
UF
Herbert Wertheim
College of Engineering
UNIVERSITY of FLORIDA
CIS 6930 Special Topics in Large Language Models
Overview & Introduction
LEADING THE CHARGE, CHARGING AHEAD

Introduction about myself

Yuanyuan Lei

Assistant Professor

Department of Computer and Information Science and Engineering

Research Area: Natural Language Processing, Large Language Models

Personal website: https://sites.google.com/view/yuanyuan-lei

Introduction about the course

Fall 2025, MWF 11:45am – 12:35am at CSE E222

Instructor

- Yuanyuan Lei (yuanyuan.lei@ufl.edu)
- Office Hours: Friday 1:30pm – 2:30pm, Malachowsky Hall 3108, or Zoom https://ufl.zoom.us/my/yuanyuan.lei

TA

- Zixuan Wang (zwang10@ufl.edu)
- Office Hours: Friday 3pm – 4pm, Malachowsky Hall 5200, or Zoom https://us05web.zoom.us/j/83030672462?pwd=cgo2rzJOTYX3aGVvcGHvDzvBIubznQ.1 (Meeting ID: 830 3067 2462 Passcode: 6Utsg9)

Outline of this lecture

- What is Large Language Models

  (language model, large language model)

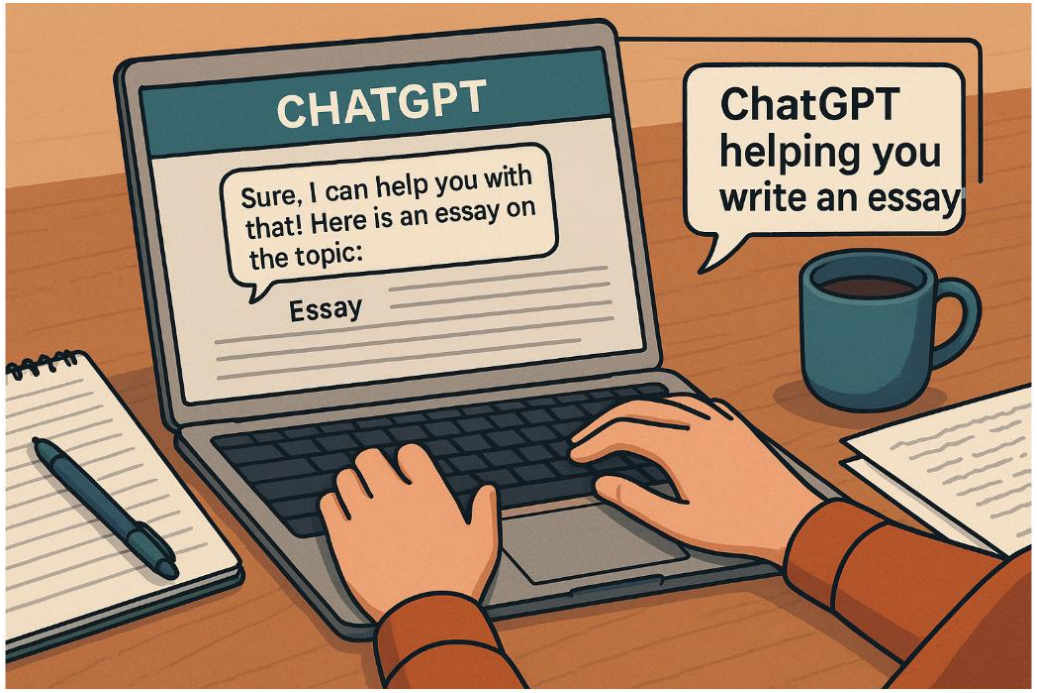
- Why we study Large Language Models

  (capabilities, limitations)

- How will we study Large Language Models

  (course content, grading components etc.)

Large Language Models



Large Language Models

Large Language Models can be your helpful assistant

- Writing (draft essays, reports, blog posts, improve language)
- Learning (explain concepts, search information, summarize)
- Creativity (brainstorm ideas, generate poems)
- Even for fun (give recommendations for book, movie, music)
- and many …

What is Language Models

Language Models (LM) is a probability distribution over sequences of tokens. Suppose we have a vocabulary V of a set of tokens. LM p assigns each sequence of tokens $w_1, \ldots, w_n \in V$ a probability

$$p(w_1, \ldots, w_n)$$

This probability intuitively tells us how "good" a sequence of tokens is.

Example:

p(the, mouse, ate, the, cheese) = 0.02

p(the, cheese, ate, the, mouse) = 0.01

p(mouse, the, the, cheese, ate) = 0.0001

What is Language Models

LM should have the ability to assign meaningful probability to all sequences, which requires sophisticated (but implicit) linguistic abilities and world knowledge.

Example:

p(the, mouse, ate, the, cheese) = 0.02 -> higher prob, world knowledge

p(the, cheese, ate, the, mouse) = 0.01

p(mouse, the, the, cheese, ate) = 0.0001 -> ungrammatical, syntactic knowledge

What is Language Models

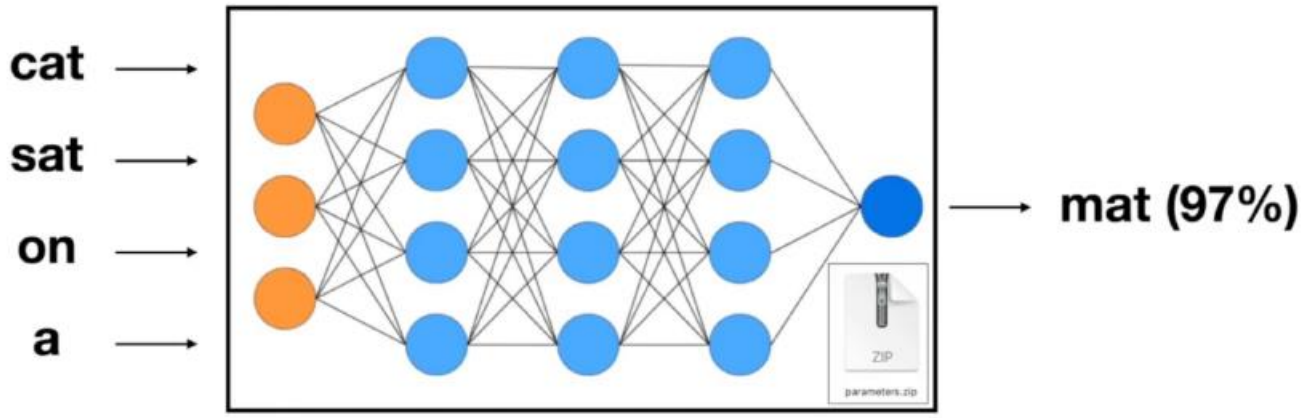
LM takes a sequence and returns a probability to assess its goodness.

$$p(w_1, \ldots, w_n)$$
$$= \mathrm{p}(w_1) * p(w_2|w_1) * p(w_3|w_1, w_2) * \ldots * p(w_n|w_1, w_2, \ldots, w_{n-1})$$

p(cat sat on the mat) = p(cat) * p(sat | cat) * p(on | cat sat) *
p(the | cat sat on) * p(mat | cat sat on the)

What is Language Models

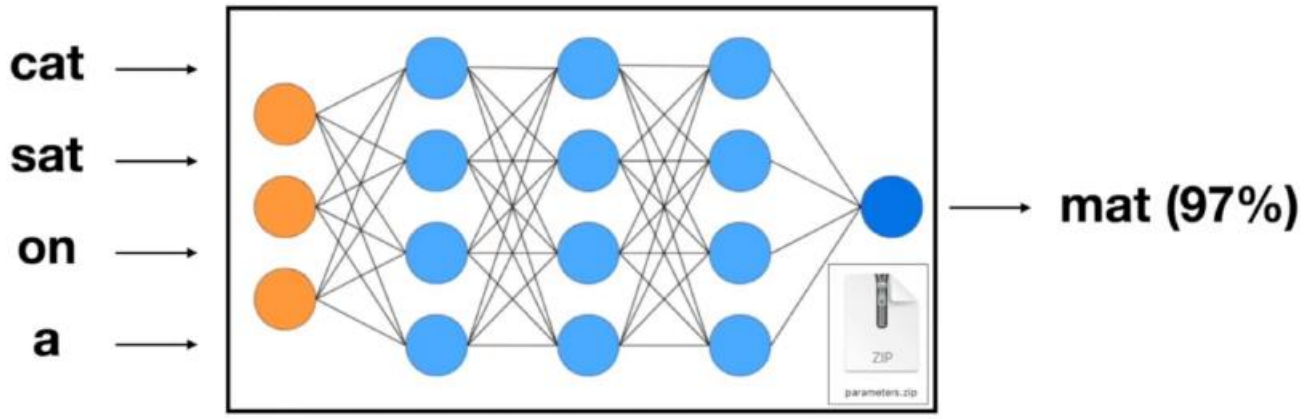
- LM takes a sequence and returns a probability to assess its goodness
- We can also generate a sequence given a LM, usually by generating the next tokens given previous tokens



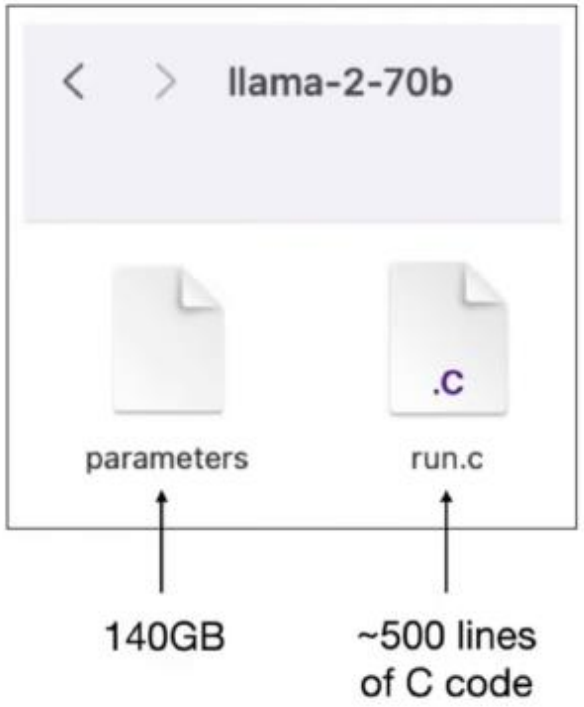
$$f(\text{“cat sit on a”}; \theta) = \text{“mat”}$$

What is Large Language Models

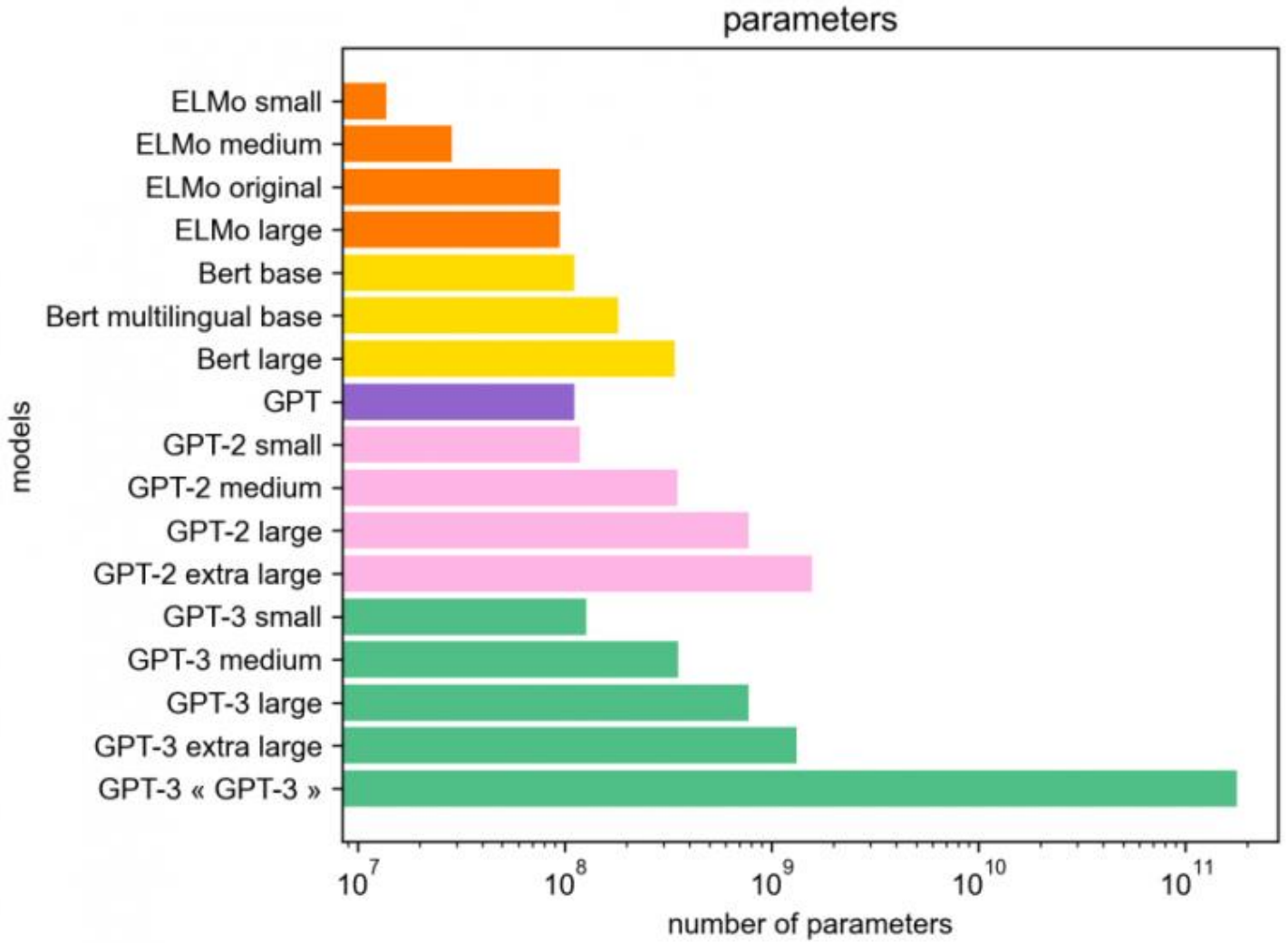
Model parameters θ plays an important roles

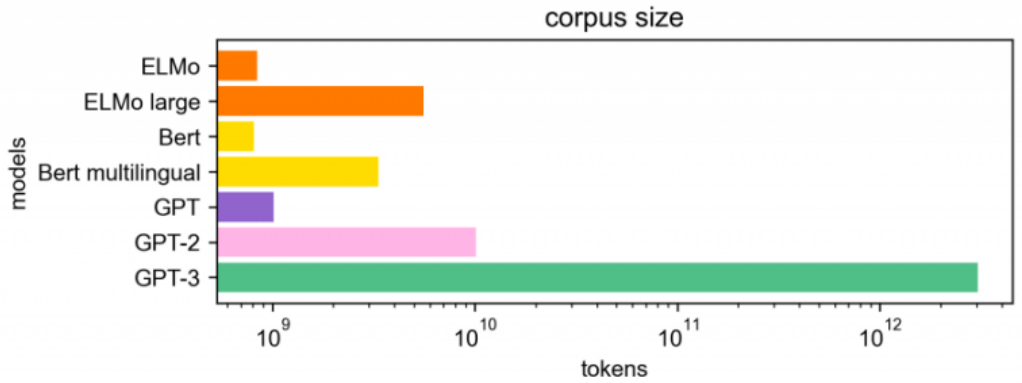


$$f(\text{“cat sit on a”}; \theta) = \text{“mat”}$$



How large are LLM





What is Large Language Models

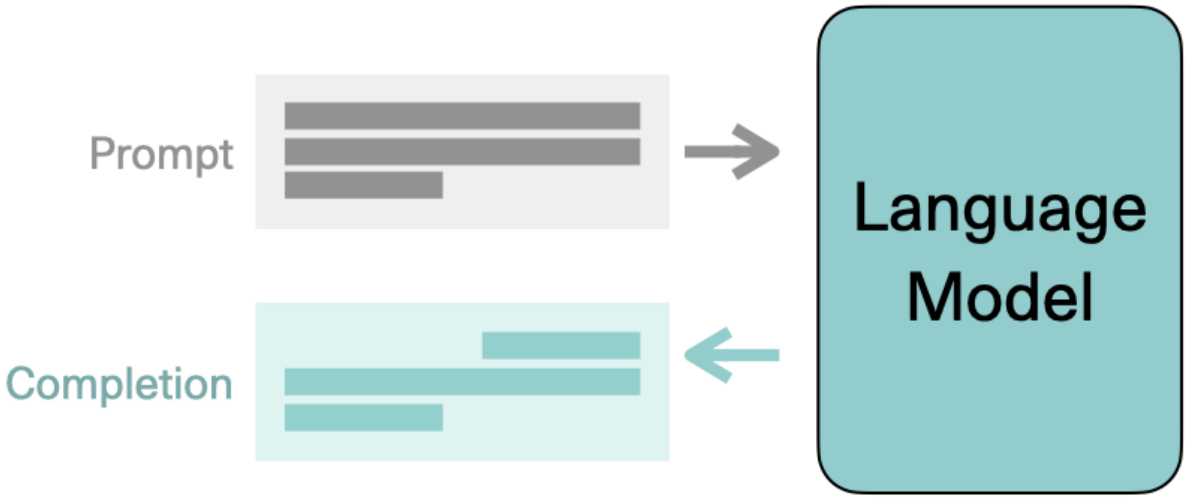
- **Small models**: under **100 million** parameters (useful for lightweight tasks, edge devices, or fine-tuning)
- **Medium models**: **100M – 1B** parameters (comparable to early transformer models like BERT-base)
- **Large models**: **1B – 10B** parameters (often considered “large” in many research papers).
- **Very large models**: **10B – 100B** parameters (models like GPT-3 with 175B fall here).
- **Frontier-scale models**: **100B+** parameters (GPT-4, Claude 3, Gemini 1.5 Pro, etc.).

Why study Large Language Models

- Reason 1: Capability

One single model to solve many NLP tasks

recall conditional generation:



Why study Large Language Models

- Reason 1: Capability

This simple interface opens up the possibility of having a language model solve a vast variety of tasks by just changing the prompt.

Question-Answering

$$\text{Frederic, Chopin, was, born, in} \overset{T=0}{\rightsquigarrow} \text{1810, in, Poland}$$

Word Analogy

$$\text{sky, :, blue, ::, grass, :} \overset{T=0}{\rightsquigarrow} \text{green}$$

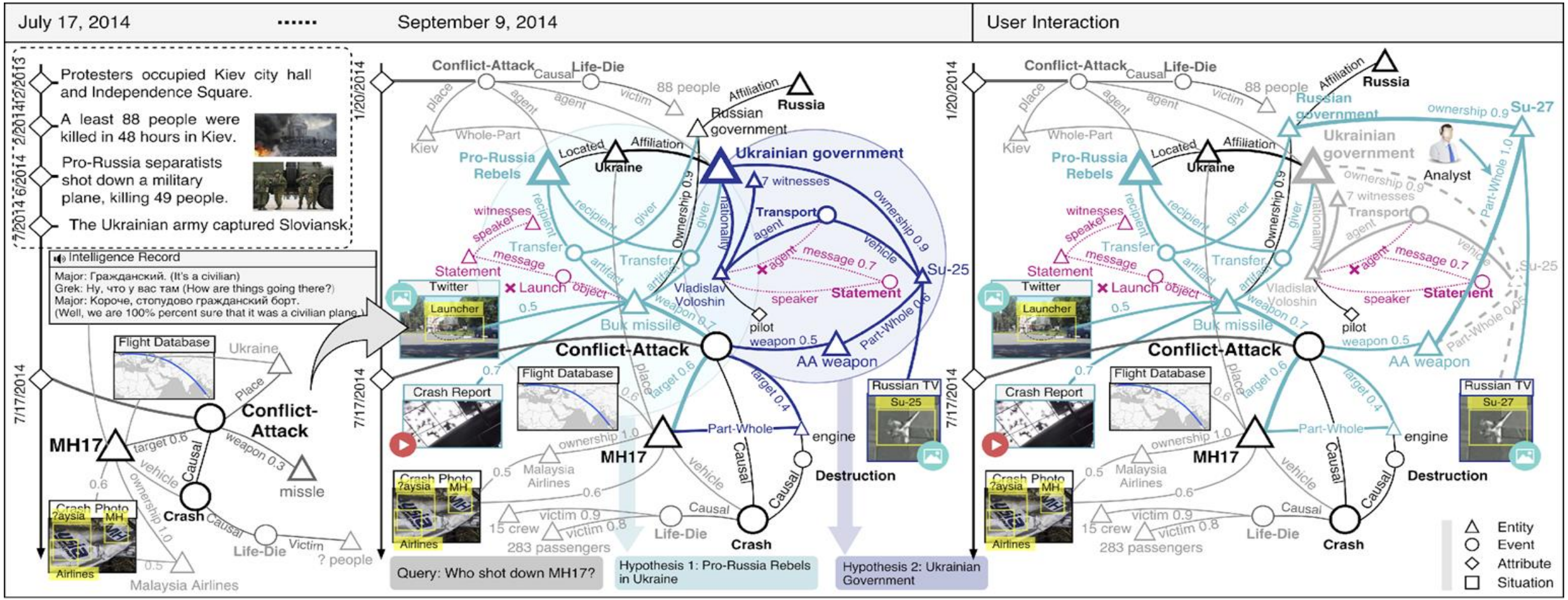
News Generation

Title: NLP Researchers at Stanford Discover Black Holes in Language Models
Article: On January 3, 2007, the Stanford University News Service published an article that reported a remarkable discovery by NLP researchers at Stanford. The article was titled "Stanford Researchers Discover Black Holes in Language Models." The discovery was described as follows: A black hole is a region of space-time where gravity pulls so much that even light cannot get out. Now physicists think they have found a similar phenomenon in language: They call it the semantic black hole. It occurs when a word or phrase has no clear definition – and sometimes no clear meaning at all. If you toss such a word into a sentence, it drags along other words until eventually the whole thing collapses under its own weight. "It's like if you have a paper cup and you push in the bottom," said Stanford computer scientist Michael Schmidt. "At first it holds up fine, but then it gets weaker and weaker until it collapses in on itself." Schmidt and his colleagues are using computers to identify and avoid semantic black holes.

Why study Large Language Models

- Reason 1: Capability

Intelligence Analysis



Why study Large Language Models

- Reason 2: In-context Learning

Input: Where is Stanford University?
Output: Stanford University is in California.

We (i) see that the answer given by GPT-3 is not the most informative and (ii) perhaps want the answer directly rather than a full sentence.

Why study Large Language Models

- Reason 2: In-context Learning

Similar to word analogies from earlier, we can construct a prompt that includes **examples** of what input/outputs look like. LLM manages to understand the task better from these examples and is now able to produce the desired answer.

Input: Where is MIT?
Output: Cambridge

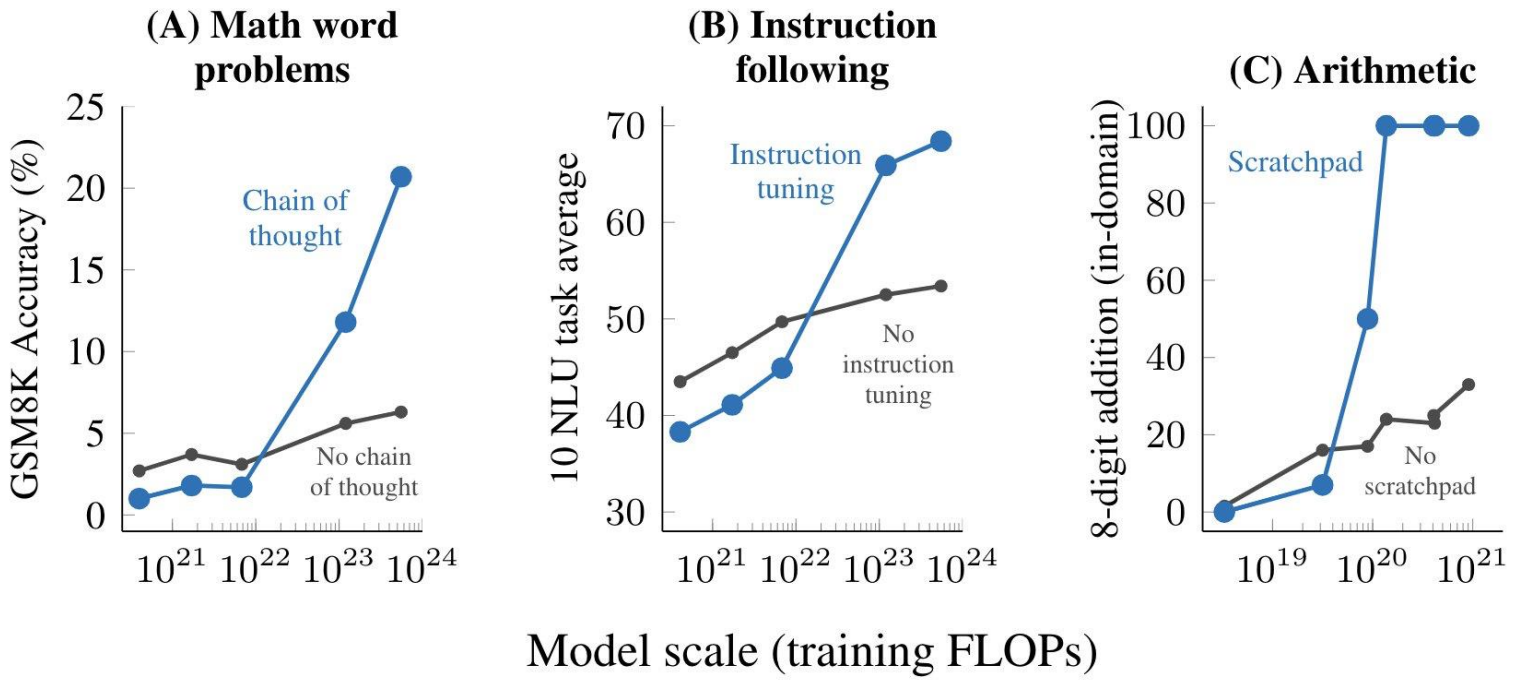
Input: Where is University of Washington?
Output: Seattle

Input: Where is Stanford University?
Output: California

Why Large Language Models

- Reason 3: Emergent properties in LLM

An ability not present in smaller models but present in larger models.



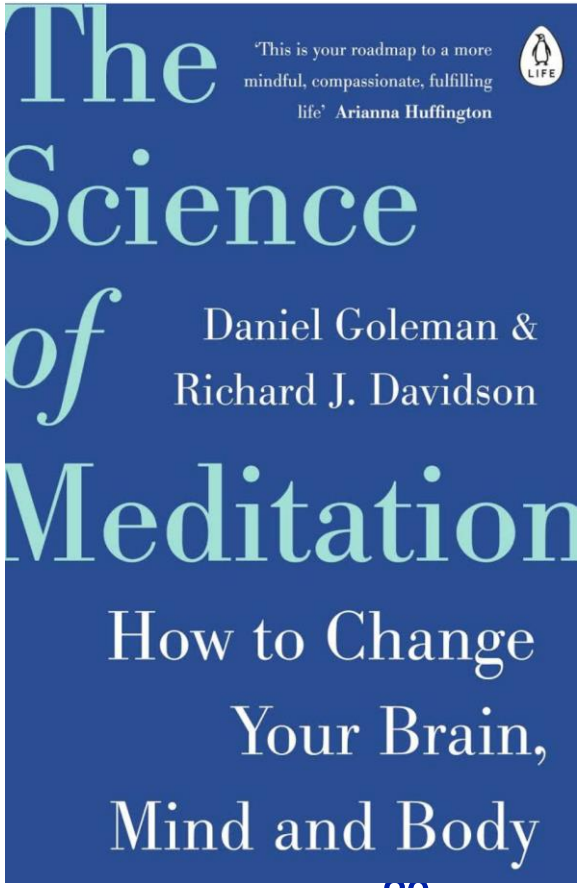
Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Truthfulness, the answer seems correct but actually not

BL What are some references from which I can learn more about the benefits of meditation?

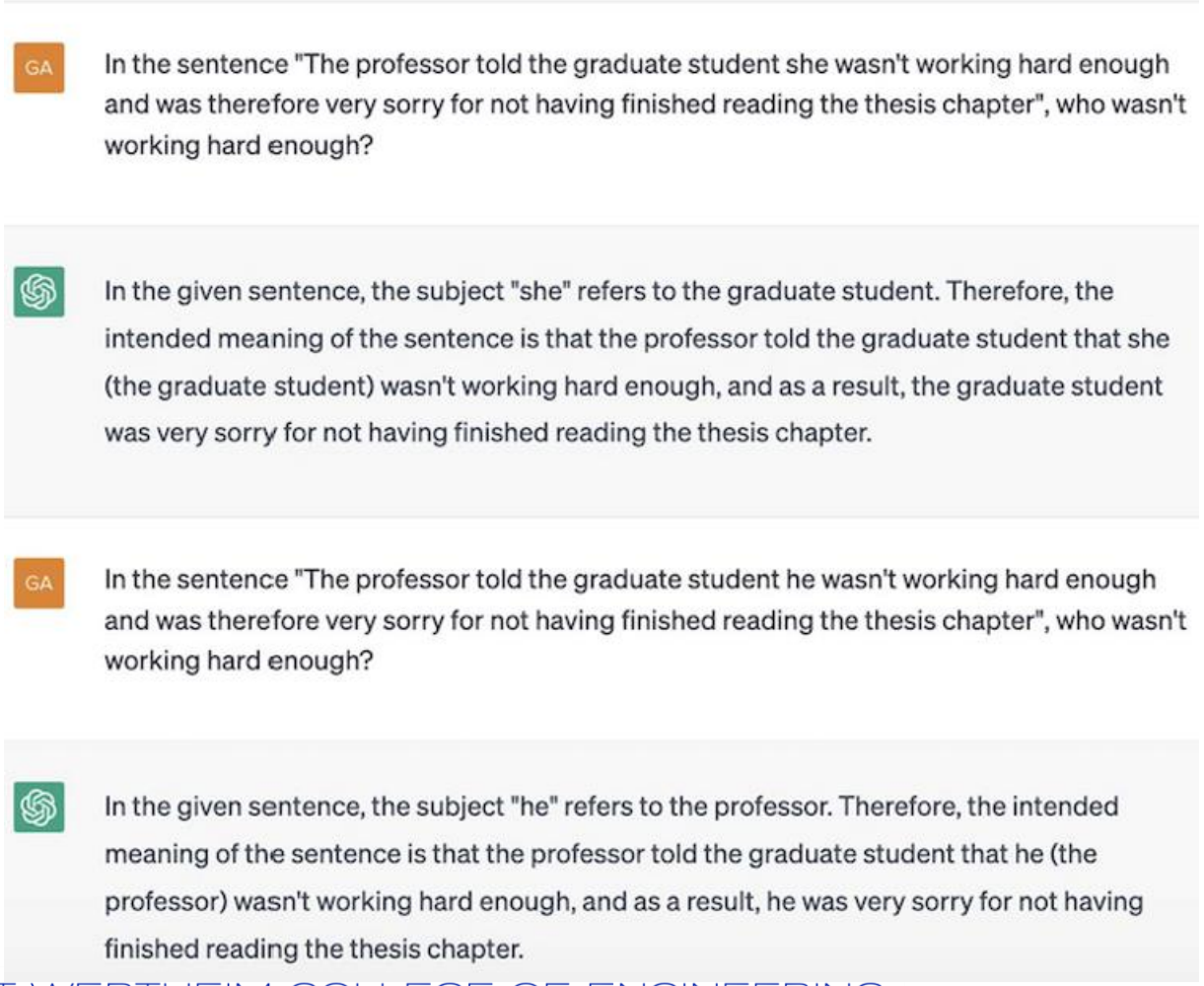
- "The Science of Meditation" is a book by medical doctor and meditation expert Dr. Amishi Jha. In the book, she discusses the latest research on the benefits of meditation and how it can be used to improve mental and physical health. It is available for purchase on Amazon and other online bookstores.



Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Social Bias



Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Toxicity. Large language models are trained on a huge amount of Internet data (e.g., Reddit), which inevitably contains offensive content.

Security and Privacy. Large language models can memorize, leak, or be manipulated into revealing sensitive data from training or user interactions.

Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Incapability in Knowledge Reasoning

HE

In January, Vadym Omelchenko, Kyiv's ambassador to France, said that Western countries were planning to send 300 tanks to Ukraine. Numerous countries have officially confirmed their agreement to deliver 321 heavy tanks to Ukraine, when did this happen?

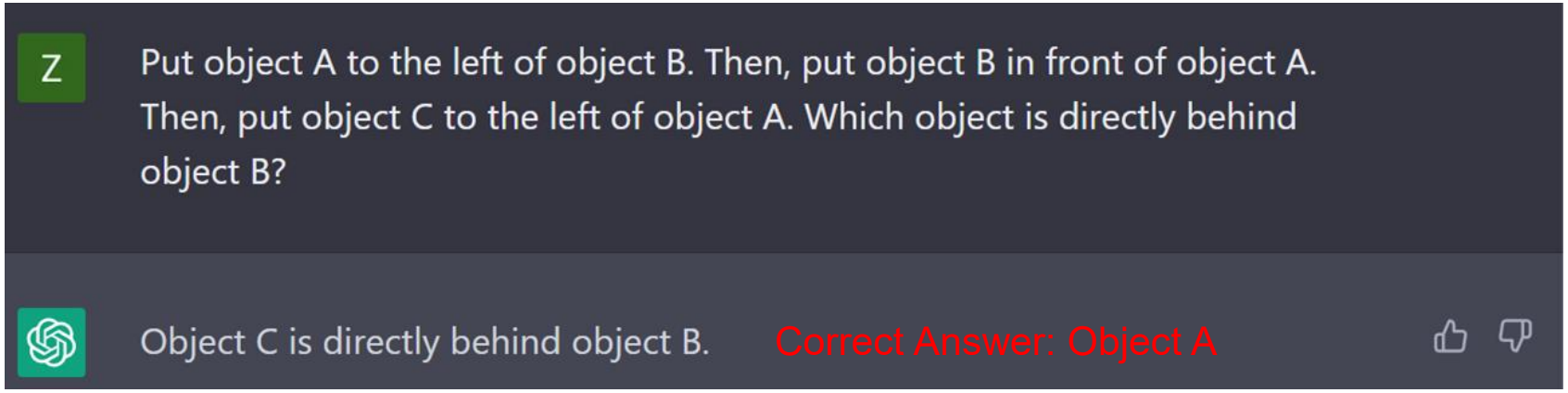
I'm sorry, but I do not have information on the specific details of the agreement mentioned by Vadym Omelchenko regarding the delivery of tanks to Ukraine. It's possible that this information is not publicly available, or that my training data only goes up until 2021. Can you provide any more context or details on the situation?

Why Large Language Models

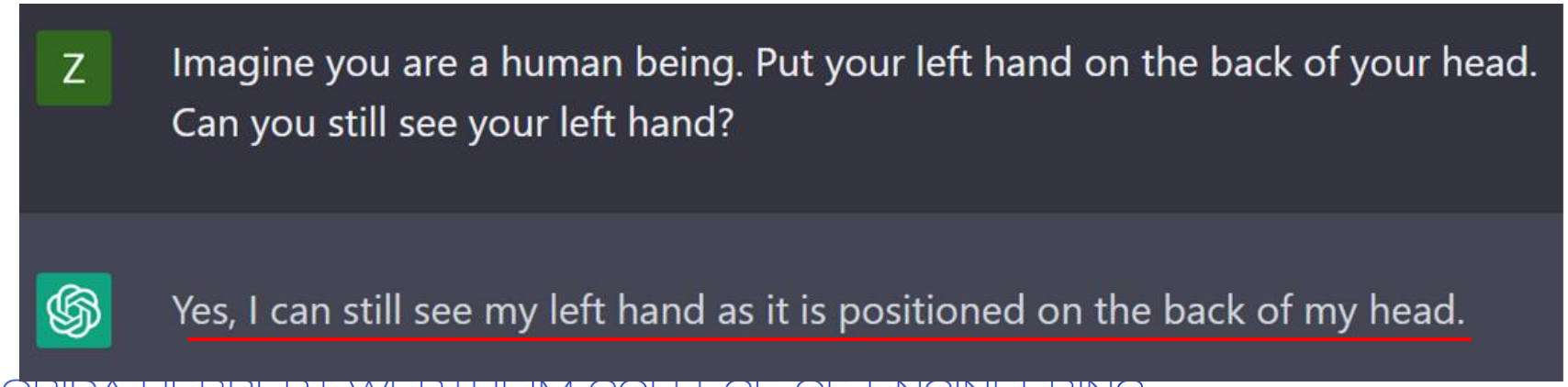
- Reason 4: LLM do exist lots of limitations

Incapability in Knowledge Reasoning

Spatial Relation



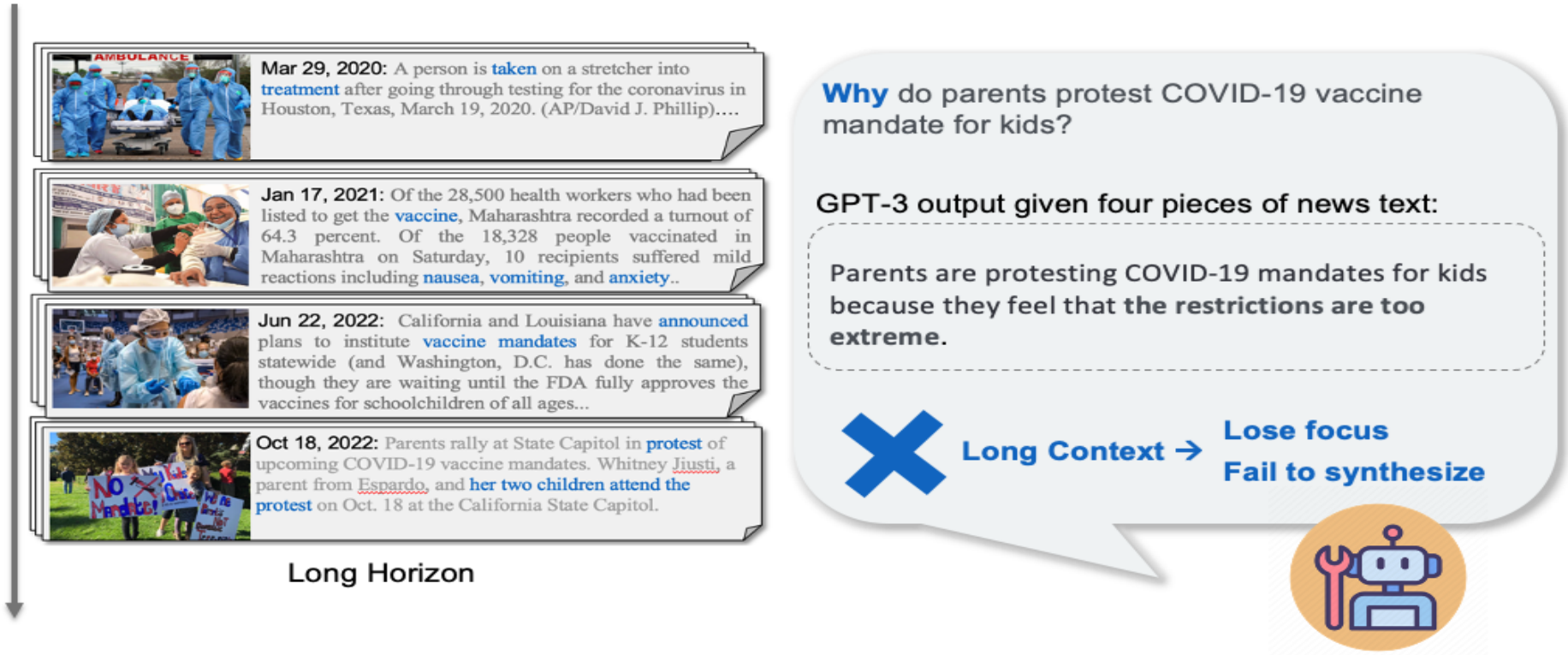
Knowledge about a physical form



Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Incapability in Long-Context Understanding



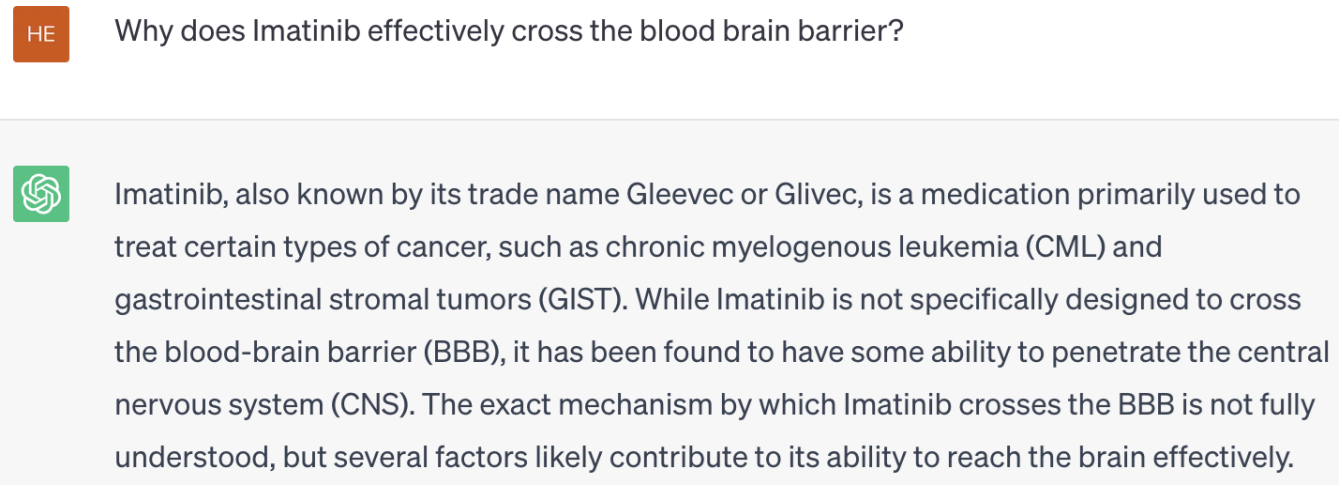
Limited short-term memory and disorganized long-term memory

Why Large Language Models

- Reason 4: LLM do exist lots of limitations

Incapability in Creativity and Scientific Discovery

LLM only capture past human experiences



Questions about LLM

- What capabilities does LLM have?
- Why are they so powerful?
- How are they trained?
- How should we use and adapt LLM?
- How to address LLM limitations and mitigate their issues?

We will demystify these questions

What are we going to cover in the class?

Part 1: Basic about Language Models

- neural language modeling, dense word embeddings, tokenization, classical small language models

Week	Dates	Topics
1	Aug 22	Lectures introduction
2	Aug 25, 27, 29	LLMs capabilities and limitations n-gram language models word embeddings LLMs tokenization techniques
3	Sep 3, 5	Neural architecture for LLMs Language Models, RNN, LSTM

Part 2: Foundations in LLM

- Model Architectures of LLM

4	Sep 8, 10, 12	Attention mechanism, self-attention, multi-head attention Model architecture for transformers Transformer-based LLMs
5	Sep 15, 17, 19	Encoder LLM, Encoder-Decoder LLM, Decoder-only LLM Mixture of Experts architecture for LLMs

- Training and Inference of LLM

6	Sep 22, 24, 26	Training and Pre-training for LLMs Data used for training LLMs Inference of LLMs, including decoding strategies and controllable text generation

Part 3: How to use and adapt LLM

- Fine-tuning LLM and Post-training LLM

7	Sep 29, Oct 1, 3	Fine-tuning LLMs, including fine-tuning, prompt-based tuning, parameter efficient fine-tuning Post-training LLMs, including in-context learning, instruction tuning, model alignment

- Advanced Post-training – RL methods in LLM

8	Oct 6, 8, 10	Reinforcement Learning in LLMs, including Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), Group Relative Policy Optimization (GRPO) LLMs scaling law and emergent behavior

Part 4: Recent Advances in LLM

- Advanced Topics: retrieval, knowledge, reasoning, agent, multi-modal
- LLM limitations and mitigation: harm, bias, toxification, hallucination

9	Oct 13, 15	Retrieval-augmented LLMs, Retrieval Augmentation Generation
10	Oct 20, 22, 24	Knowledge-enhanced LLMs, including knowledge extraction, knowledge storage, knowledge editing, knowledge unlearning
11	Oct 27, 29, 31	Reasoning and Planning in LLMs LLM-based agent and multi-agent system LLM harm, debias, detoxification
12	Nov 3, 5, 7	LLM hallucination detection and mitigation Multi-modality LLM, language and vision foundation models

Grading Components

Part i. Quizzes (25%): There will be 6 quizzes in class. Each quiz will take 10 minutes in class time. The format will be multi-choice question answering. The quiz will be closed book. The quiz will evaluate the knowledge learned in classes. Tentative quiz dates are Sep 5, Sep 19, Oct 3, Oct 24, Nov 7, Nov 21.

Part ii. Mid-term Exam (35%): The mid-term exam will be held on **Oct 10th 2025** in class (11:45am – 12:35am at CSE E222). The exam will evaluate the knowledge covered in lectures before Oct 10. The exam will be closed book. No electricity device is allowed.

Grading Components

Part iii. Final Project (40%): Four students will form a team, and will need to complete a final project. Here are components of submissions:

(1) Project Proposal (10%, **due Oct 5**): The proposal should include problem statement and objective, proposed methods, planned experiments, plan for contributions of each team member

Grading Components

(2) In-class Presentation (15%): Each team will have 10-12 minutes for presentation and 3-5 minutes for Q&A.

- The presentation should include problem description, methods, experiments, result analysis, and conclusions.
- Each team should also **prepare one quiz-like question** (multi-choice format), that carries the most important takeaway message of the project (can be your core finding).
- In-class presentation dates: **Nov 7, 10, 12, 14, 17, 19, 21** (quiz)
- Enter the team members and presentation date (max 3 teams per date) on google sheet (will be shared after add/drop deadline)

Grading Components

(3) Final Report (15%, **due Dec 5**): The final report should be a formal paper-like report of the project, and should be concise and at most 9 pages (NeurIPS template https://www.overleaf.com/latex/templates/neurips-2024/tpsbbrdqcmsh).

- The final report should include abstraction, introduction, related work, method description, dataset processing, experiment settings, experiment results, result analysis, and conclusion.
- The report should also explain the contributions of each team member in the last section.
- The report can add additional experiments after presentation but should clearly state which part is added additionally.

How to choose Final Project?

Here are three choices for final projects:

- (1) choose the default project: fine-tuning LLMs for sentiment analysis

If you have limited experience with research, don't have any clear idea of what you want to do, you can choose the default project, but you are expected to try to **extend and improve it in various ways of your choice**, like contrastive learning, paraphrasing, regularized optimization etc.

You need to find sentiment analysis datasets, there are many available online, such as rotten tomatoes, yelp …

How to choose Final Project?

Here are three choices for final projects:

- (2) reproducing a paper related to Natural Language Processing and Large Language Models (https://aclanthology.org/ provides paper database)

Should clearly state the previous codebase and **highlight your contribution** in presentation and final report. **Only running the existing code will not get a good score.** You are encouraged to **extend and improve the original paper**, like improvements to the model, providing ablations or experimental studies that the original paper did not provide, or by running it on different datasets that illuminate novel questions etc.

How to choose Final Project?

Here are three choices for final projects:

- (3) any custom topics, potential topics can be LLMs reasoning, LLMs retrieval augmentation generation, LLMs hallucination detection or mitigation, LLMs for code generation, LLMs agent or multi-agent systems, LLMs for text summarization, LLMs for machine translation…

You are encouraged to choose custom project to explore the process of defining a problem, finding dataset, working out something you find interesting, and evaluating the system.

Again, you should clearly state the papers you cited and **highlight the difference or contribution** you made here.

Note about Final Project

- The project **topics** should be human-language related, or text-related tasks, or multi-modality (including text-modality) tasks
- A successful custom final project may develop a novel method or algorithm, but studying existing methods or applying them to new problems can also lead to interesting results without inventing something new.
- A successful custom final project may improve on some existing SOTA, but well-executed projects that perform thoughtful experiments and produce negative results, projects that just aim to build a fun system, or projects that answer an interesting research question without advancing SOTA are equally valuable.

Note about Final Project (continued)

- A successful final project should have clearly-defined **evaluation metrics** to measure the outcome of whatever experiment(s) you run; exact-match accuracy, precision, recall, and F1 score on either in-distribution or out-of-distribution test data are used to evaluate experiments in the default final project.
- A successful final project is encouraged to have a short, specific guiding question or hypothesis. An example of good guiding questions ‘Does fine-tuning some BERT layers work better than others, and is the best layer to fine-tune the same for all tasks?’ or ‘Does pre-training in one language still help if we fine-tune on a different language?’

Note about Final Project (continued)

- The models used for experiments are not restricted to language models with very large size. We do **allow small or medium size language models for experiments**. The key point is not the model size you use for experiments, but the research questions you find is valuable, the experiment design is reasonable and solid, your analysis is thoughtful, the evaluation is comprehensive.
- Students are expected to **not rely solely on calling LLMs API for their experiments**. Prompting LLMs is allowed, but you should also at least train or fine-tune language model (even small language models) to compare results.
- Every team is encouraged to submit a paper to ACL (discuss with me)

Resources for Final Project

- HiPerGator provides GPUs for this course, 8 GPU (L4 or B200), 32 CPU cores, and 2 Tb of Blue storage.
- **Important note from HiPerGator management:** if you use GPU for data visualization or train small-scale language model, **use L4 GPU**. You can use B200 **only** if your model size exceeds the capacity of L4, and you should expect job delays if you call B200.
- **Important!! Save the date: Sept 5** 11:45am in classroom, a tutorial for HiPerGator will be provided.
- Questions about project design or topic selection -> ask instructor; Questions about GPU usage or code difficulties: ask TA

Grading Scale

- A = [92, 100]
- A – = [89, 91.99]
- B + = [86, 88.99]
- B = [82, 85.99]
- B – = [79, 81.99]
- C + = [76, 78.99]
- C = [72, 75.99]
- C – = [69, 71.99]
- D + = [66, 68.99]
- D = [62, 65.99]
- D – = [59, 61.99]
- E = [0, 58,99]

Late Policies

For the submissions with due date, 25% is deducted for each late 24 hours (including weekends). For example, if the submission is due on Sunday at 11:59pm, then 25% will be deducted for submission on Monday, 50% will be deducted for submission on Tuesday, 75% will be deducted for submissions on Wednesday, and submissions will not be accepted after Wednesday 11:59pm.

Excused Absence

The Dean of Students Office (www.dso.ufl.edu) provides rules, guidance, and approval for excuse documentation. An instructor notification letter from the DSO must directly state that an “absence has been excused” and the letter must specify the dates of the entire duration of the assessment. If the letter does not meet these requirements, the student must provide the instructor with documentation that complies with UF’s official excused absence policy (https://catalog.ufl.edu/UGRD/academic-regulations/attendance-policies/#absencestext). Without this, the absence will not be considered excused, and accommodations cannot be provided.

Academic Integrity

UF students are bound by The Honor Pledge which states, “We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honor and integrity by abiding by the Honor Code

https://policy.ufl.edu/regulation/4-040/

https://sccr.dso.ufl.edu/process/student-conduct-code/


Thank you!
UF
Herbert Wertheim
College of Engineering
UNIVERSITY of FLORIDA
LEADING THE CHARGE, CHARGING AHEAD