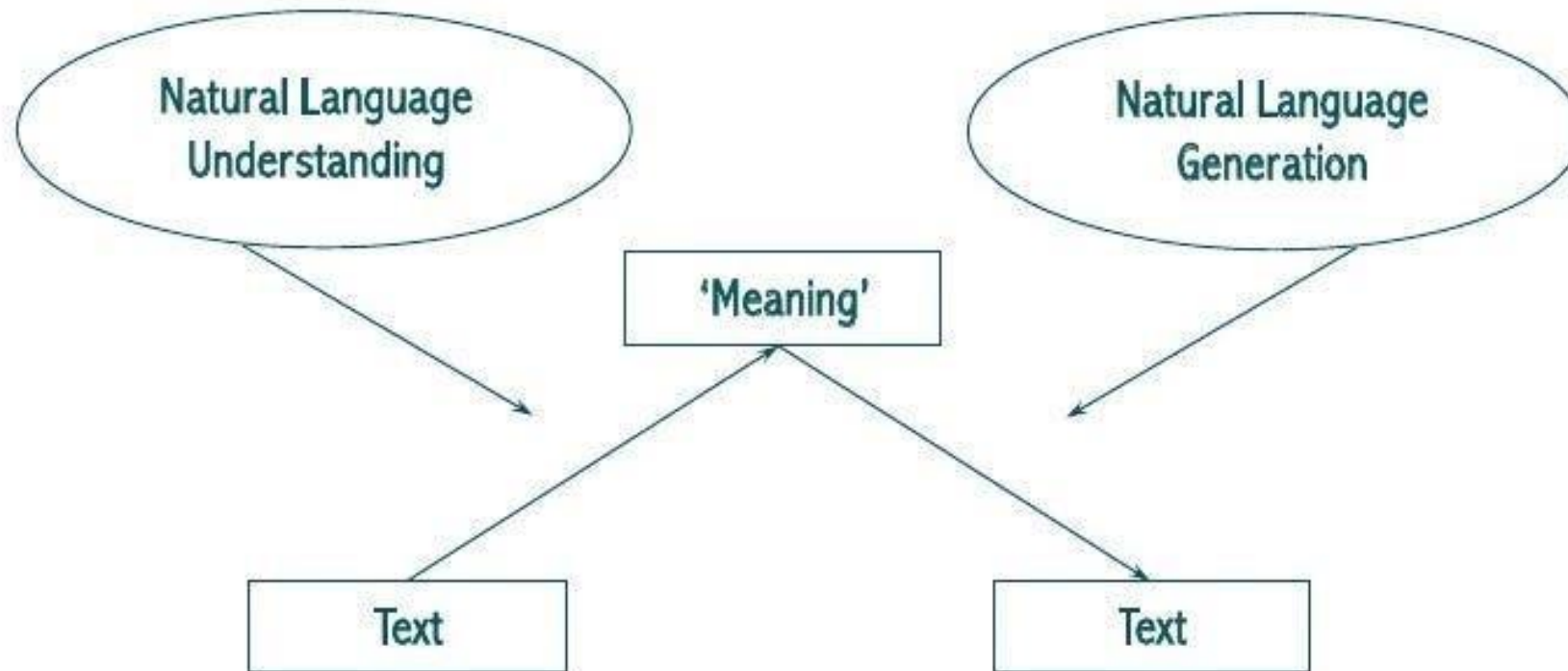


But in the large language models era, NLP also deals with programming language (see Prof. Pinjia He)

What is NLP?

NLP — Natural Language “Processing” **NLU** — Natural Language “Understanding” **NLG** — Natural Language “Generation”

NLP = NLU + NLG



Go **beyond** the keyword matching



- Identify the **structure** and meaning of words, sentences, texts and conversations
- Deep understanding of broad language
- NLP is all around us

Machine translation

Google

All Images Shopping Apps Videos More

About 20,800,000 results (0.54 seconds)

Spanish English

3 more translations

Haaretz הארץ
אולי הפעם הוא יצליח?



היום: נתניהו שוב ינסה להעביר את דו"ח טרכטנברג בממשלה
www.haaretz.co.il

אחרי הפאריסה בשבוע שעבר, בלשכת רה"מ מסרבים להתחייב שתהיה הצבעה בתום הדיון. בישראל ביתנו, ש"ס והעצמאות דורשים ריכוך או השמטה של חלק מההמלצות

Like · Comment · Translate · Share · Yesterday at 06:00

9 people like this.

View 1 share

 שוקי זקס עם שלם משלם את מחיר החיחות והקמבינות של ראש הממשלה. שלא מצא לנכון להתגייס להבראת מערכת הבריאות. סיכן חיים אישים ולאחרים גם לסבל. חובה לגרש את החשף הזה מהעיר. מארה"ב כבר זרקו אותו. מאירוסה גם. האיש הזה הורס בשקדנות את המדינה ומכלה את אורחיה. בוד לך בינמין נתניהו. נתניהו נרתע: לא מתחייב להצבעה על טרכטנברג
www.ynet.co.il

השותפות הקואליציוניות הצליחו להטיל מורא על ראש הממשלה, שכבר אינו מתחייב להביא להצבעה היום את מסקנות דו"ח טרכטנברג. ש"ס לפי שעה מתנגדת. ישראל ביתנו ומפלגת העצמאות יקימו בבוקר ישיבת שרים ויחליטו רק בה - המערכת הפוליטית, חדשות

Expand preview

Yesterday at 06:31 · Like · 2 people · Translate

 עמי וקנין או שהדו"ח יעבור או שביבי יעבור
Yesterday at 07:10 · Like · Translate

 Dalya Gumis שפסיקו להצביע ותחילו לבצע
Yesterday at 08:11 · Like · Translate

 Yuval Gilor נמאס כבר לך הביתה
Yesterday at 08:49 · Like · Translate

Haaretz הארץ
Maybe this time he succeeds? [?]



היום: נתניהו שוב ינסה להעביר את דו"ח טרכטנברג בממשלה
www.haaretz.co.il

אחרי הפאריסה בשבוע שעבר, בלשכת רה"מ מסרבים להתחייב שתהיה הצבעה בתום הדיון. בישראל ביתנו, ש"ס והעצמאות דורשים ריכוך או השמטה של חלק מההמלצות

Like · Comment · Original · Share · Yesterday at 06:00

9 people like this.

View 1 share

 שוקי זקס With full pay the price for hzhiloth vakombinot of the Prime Minister. Not find it appropriate to action lhabrat the health system. Endangered animals, people and caused the suffering that must drive with ... city. Foundation of Minneapolis had already thrown him. Europe also. This man destroys the country with the citizens in its path. Contempt you Binyamin Netanyahu. Netanyahu was not committed to voting the Trachtenberg managed to impose a "partnership hloalicioniot terror the Prime Minister, who is pledging to bring voting day the conclusions report Trachtenberg. hourly opposes. Israel Beiteinu-independence morning ministerial session will vihalifo only it-political system, news

Expand preview

Yesterday at 06:31 · Like · 2 people · Original

 עמי וקנין Or acknowledging the report moves or the chips will go
Yesterday at 07:10 · Like · Original

 Dalya Gumis Stop vote once to
Yesterday at 08:11 · Like · Original

 Yuval Gilor Tired already go home
Yesterday at 08:49 · Like · Original

Dialog Systems

Gift shop

Items such as caps, t-shirts, sweatshirts and other miscellanea such as buttons and mouse pads have been designed. In addition, merchandise for almost all of the projects is available.



Hi. I'm your automated online assistant. How may I help you?

Ask

CD or DVD

There is a series of CDs/DVDs with selected

Wikipedia content being produced by Wikipedians and [SOS Children](#).

Downloading

Downloading content from Wikipedia is

free of charge.

All text content is licensed under the [GNU](#)

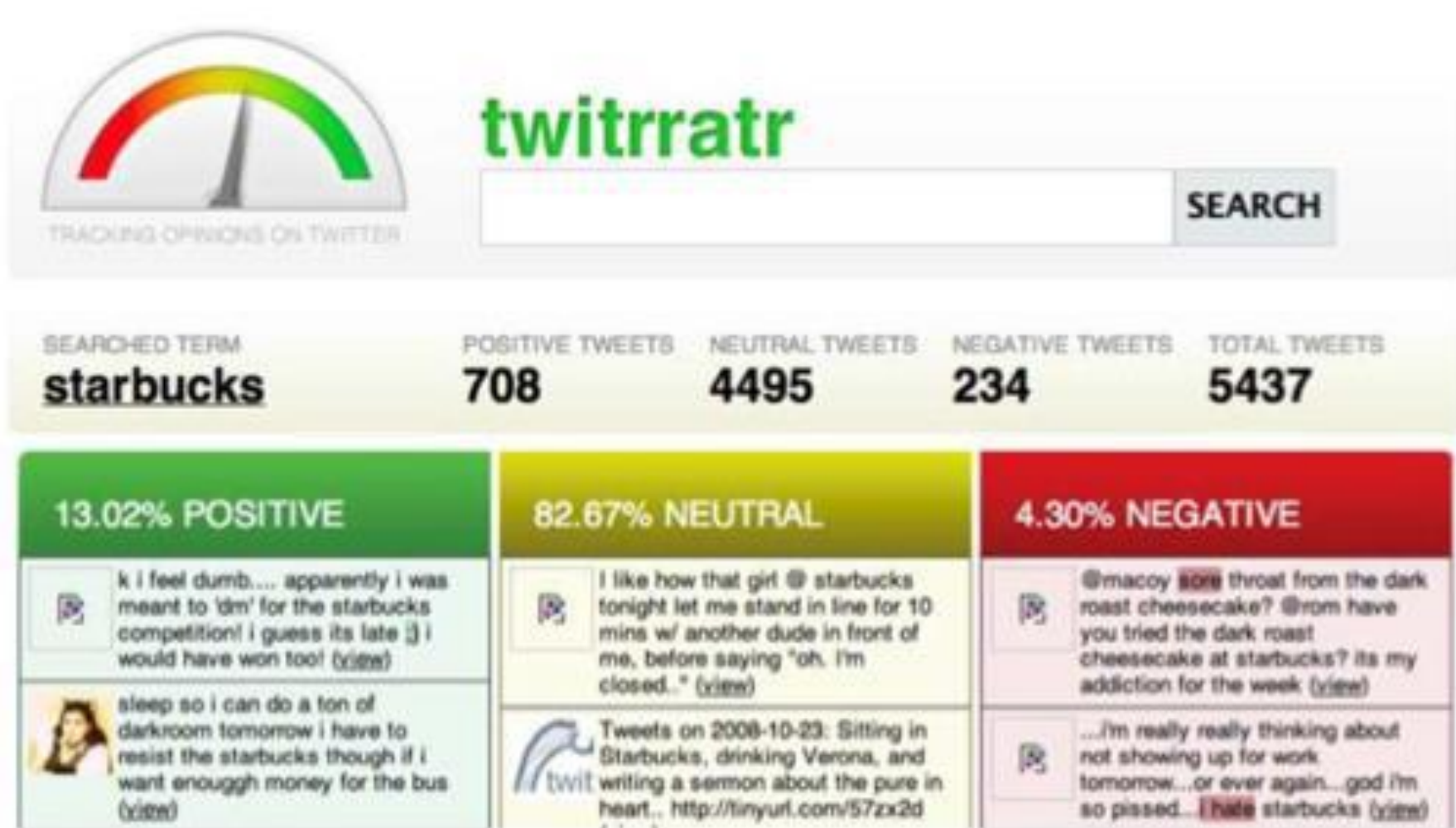
[Free](#)

[Documentation License](#)



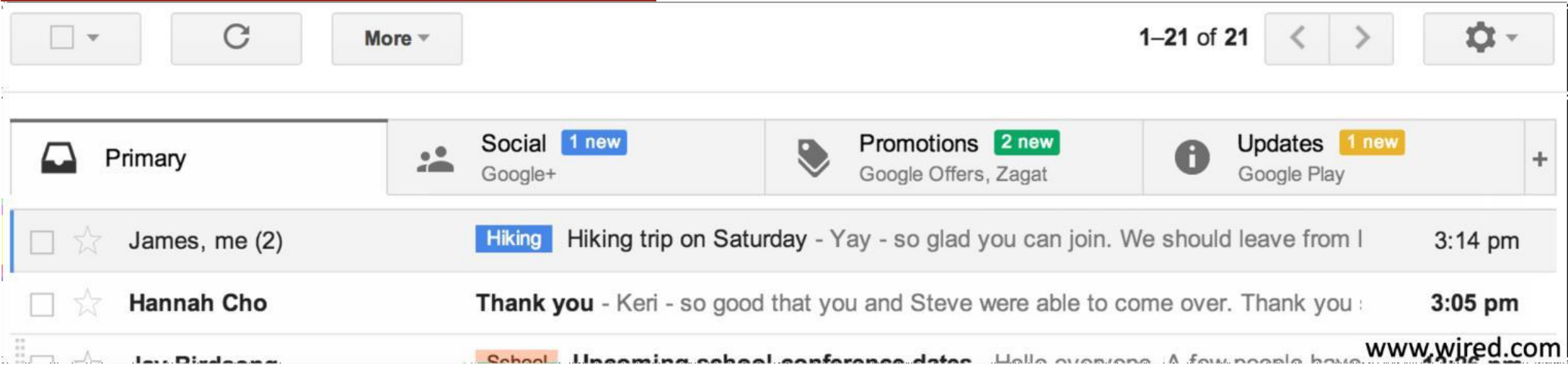
(GFDL). Images and other files are available under [different terms](#), as detailed on

Sentiment/Opinion Analysis



Text Classification

Other applications?



Question answering



- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?

Question answering


Go beyond search

What's the capital of Wyoming?

Web Maps Shopping Images News More Search tools

About 984,000 results (0.54 seconds)

Wyoming / Capital

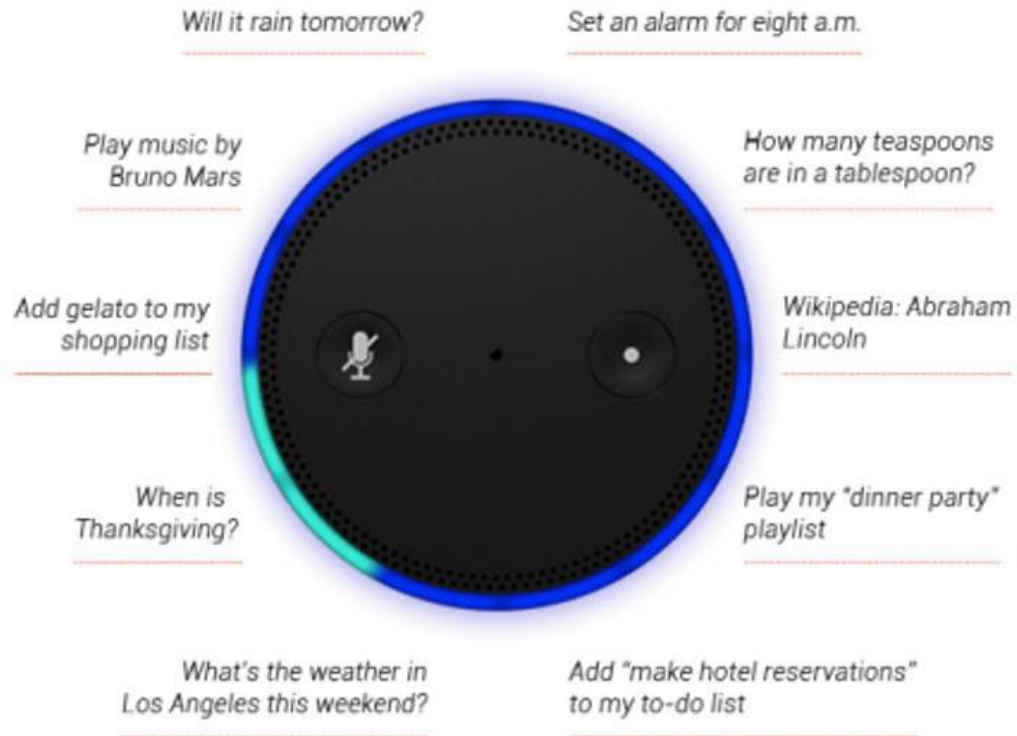


Cheyenne

Map data ©2015 Google

Natural language instruction

Digital personal assistant



The NLP would be much broader!

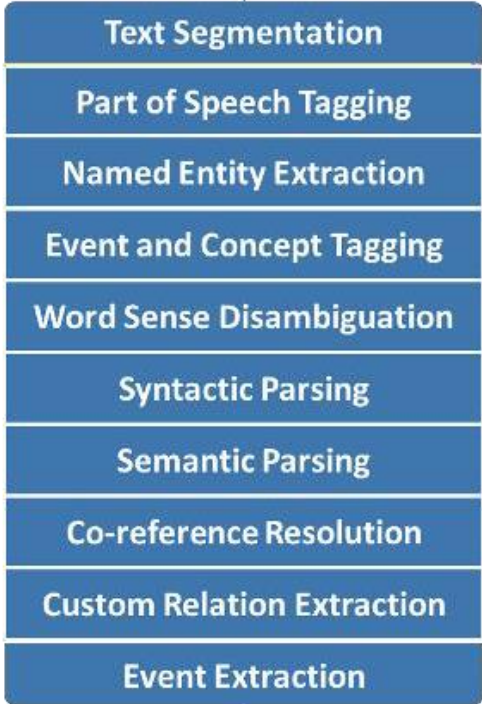
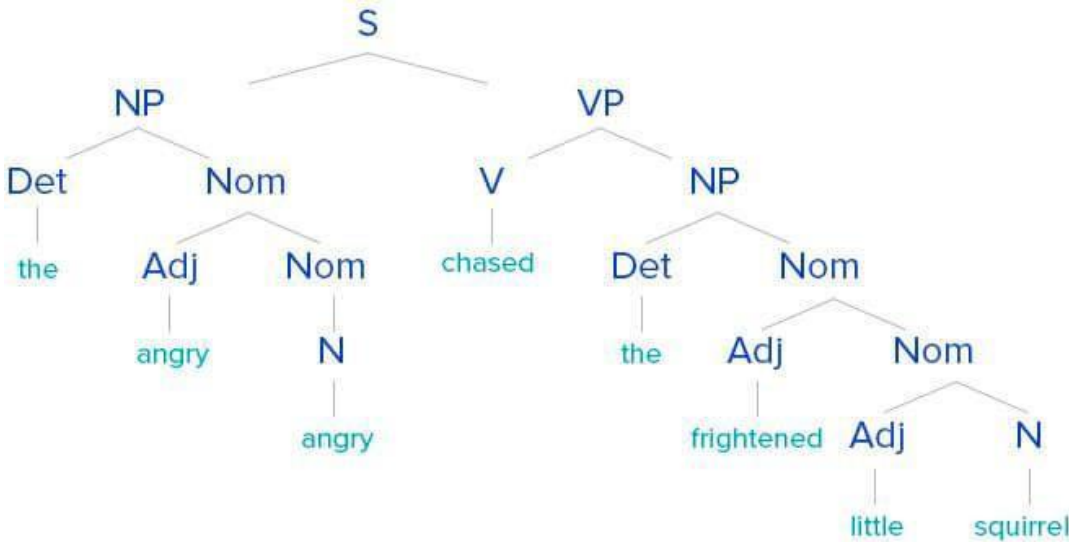
Pre-transformer era of NLP



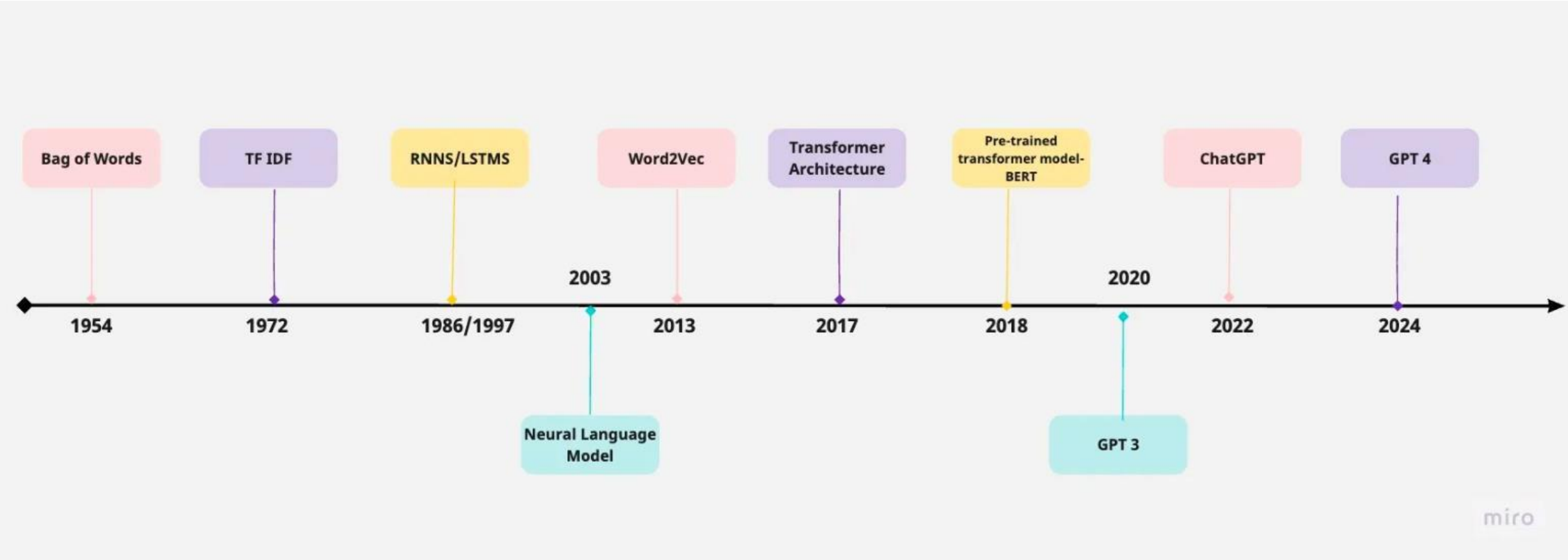
“Find me an Italian restaurant in New York City.”
action food type search type location



“And what’s the weather there tomorrow?”
search type location time

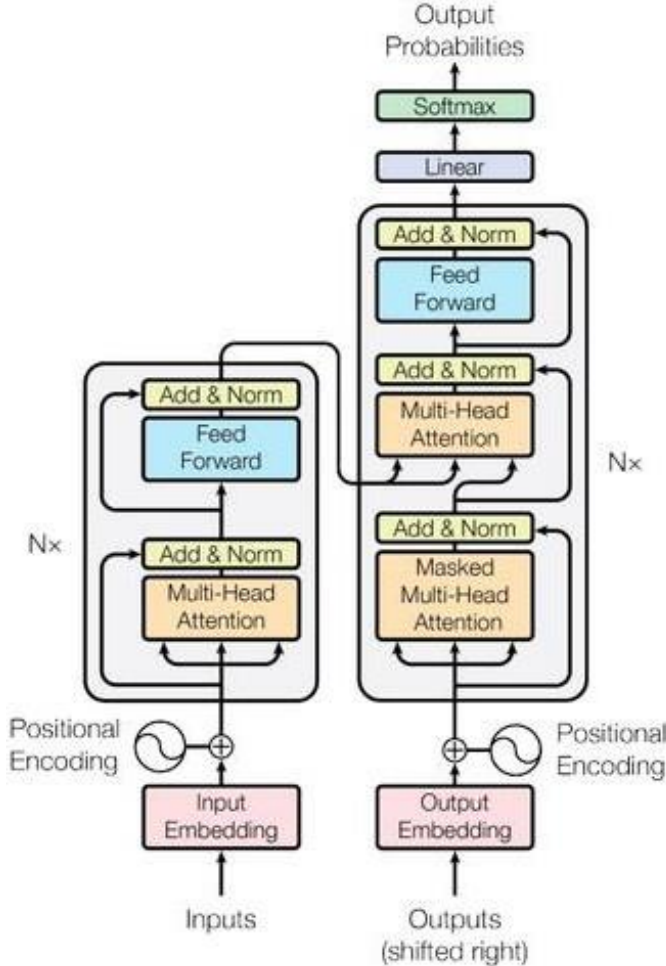
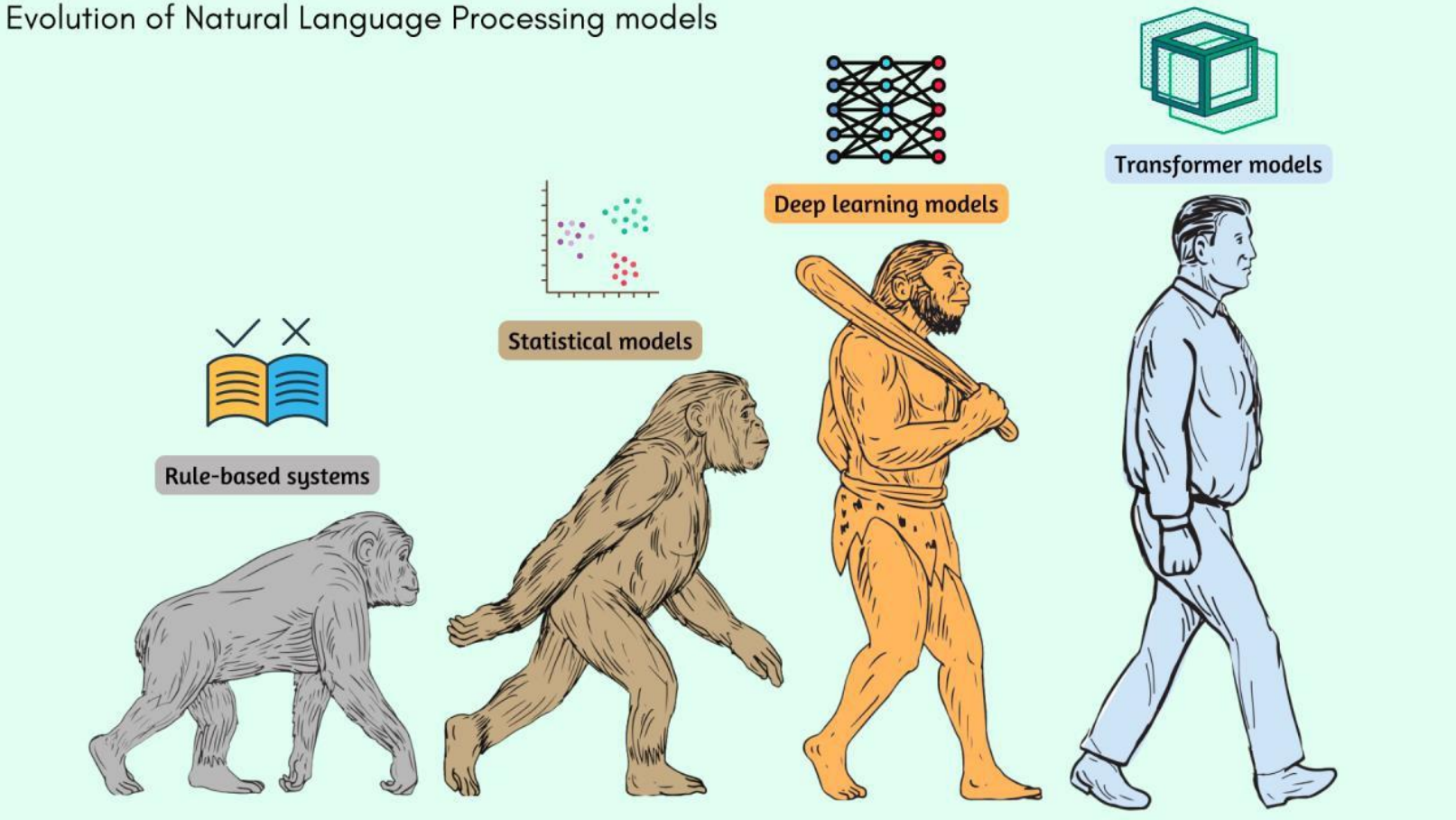


A brief timeline of the evolution of NLP



Transformer era of NLP

Evolution of Natural Language Processing models



<https://www.linkedin.cn/incareer/pulse/from-rulesets-transformers-journey-through-evolution-sota-yeddula>

Where Are We Now?



ChatGPT-5

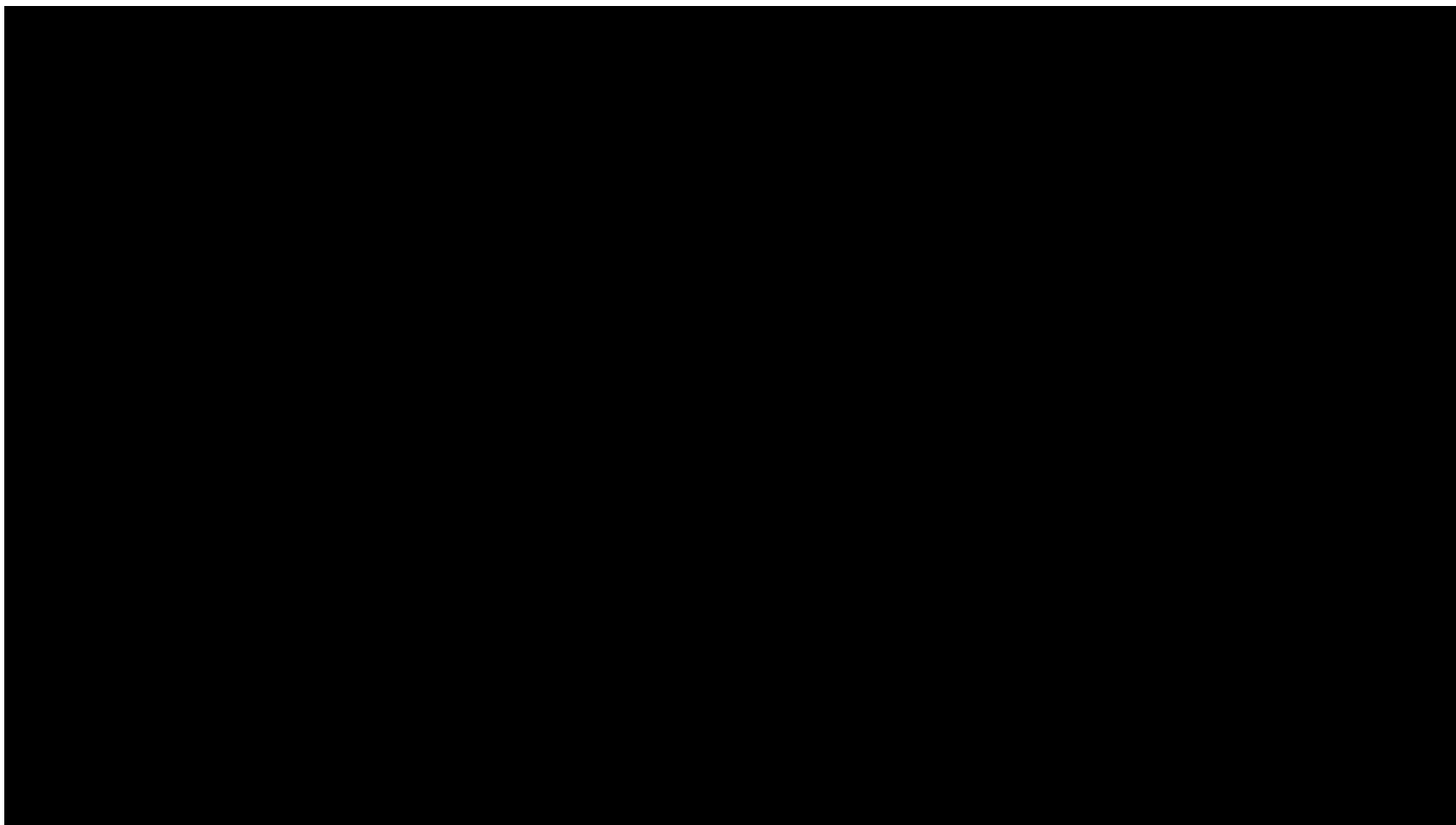


VS





<https://www.douyin.com/video/7461889667558886690>



Personal assistant: Her

Watch this movie : <https://www.imdb.com/video/vi4112492569/>

Her with video generation



Human



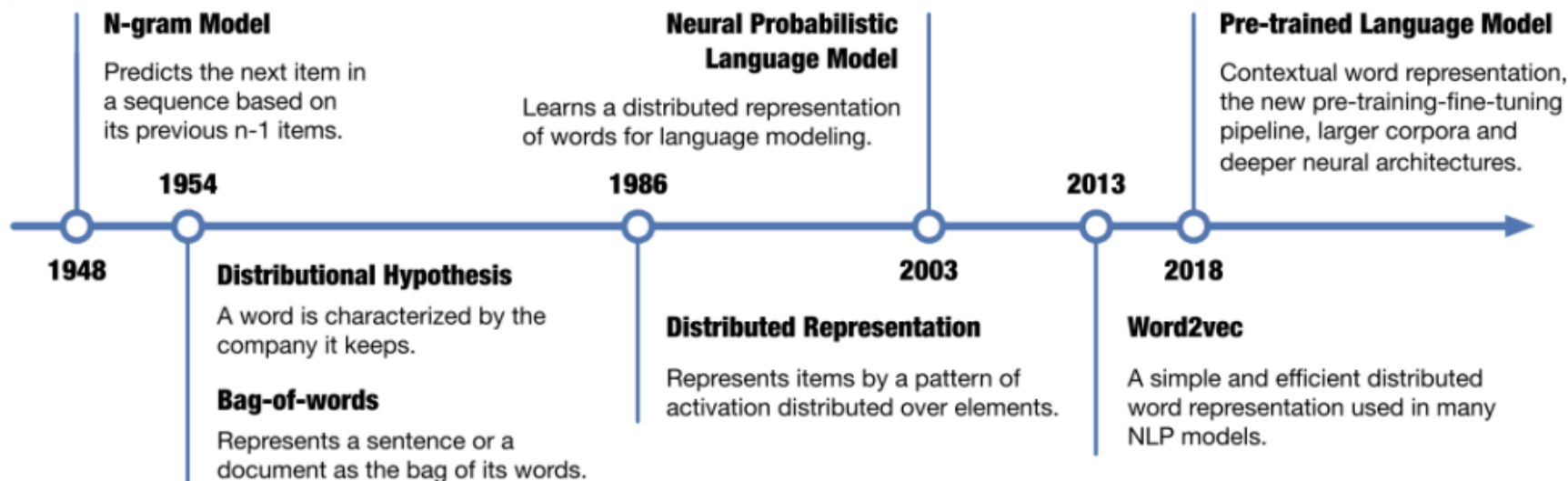
Agent (generated)

<https://aubrey-ao.github.io/BodyOfHer/>

From old NLP to Large Language models (LLMs)?

Background

- language model



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 fskjfhkjsh kjfs fs kjhkjhs fsjhfkshkjfh

Low probability



ChatGPT is all you need

high probability

What is language modeling?

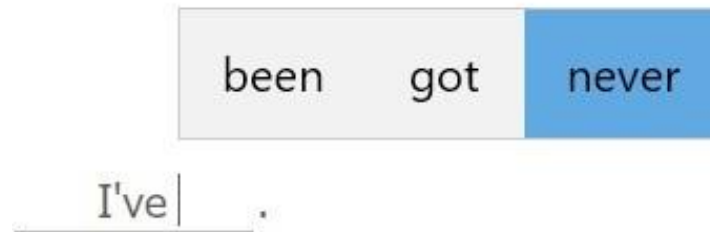
A **language model** assigns a probability to a N-gram

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

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

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

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



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

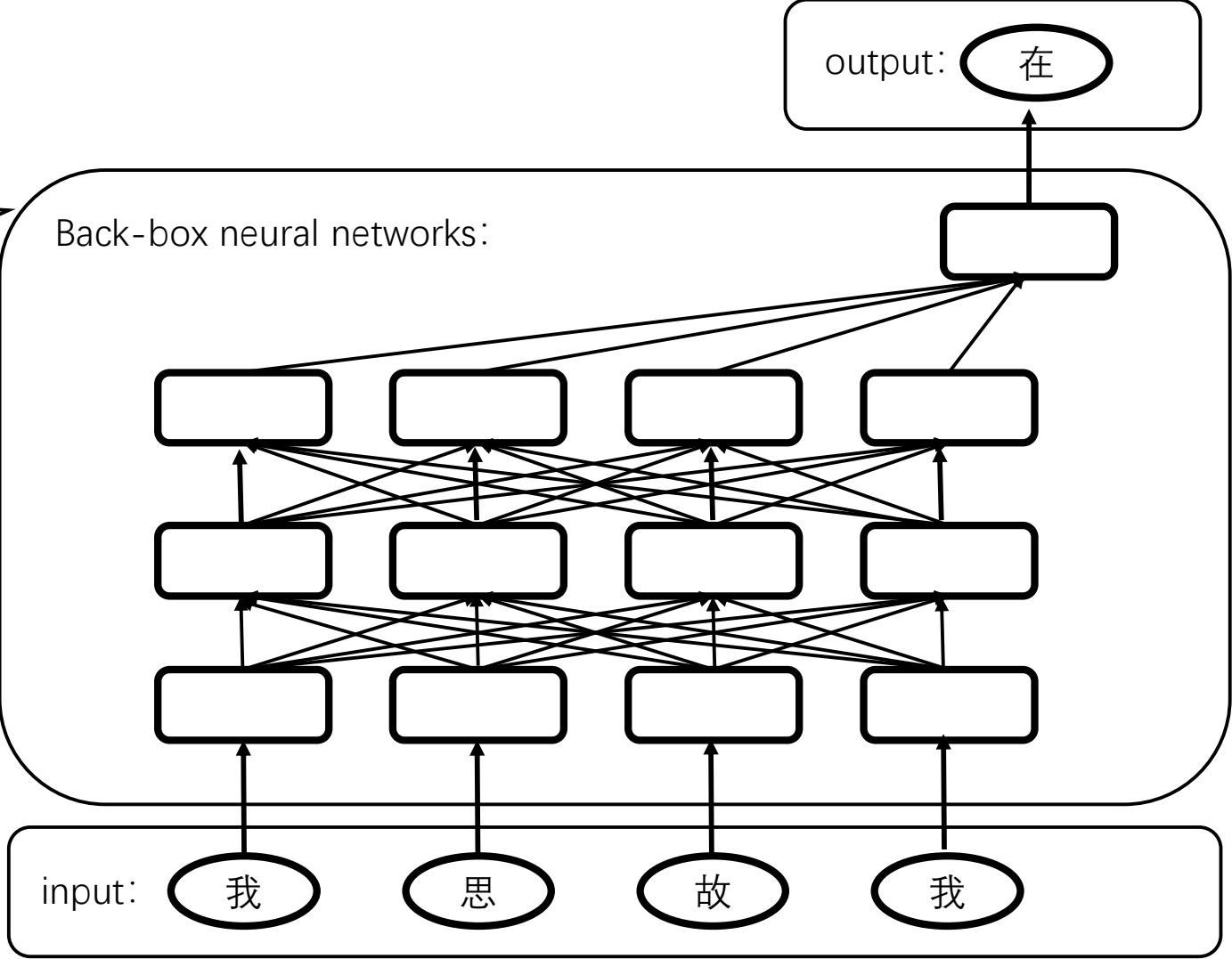
$$g: (V^{n-1}, V) \rightarrow R^+$$

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

$p(w_n | w_1 \cdots w_{n-1})$ 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"

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

Implicit order

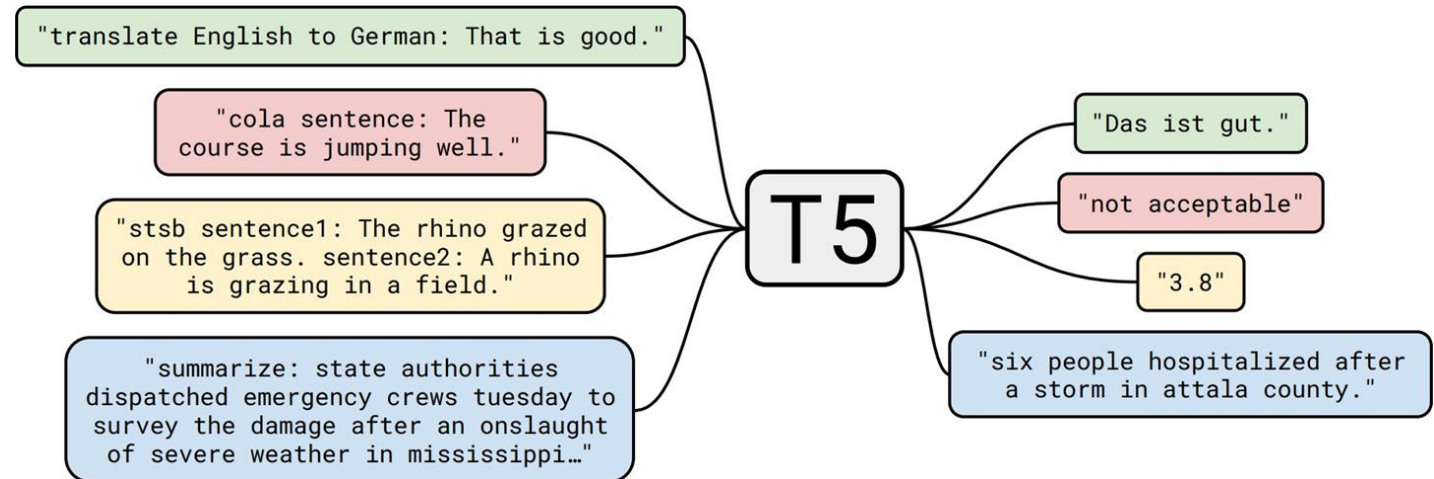
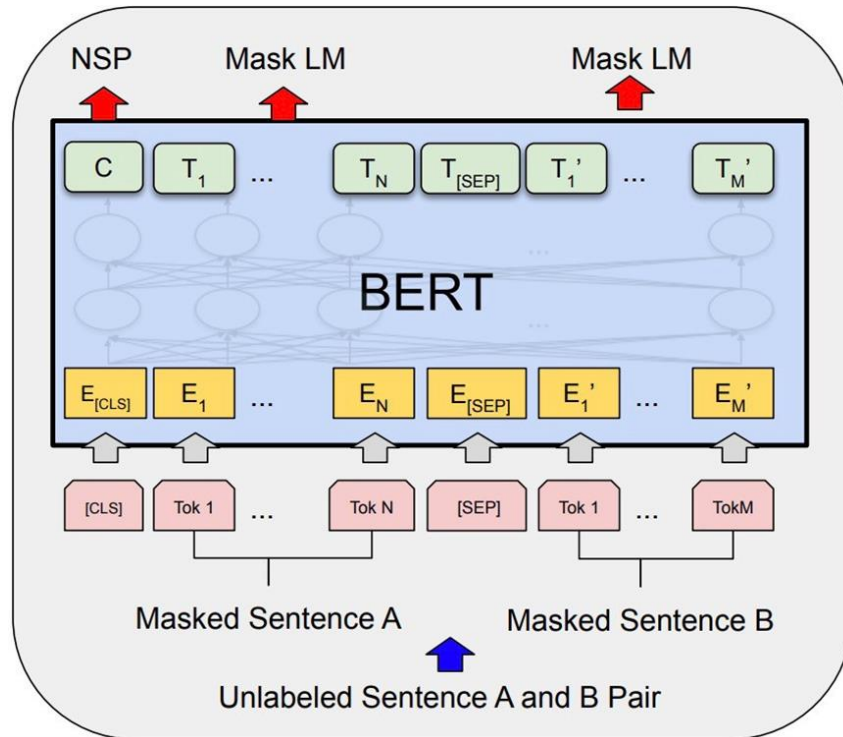


GPT-3 still acts in this way but the model is implemented as a very large neural network of 175-billion parameters!

Language models: Broad Sense

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

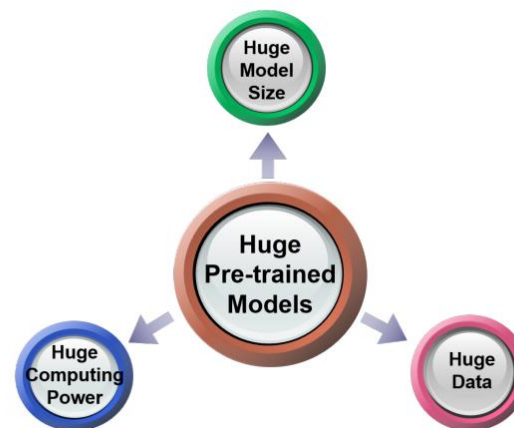
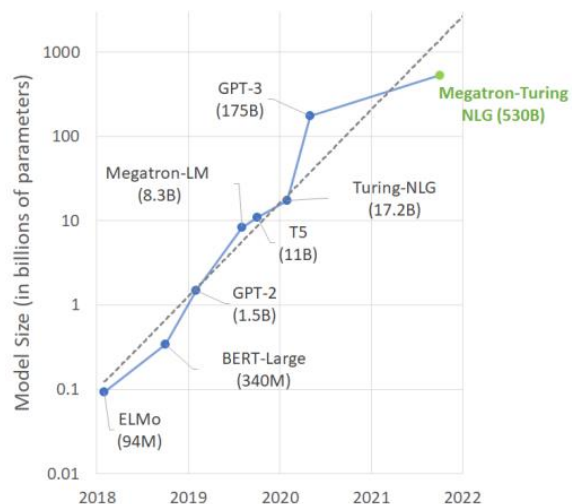
The latter two usually involve a different **pre-training** objective.



PLM vs. LLM

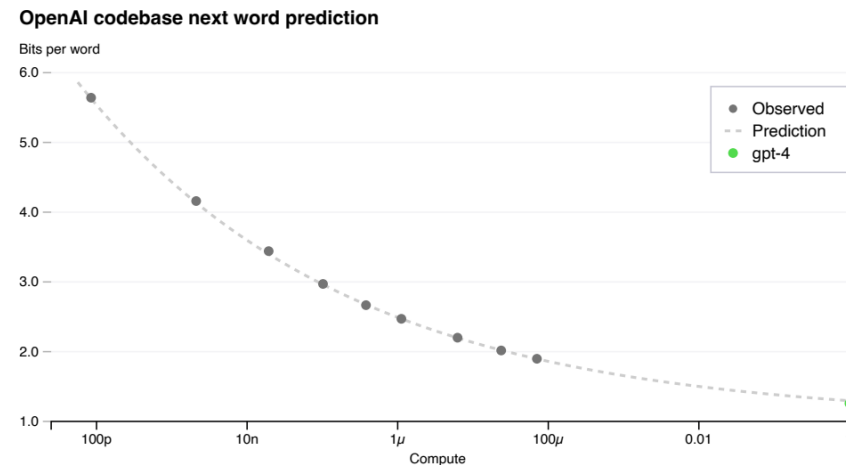
We do not explicitly mention pre-training because pre-training and training use the same language models objective (e.g., autoregressive generation)

- Pre-trained language model
- **Large pre-trained Language Model (LLM)**



Some basics for large language models

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

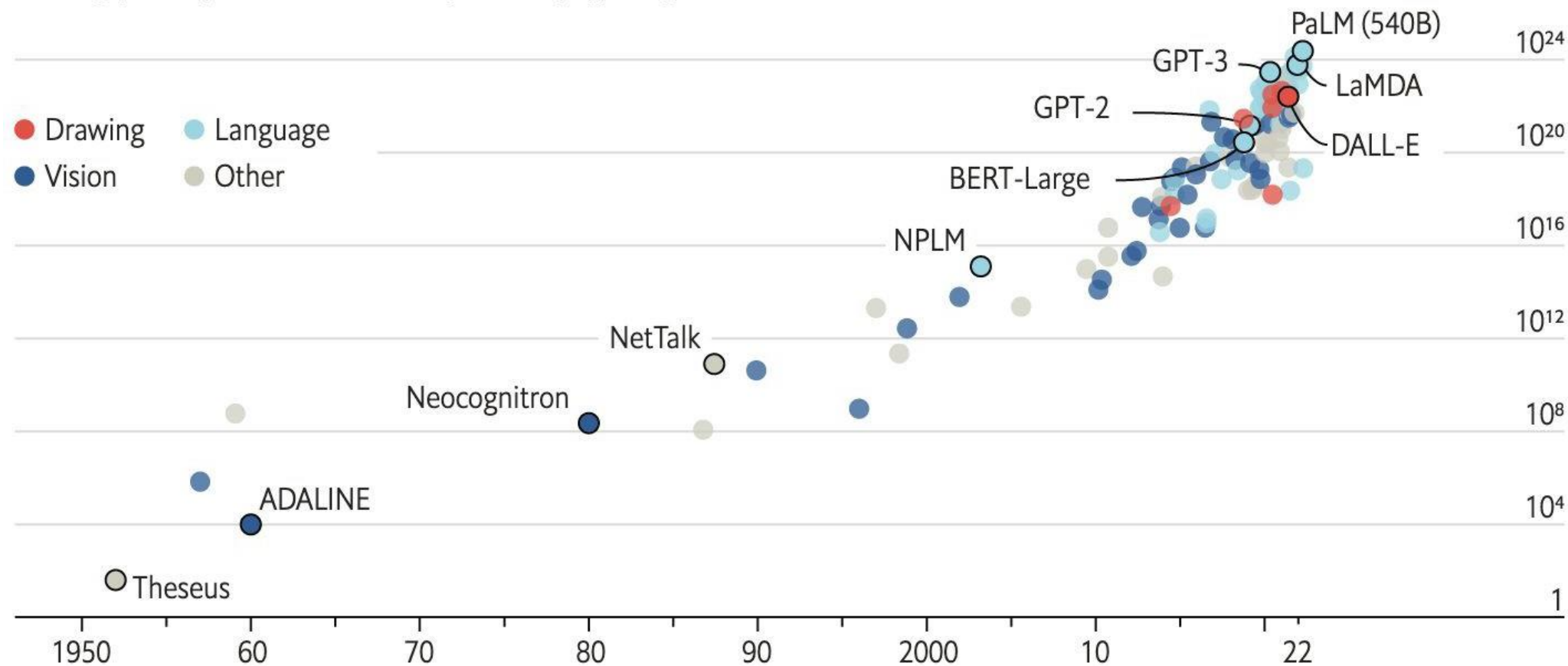


- Scalable **data** (plain texts are everywhere vs. supervised data)
 - <https://github.com/esbatmop/MNBVC>

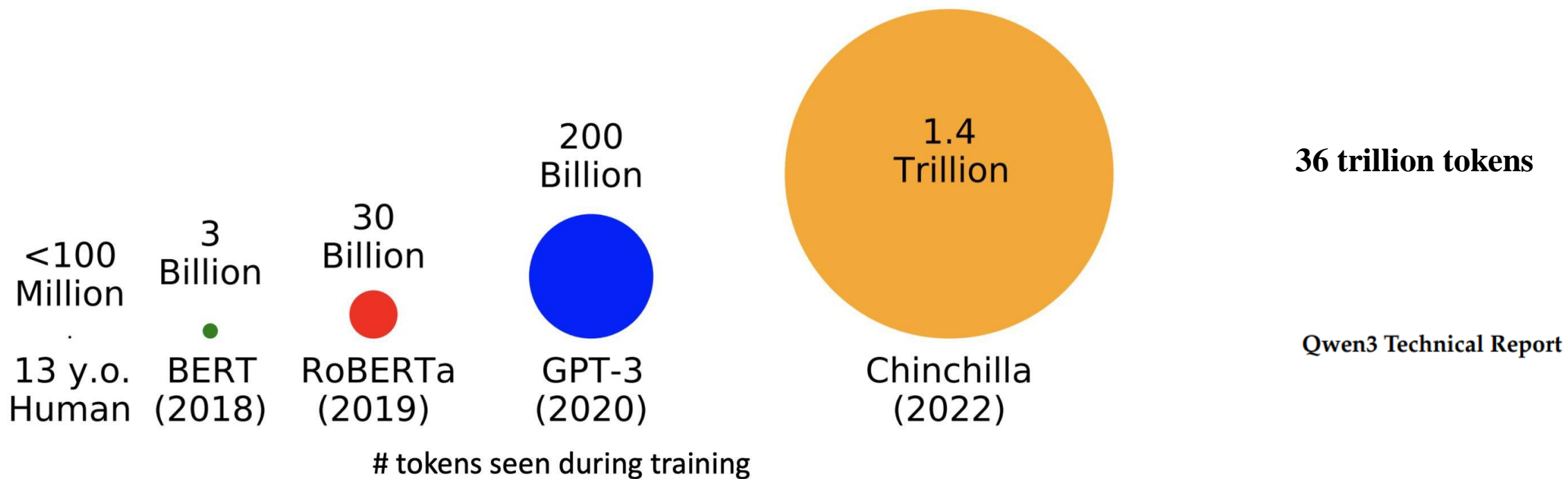
Large Language Models

AI training runs, estimated computing resources used

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



Large Language Models - **Hundreds of Billions of Tokens**



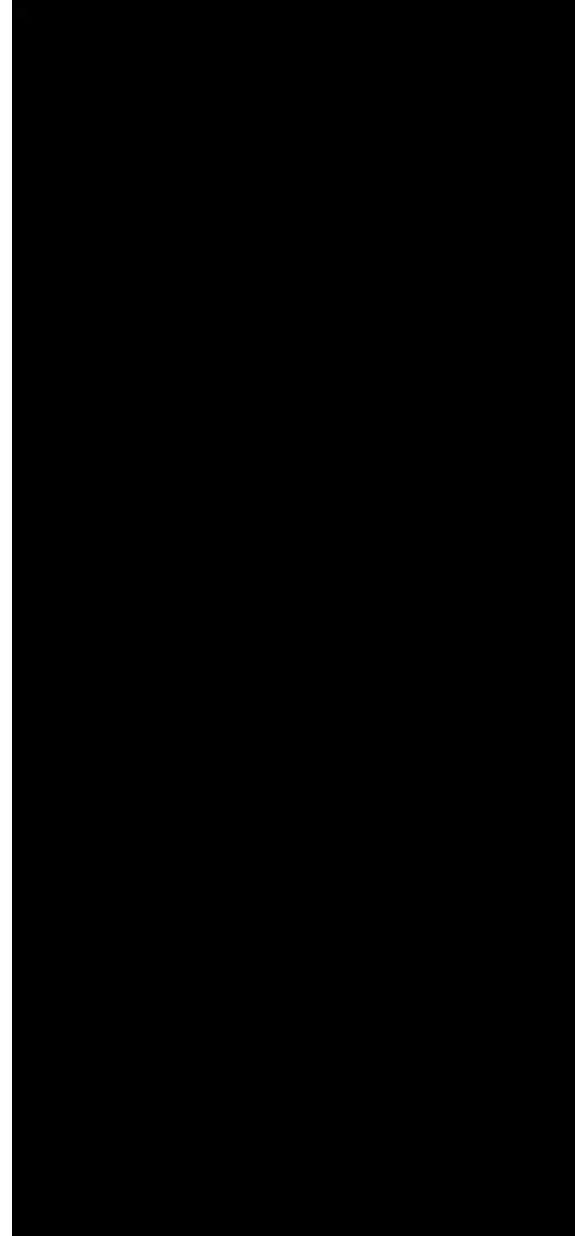
How Large are “Large” LMs?

- ❖ 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



From LLMs to **Agent**

Mobile Agent



<https://github.com/X-PLUG/MobileAgent>

Game Agent



<https://craftjarvis.github.io/ROCKET-1/>

Shaofei Cai, Zihao Wang, Kewei Lian, Zhancun Mu, Xiaojian Ma, Anji Liu, Yitao Liang. ROCKET-1: Master Open-World Interaction with Visual-Temporal Context Prompting. <https://arxiv.org/abs/2410.17856>

Embodied AI



<https://mobile-aloha.github.io/>

Contents

- About me
- Philosophy of this course
- NLP and large language models (LLM)
- **Introduction to our research**

Survey 1: Jobs in the future

- Which jobs are less likely be replaced by AI
 - A. Doctors
 - B. Mathematicians
 - C. Farmers
 - D. Software Engineers
 - E. AI Engineers
 - F. Blue-collar workers (蓝领工人)
 - G. White-collar workers (白领工人)

Survey 2: Intelligence Law

- How much human work time can AI replace at a decent quality level?
- *目前、AI能够以体面的效果完成“最好的人类”多少时间能够完成的的工作？*
 - A.** Less than 1 minute
 - B.** 1 minute to 1 hour
 - C.** 1 hour to 10 hours
 - D.** 10 hours to 100 hours
 - E.** More than 100 hours

Survey 3: AI replaced Humans?

- In the future (**to the year of 2030**), how much of human work will AI replace?
- A. Less than 20%
- B. 20–50%
- C. 50–80%
- D. 80–95%
- E. More than 95%
- F. 100%

Survey 4: AI replaced Humans?

- In the future (**to the year of 2050**), how much of human work will AI replace?
- A. Less than 20%
- B. 20–50%
- C. 50–80%
- D. 80–95%
- E. More than 95%
- F. 100%

Thanks

Join us as a RA if interested
wangbenyou@cuhk.edu.cn