

# Natural Language Processing

Introduction

Instructor: Yangqiu Song

# Logistics

- Instructor: Yangqiu Song
- Email: [yqsong@cse.ust.hk](mailto:yqsong@cse.ust.hk)
- Canvas (<https://canvas.ust.hk>)
  - Lecture notes
  - Assignments
  - Projects

# Goals of this Course

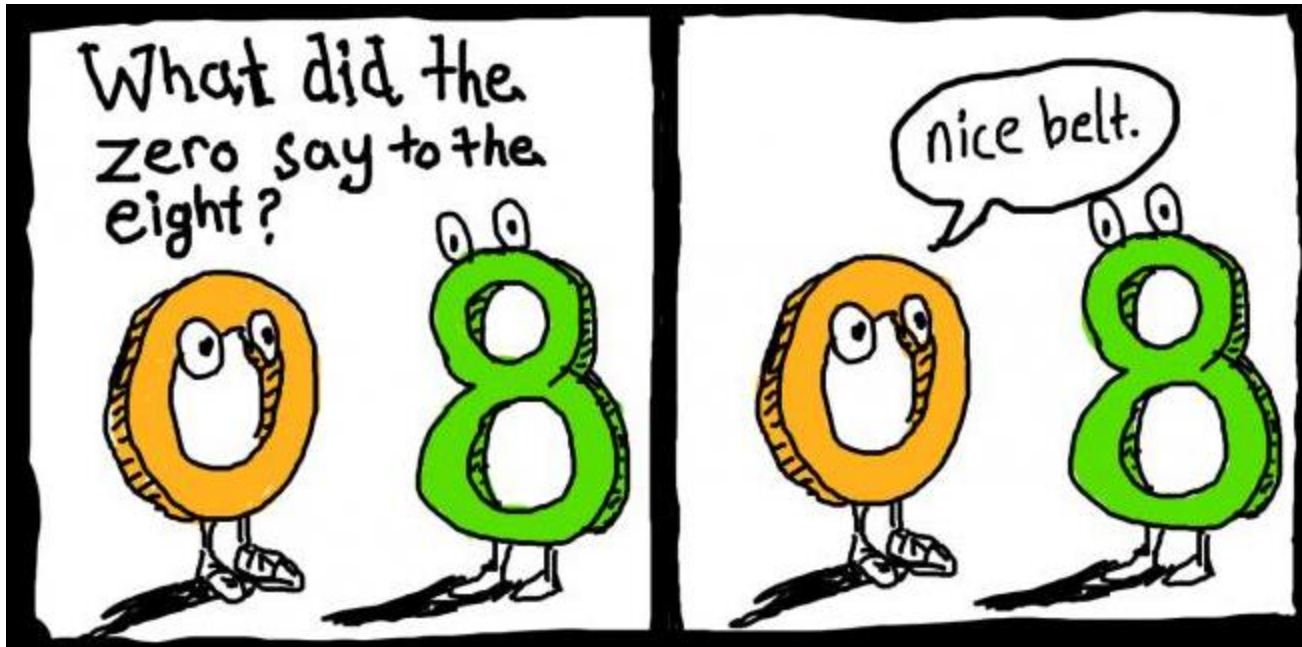
- Learn about the problems and possibilities of natural language analysis:
  - What are the major issues?
  - What are the major solutions?
    - How well do they work?
    - How do they work?
- At the end you should:
  - Agree that language is subtle and interesting!
  - Feel some ownership over the algorithms
  - Be able to assess NLP problems
    - Know which solutions to apply when, and how
  - Be able to read papers in the field

# Today

- Why NLP is difficult
- How to solve it (in general)?
- What we'll do in this course

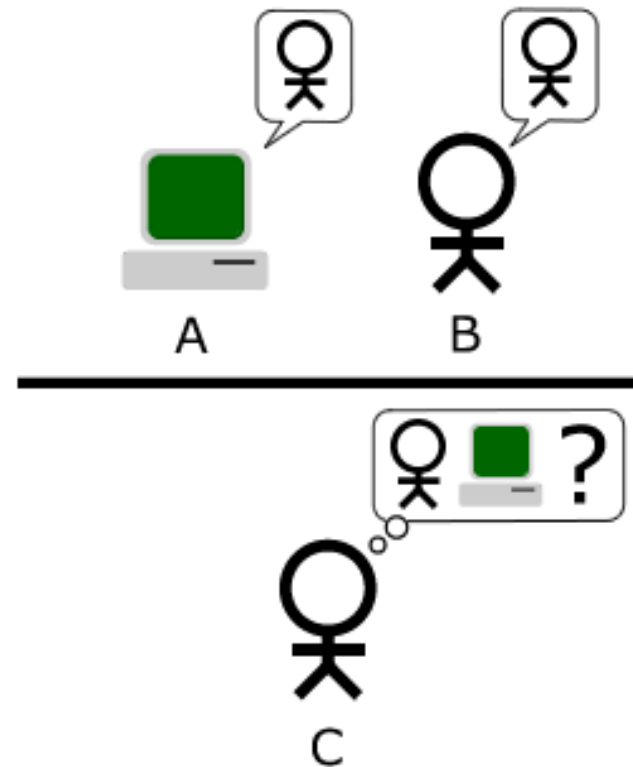
# Natural Language

- Understanding language is a very complex thing
- But something that humans are amazingly good at



# Artificial Intelligence: Turing Test

- Replacement of “Can machines think?”
  - Behavioral test
  - not just natural language understanding
- Natural language conversation
  - *Imitation game (1950)*
    - Human?
    - Computer?



# The AI Winter

- AI winter: 1974–80 and 1987–93
  - 1966: the failure of **machine translation**,
  - 1970: the abandonment of **connectionism**,
  - 1971–75: DARPA's frustration with the **Speech Understanding Research** program at Carnegie Mellon University,
  - 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill report,
  - 1973–74: DARPA's cutbacks to academic AI research in general,
  - 1987: the collapse of the Lisp machine market,
  - 1988: the cancellation of new spending on AI by the Strategic Computing Initiative,
  - 1993: expert systems slowly reaching the bottom, and
  - 1990s: the quiet disappearance of the fifth-generation computer project's original goals.

# Enabled by Big Data

- 1971–75: DARPA's frustration with the **Speech Understanding**
- “Watson is a **question answering** (QA) computing system that IBM built to apply advanced
  - natural language processing,
  - information retrieval,
  - knowledge representation,
  - automated reasoning, and
  - machine learning technologies
- to the field of **open domain question answering**.”



In 2011, Watson competed on Jeopardy! against former winners Brad Rutter and Ken Jennings. Watson received the first place prize of \$1 million.



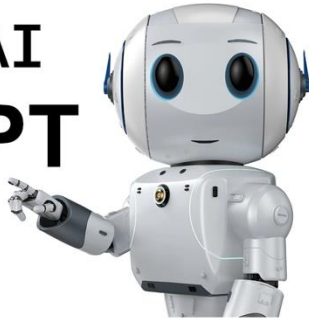


# Enabled by Big Data



# Enabled by **Big Data** and **Powerful Machines** and **Algorithms**

 OpenAI  
**ChatGPT**



MI

How are you today, ChatGPT?



As a language model trained by OpenAI, I don't have the ability to experience emotions like a human would. I am designed to assist with providing information and answering questions to the best of my ability. Is there something specific you would like to know?



MI

What is the Capital of the Klingon homeworld, Qo'noS?



In the fictional Star Trek universe, the Klingon homeworld is called Qo'noS and its capital is called First City. It is the political and cultural center of the Klingon Empire and the location of the Klingon High Council. It is a city of great size and importance in Klingon society.



# Large Language Models !

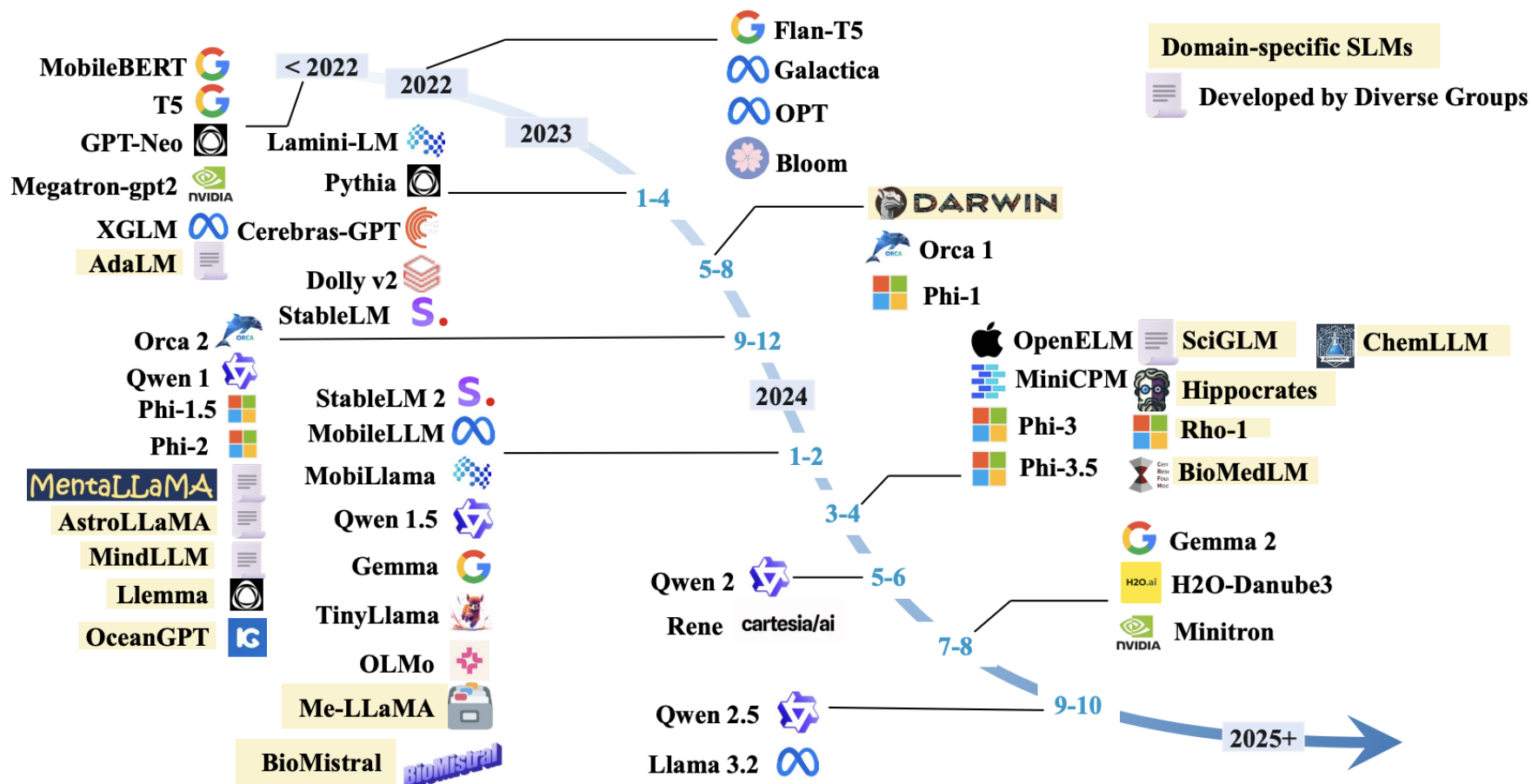
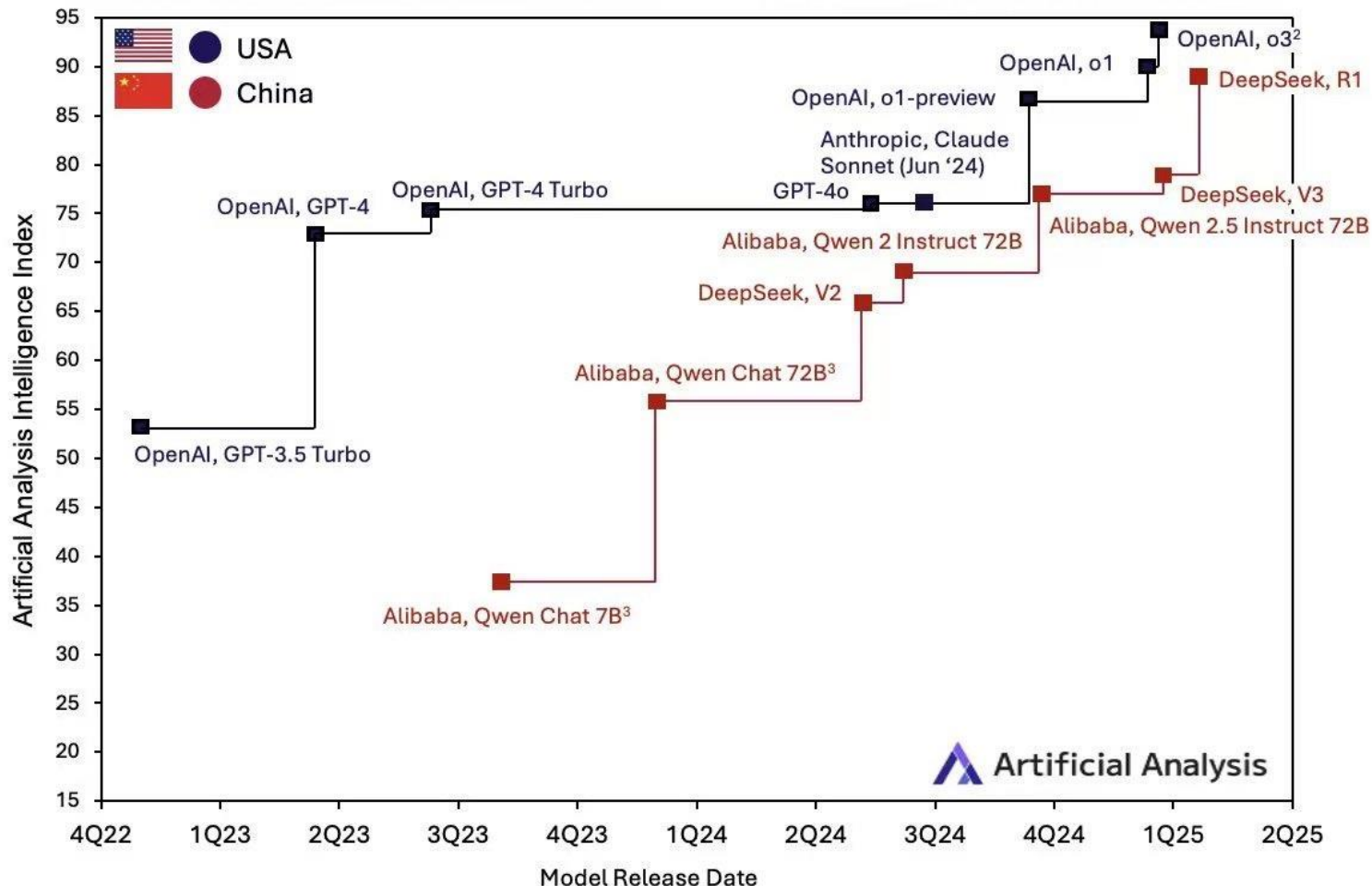


Fig. 3. A timeline of existing small language models.

# Large Language Models !

## US & China: Frontier Language Model Intelligence, Over Time<sup>1</sup>

Artificial Analysis Intelligence Index: MMLU, GPQA Diamond, MATH-500, HumanEval. Top 5 Labs by Market Share.



# Big Data

- Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, Jeffrey Dean. **Large Language Models in Machine Translation**. ACL, 2007
  - <http://www.aclweb.org/anthology/D07-1090.pdf>
  - 2 Trillion Tokens by Google
  - Using 1,500 CPUs, 1 day
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2018
  - <https://arxiv.org/abs/1810.04805>
  - 3.3 Billion Tokens by Google: 1,000 times less than 2007 but more powerful
  - Using 16 TPUs (64 Chips), 4 days
- ChatGPT: Optimizing Language Models for Dialogue. 2023
  - <https://openai.com/blog/chatgpt/>


GPT-3 training data

| Dataset      | # tokens    | Proportion within training |
|--------------|-------------|----------------------------|
| Common Crawl | 410 billion | 60%                        |
| WebText2     | 19 billion  | 22%                        |
| Books1       | 12 billion  | 8%                         |
| Books2       | 55 billion  | 8%                         |
| Wikipedia    | 3 billion   | 3%                         |

<https://en.wikipedia.org/wiki/GPT-3>

| Source                | Provenance   | New?    | Documents (millions) | OLMo tokens (billions) | Sample Proportion | Cutoff Date |
|-----------------------|--|---------|----------------------|------------------------|-------------------|-------------|
| Dolma's CC            | <a href="#">Common Crawl via Dolma v1.6</a>  | Updated | 875.2                | 1,195.50               | 50%               | Mar-23      |
| Refined Web           | <a href="#">Refined Web</a>  | Yes     | 664                  | 456.4                  | 100%              | Feb-23      |
| StarCoder             | <a href="#">StarCoder</a>  | Yes     | 206.6                | 263.8                  | 100%              | May-23      |
| C4                    | <a href="#">C4 via Dolma v1.6</a>  | Updated | 249.9                | 138.4                  | 50%               | Apr-19      |
| Reddit                | <a href="#">PushShift API</a>  | Updated | 377.4                | 79.9                   | 100%              | Mar-23      |
| Semantic Scholar      | <a href="#">peS2o via Dolma v1.6</a>   | No      | 38.8                 | 57.2                   | 100%              | Mar-23      |
| arXiv                 | <a href="#">RedPajama v1</a>   | Yes     | 1.5                  | 28                     | 100%              | Mar-23      |
| StackExchange         | <a href="#">RedPajama v1</a>   | Yes     | 29.3                 | 19.6                   | 100%              | Mar-23      |
| Flan                  | Flan Collection, reproduced following the original code, as performed by <a href="#">Dettmers et al., (2023)</a> | Yes     | 52.1                 | 16.5                   | 100%              | Feb-23      |
| CC News               | <a href="#">Common Crawl</a>   | Yes     | 22                   | 14.3                   | 100%              | Mar-23      |
| OpenWebMath           | <a href="#">OpenWebMath via Proof Pile II</a>  | Yes     | 2.9                  | 12.6                   | 100%              | May-23      |
| Algebraic Stack       | <a href="#">Proof Pile II</a>  | Yes     | 2.8                  | 12.6                   | 100%              | Oct-23      |
| Project Gutenberg     | <a href="#">Project Gutenberg via Dolma v1.6</a>   | No      | 0.0556               | 5.3                    | 100%              | Mar-23      |
| MegaWika              | <a href="#">MetaWika</a>   | Yes     | 3.2                  | 4.6                    | 100%              | Jul-23      |
| Wikipedia & Wikibooks | <a href="#">Wikimedia via Dolma v1.6</a>   | No      | 6.2                  | 3.7                    | 200%              | Mar-23      |
| Total                 |  |         | 2532                 | 2,308.50               | 1,715.10          | Oct-23      |

<https://allenai.org/dolma>

v1\_7  4/15/2024 4.5 TB Used to train OLMo-7B-v1.7. New sources, more quality filtering, fuzzy deduplication.



# The Scaling Law

- A empirical function of the **model parameters  $N$**  (excluding embeddings), the **size of the dataset  $D$** , and the **amount of compute  $C$**  used for training

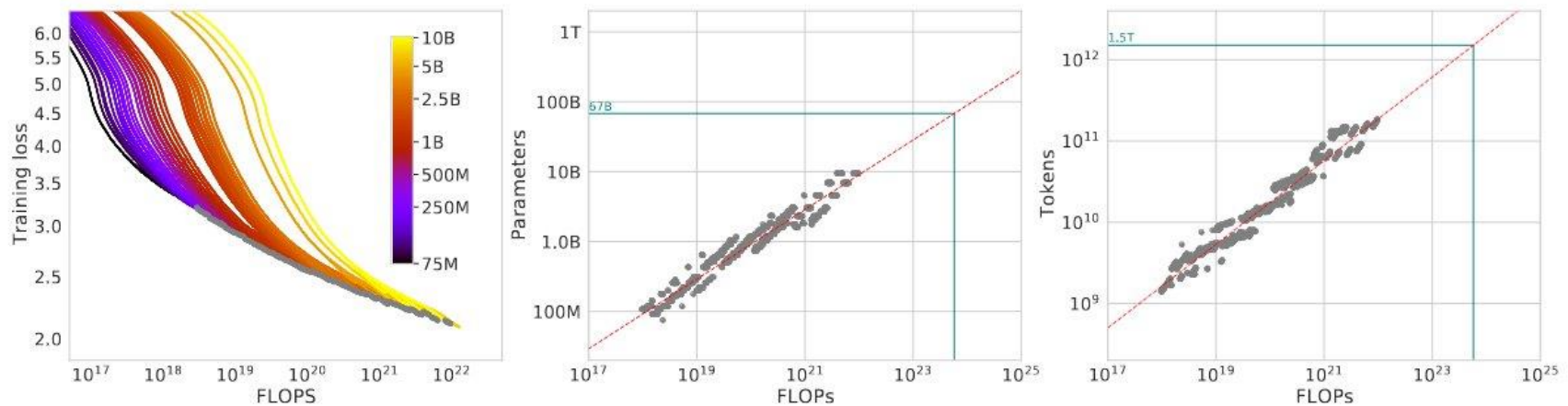


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate the optimal model size (**center**) for a given compute budget and the optimal number of training tokens (**right**). In green, we show projections of optimal model size and training token count based on the number of FLOPs used to train *Gopher* ( $5.76 \times 10^{23}$ ).

# The DeepSeek V3!

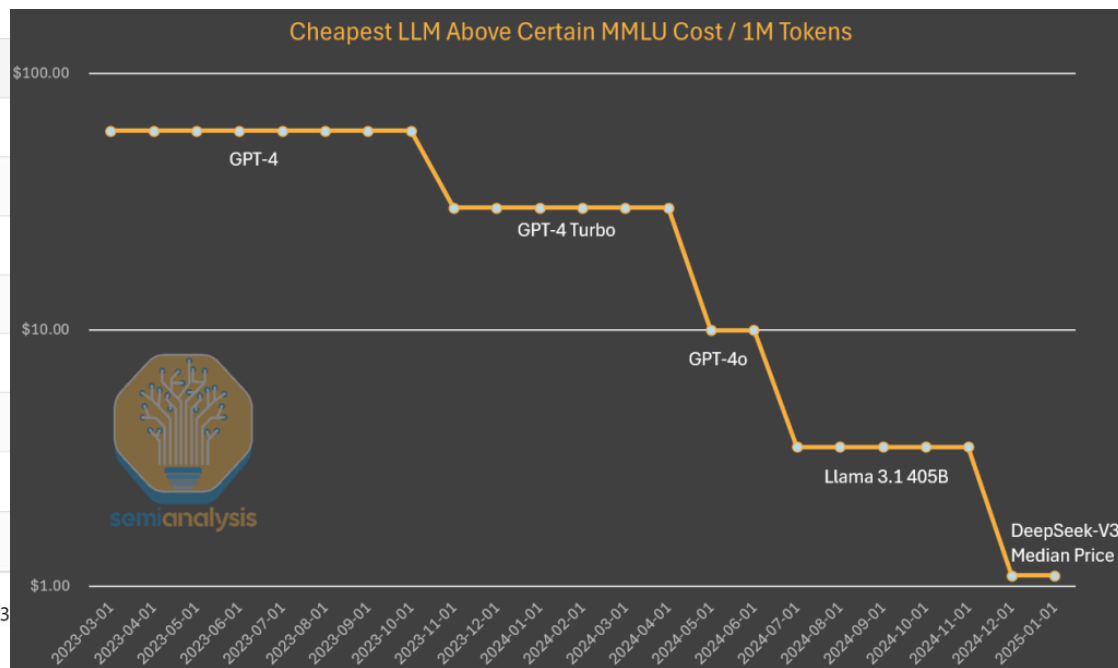
| Training Costs    | Pre-Training | Context Extension | Post-Training | Total    |
|-------------------|--------------|-------------------|---------------|----------|
| in H800 GPU Hours | 2664K        | 119K              | 5K            | 2788K    |
| in USD            | \$5.328M     | \$0.238M          | \$0.01M       | \$5.576M |

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.

## Metrics Comparison

| Metric                   | DeepSeek V3                       |
|--------------------------|-----------------------------------|
| Parameters               | 671B total (37B active per token) |
| GPU Type                 | NVIDIA H800                       |
| GPU Count                | 2,048                             |
| Training Duration        | ~2 months                         |
| Tokens Processed         | 14.8T                             |
| GPU Hours                | 2.788M                            |
| Training Cost            | ~\$5.6M                           |
| Cost per Trillion Tokens | ~\$378K                           |

**Note:** Cost estimations uses an average of **\$2/hour** for H800 GPUs (DeepSeek V3 based on rental GPU prices.

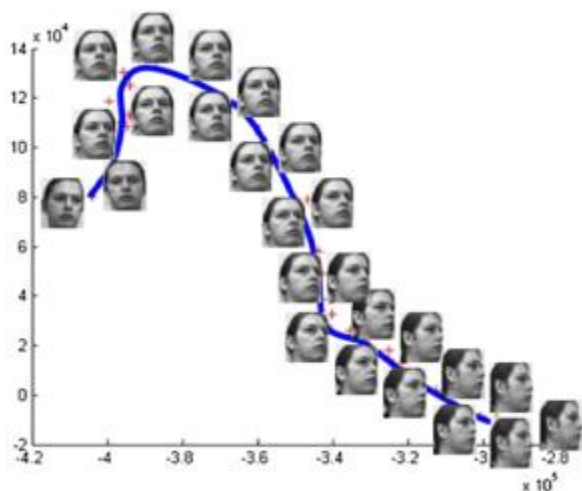


“DeepSeek has achieved this level of cost but by the end of the year do not be shocked if costs fall another 5x.”

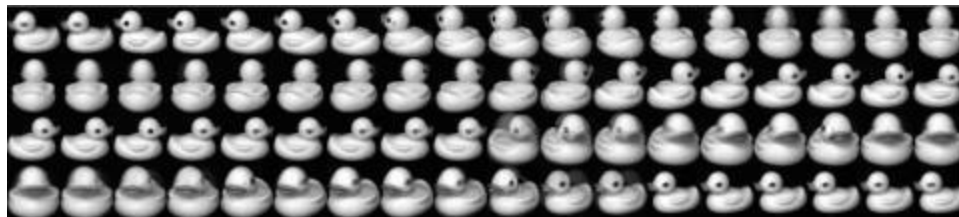
# What's Special about Human Language?

- A human language is a discrete/symbolic/categorical signaling system
  - With very minor exceptions for expressive signaling (“I loooove it.” “Whoomppaaa”)
- Large vocabulary, symbolic encoding of words creates a problem for machine learning – sparsity!

Face



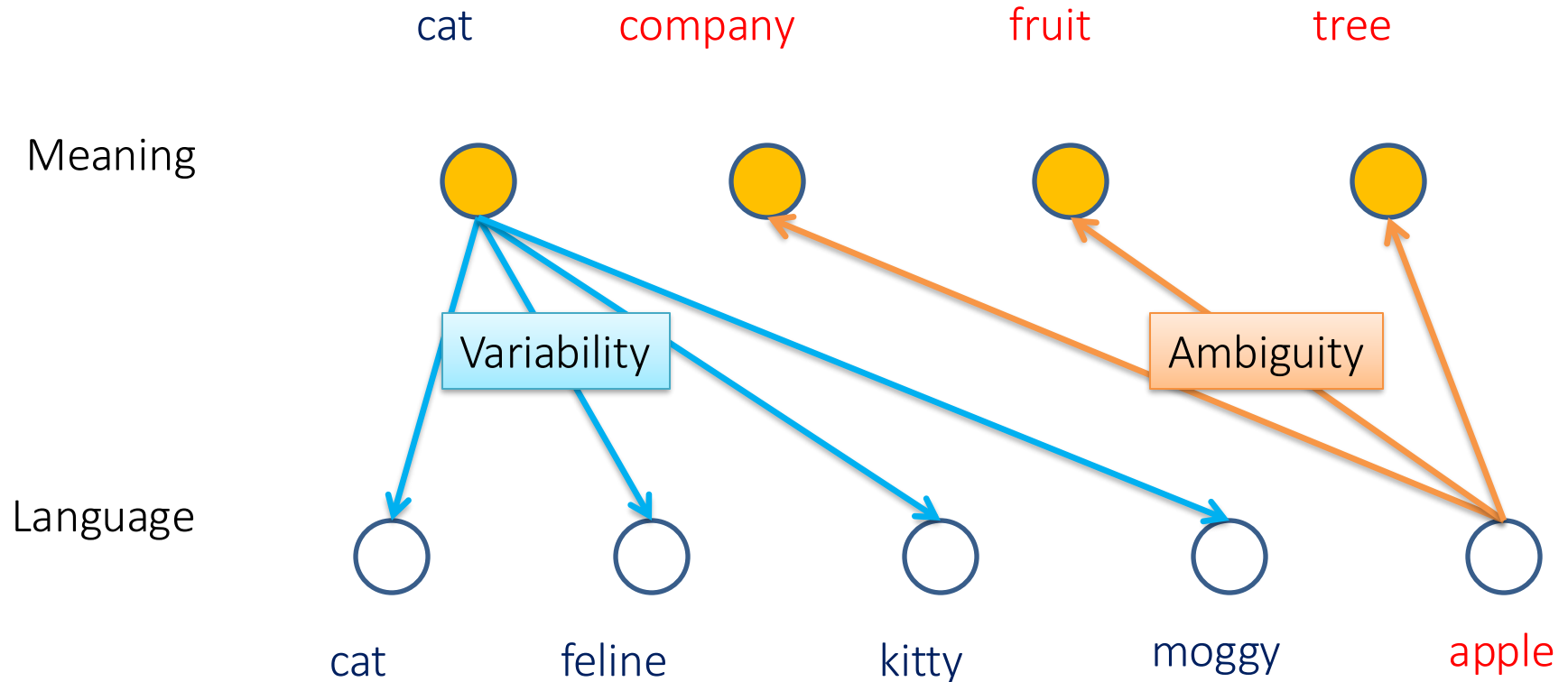
Toy





# Why is NLP Difficult?

- Variability and ambiguity everywhere



# Words are Ambiguous (have multiple meanings)

- I know that.
- I know that block.
- I know that blocks the sun.
- I know that block blocks the sun.

# More Examples of Ambiguity

- Get the cat with the gloves.



# Language Subtleties

- Adjective order and placement

- A big black dog
- A big black scary dog
- A big scary dog
- A scary big dog
- ✗ A black big dog

- Antonyms

- Which sizes go together?
  - Big and little
  - Big and small
  - Large and small
  - ✗ Large and little

# Levels of Linguistic Analysis: Analogy with Programming Languages

**Pragmatics:** what does it do?

← implemented the right algorithm

↑  
**Semantics:** what does it mean?

← no implementation bugs

↑  
**Syntax:** what is grammatical?

← no compiler errors

↑  
**Morphology:** basic unit of words

← naming your world

# Analogy with Programming Languages

- **Syntax**: no compiler errors
- **Semantics**: no implementation bugs
- **Pragmatics**: implemented the right algorithm
  
- Different **syntax**, same **semantics** (5):
  - $2 + 3 \leftrightarrow 3 + 2$
  
- Good **semantics**, bad **pragmatics**:
  - correct implementation of deep neural network
  - for estimating coin flip prob.

How to do natural language processing?

# The Role of Memorization

- Children learn words quickly
  - As many as 9 words/day
  - Often only need one exposure to associate meaning with word
    - Can make mistakes, e.g., overgeneralization
      - “I goed to the store.”
  - Exactly how they do this is still under study



# The Role of Memorization

- Dogs can do word association too!
  - Rico, a border collie in Germany
  - Knows the names of each of 100 toys
  - Can retrieve items called out to him with over 90% accuracy.
  - Can also learn and remember the names of unfamiliar toys after just one encounter, putting him on a par with a three-year-old child.



[https://en.wikipedia.org/wiki/Rico\\_\(dog\)](https://en.wikipedia.org/wiki/Rico_(dog))

# But there is too much to memorize!

establish

establishment

the church of England as the official state church.

disestablishment

antidisestablishment

antidisestablishmentarian

antidisestablishmentarianism

is a political philosophy that is opposed to the separation of church and state.

# Rules and Memorization

- Current thinking in psycholinguistics is that we use a combination of rules and memorization
  - However, this is very controversial
- Mechanism:
  - If there is an applicable rule, apply it
  - However, if there is a memorized version, that takes precedence. (Important for irregular words.)
    - Artists paint “still lifes”
      - Not “still lives”
    - Past tense of
      - think → thought
      - blink → blinked

# Representation of Meaning

- I know that block blocks the sun.
  - How do we represent the meanings of “block”?
  - How do we represent “I know”?
  - How does that differ from “I know that.”?
  - Who is “I”?
  - How do we indicate that we are talking about earth’s sun vs. some other planet’s sun?
  - When did this take place? What if I move the block? What if I move my viewpoint? How do we represent this?

# How to tackle these problems?

- The field was stuck for quite some time.
- A new approach started around 1990
  - Well, not really new, but the first time around, in the 50's, they didn't have the text, disk space, or GHz
- Main idea: combine memorizing and rules
- How to do it:
  - Get large text collections (corpora)
  - Compute statistics over the words in those collections
- Surprisingly effective
  - Even better now with the Web

# NLP ?= Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Computer representation of language is not directly observable.
- Early connections to information theory (1940s)
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

# NLP $\neq$ Linguistics

- NLP must contend with NL data as found in the world
- NLP  $\approx$  computational linguistics
- Linguistics has begun to use tools originating in NLP!

# Desiderata for NLP Methods

(ordered arbitrarily)

- Sensitivity to a wide range of the phenomena and constraints in human language
- Generality across different languages, genres, styles, and modalities
- Computational efficiency at construction time and runtime
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency, robustness, etc.)
- High accuracy when judged against expert annotations and/or task-specific performance



# Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Theory of computation
- Data science
- Social and political science
- Psychology
- Economics
- Education

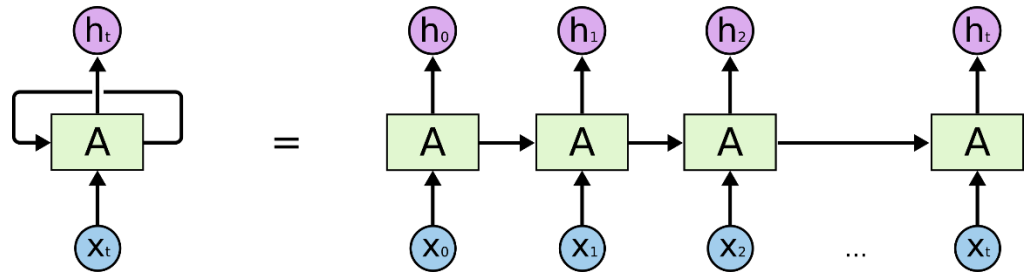
# Machine learning for NLP

- mid-1970s: **HMMs** for speech recognition → probabilistic models
- early 2000s: **conditional random fields** for part-of-speech tagging → structured prediction
- early 2000s: **latent Dirichlet allocation** for modeling text documents → topic modeling
- mid 2010s: **sequence-to-sequence** models for machine translation → neural networks with memory/state
- 2020: **large language models** → in-context learning; instruction tuning; generation-based AI

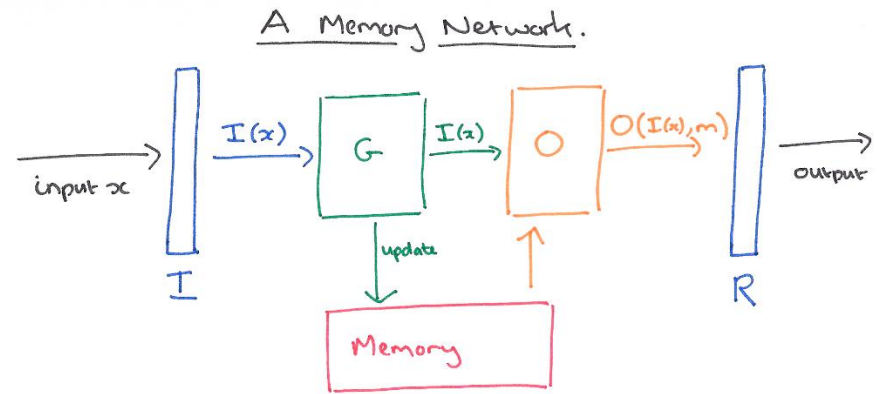
We will select some of the important topics

# Nowadays: Deep learning for NLP

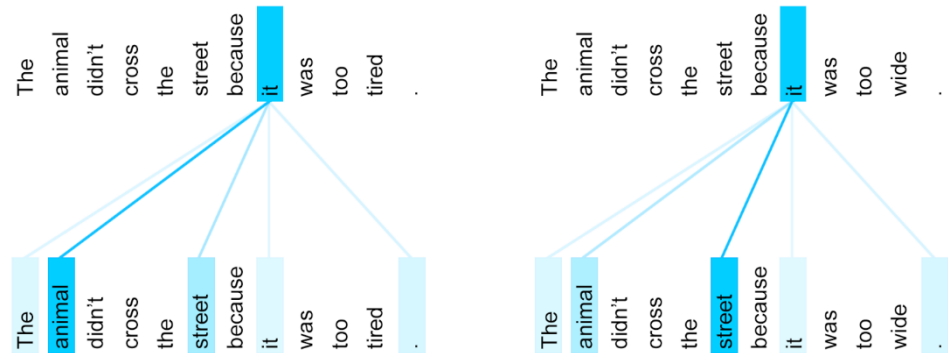
- Sequence models



- External memory models



- Attention models



# Real-World Applications of NLP

- Spelling suggestions/corrections
- Grammar checking
- Information extraction
- Text categorization
- Automated customer service
  - Conversational agents
  - Question answering
- Speech recognition (limited)
- Machine translation
- Social media analysis
- Rich visual understanding
- Mining legal, medical, or scholarly literature

# What We'll Do in this Course

- Learn fundamental machine learning models for NLP
  - Classification models
  - Language models
  - Sequence labeling models
  - Advance NLP tasks using deep learning
- Use NLTK (Natural Language ToolKit) and PyTorch to try out various algorithms
  - Some assignment will be to do some exercises

# What We'll Do in this Course

- Adopt a large text collection
- Use a wide range of NLP techniques to process it
- Release the results for others to use

# How to analyze a big collection?

- Your ideas go here with a project
- Some free computational resources to use
  - <https://www.kaggle.com/docs/notebooks>
  - <https://www.kaggle.com/dansbecker/running-kaggle-kernels-with-a-gpu>
  - SuperPOD
    - Our TAs are experienced users

# Course Information

- Work load and grading:
  - Assignments (10%)
    - Reading notes
  - Group Projects and Presentation (50%)
    - Project (40%): a team-based project
      - Up to 3 students in each group
      - Due: in the 11<sup>th</sup> week
    - In-class Presentation (10%): present what you have done in the project/survey
  - Individual Project (40%)
    - Similar to this project, with some modifications
    - Evaluating Large Language Models
      - [https://stanford-cs324.github.io/winter2022/projects/CS324\\_P1.pdf](https://stanford-cs324.github.io/winter2022/projects/CS324_P1.pdf)
    - Due: in the 14<sup>th</sup> week
  - Academic integrity policy
    - Late submission: score got reduced (time based)
    - Plagiarism: all involved parties will get zero



15 minutes break