

# CSCE 638 Natural Language Processing Foundation and Techniques

Spring 2025

Lecture 1: Course Overview

Kuan-Hao Huang



# Instructor

- [Kuan-Hao Huang](#)
- Assistant Professor
  - Department of Computer Science and Engineering
- Research focus: Natural Language Processing (NLP)
  - Reliability, Privacy, and Fairness in NLP models
  - Large Language Models (LLMs)
  - Multilingual NLP
  - Multimodal Understanding
  - NLP applications



# Course Information

- Time: Monday/Wednesday 4:10pm – 5:25pm
- Location: EABB 106
- Course Information:
  - Canvas: <https://canvas.tamu.edu/courses/356764>
  - Course Website: <https://khhuang.me/CSCE638-S25/>

# Course Staff

## Instructor



Kuan-Hao Huang

- Email: [khhuang@tamu.edu](mailto:khhuang@tamu.edu)
- Office Hour: Wed. 2pm – 3pm
- Office: PETR 219

## TA



Rahul Baid

- Email: [rahulbaid@tamu.edu](mailto:rahulbaid@tamu.edu)
- Office Hour: Wed. 12pm – 1pm
- Office: TBD

For questions, send emails to [csce638-ta-25s@lists.tamu.edu](mailto:csce638-ta-25s@lists.tamu.edu)

# Lecture Plan

- Course introduction
  - Prerequisite knowledge
  - Topics
- Course logistics
  - Assignments
  - Quizzes
  - Course Project
  - Grading

# Prerequisite Knowledge

- Linear Algebra
  - Vectors, matrices, matrix operations, and singular value decomposition (SVD)
- Calculus
  - Differentiation and gradients, partial derivatives, and chain rule
- Probability
  - Basic probability theory (conditional probability, Bayes' theorem)
  - Concepts of maximum likelihood estimation (MLE)
- Machine Learning
  - Supervised learning, training framework, validation and overfitting

# What is Natural Language Processing (NLP)?

- One field of AI that focuses on the interaction between machines and human languages
- Enable machines to **understand**, **reason**, and **respond** to human languages



(Generated by ChatGPT)

# Large Language Models

ChatGPT Share EJ

Bob has 5 tennis balls. He buys 2 more cans of balls. Each can has 3 balls. How many balls does he have now?

Bob starts with 5 tennis balls. Each can contains 3 balls, so 2 cans provide:

$$2 \times 3 = 6 \text{ balls}$$

Adding these to the original 5 balls:

$$5 + 6 = 11 \text{ balls}$$

Bob now has 11 tennis balls.

Generate response

Understand user instruction

Reason and think






# NLP is Everywhere

The image shows a Google search interface with the search term "texas a&m". The search results are displayed in a list format. The first result is for "Texas A&M" with the URL "https://www.tamu.edu". The second result is for "Texas A&M Athletics" with the URL "https://12thman.com" and includes a small image of a person in a red hat. The third result is for "Texas A&M University-Corpus Christi" with the URL "https://www.tamucc.edu". The fourth result is for "Texas A&M Athletics" with the URL "https://12thman.com" and includes a link to "2024 Football Schedule".


Google

texas a&m


All News Images Maps Videos Shopping Forums More Tools

 Texas A&M  
https://www.tamu.edu


**Texas A&M University**  
Howdy from Texas A&M University. Texas A&M University is an engine of imagination, learning, discovery and innovation. Here, you'll learn essential career ...

 Texas A&M Athletics  
https://12thman.com


**Texas A&M Athletics - 12thMan.com**  
The official athletics website for the Texas A&M Aggies.  
[Football](#) · [Staff Directory](#) · [2024 Football Schedule](#) · [Composite Calendar](#)

 Texas A&M University-Corpus Christi  
https://www.tamucc.edu

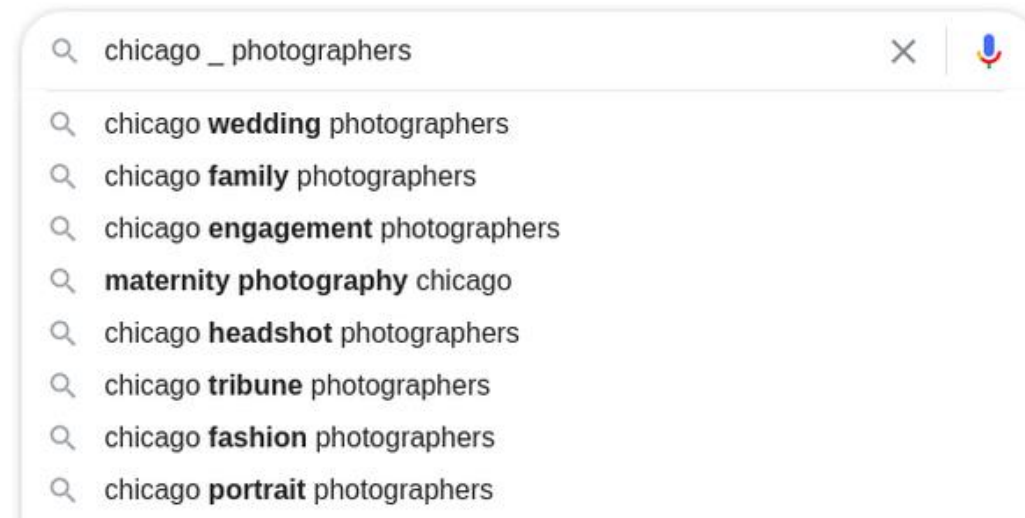
**Texas A&M University-Corpus Christi: Welcome Home**  
Welcome to THE ISLAND! Discover the Island University, the only university in the nation located on its own island, at the heart of the Texas Gulf Coast.

 Texas A&M Athletics  
https://12thman.com › sports › football › schedule

**2024 Football Schedule**  
2024 Football Schedule · Early: Game will have a start time between 11AM-Noon CT · Afternoon: Game will have a start time between 2:30PM – 3:30PM CT · Night: ...



# NLP is Everywhere



# NLP is Everywhere

## Customer reviews

★★★★☆ 4.6 out of 5

10,134 global ratings



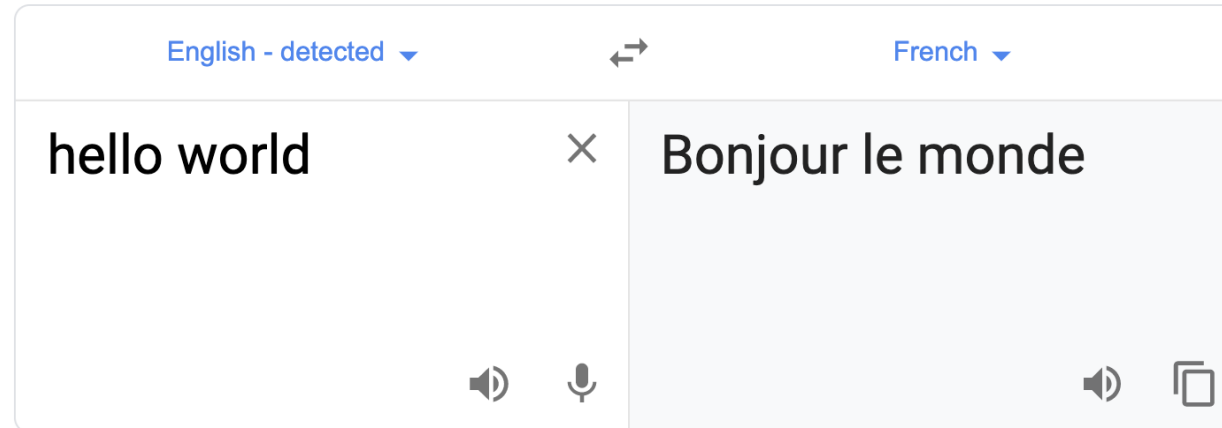
## Customers say

Customers like the sound quality, quality, and ease of installation of the sound and recording equipment. They mention that it does the job quite well as a pop filter and is good value for money. Customers are also satisfied with the sound clarity, quality and ease to installation. However, some customers are mixed on stability, fit, and flexibility.

AI-generated from the text of customer reviews

- ✓ Quality
- ✓ Value
- ✓ Sound quality
- ✓ Ease of installation
- ✓ Filter
- ✓ Fit
- Stability
- Flexibility

# NLP is Everywhere



# NLP is Everywhere

### Your recently viewed items and featured recommendations

Sponsored products related to this search [What's this?](#)

							
All-new Echo Show (2nd Gen) + Ring Video Doorbell 2- Charcoal 1 offer from <b>\$428.99</b>	AmazonBasics Microwave, Small, 0.7 Cu. Ft, 700W, Works with Alexa ★★★★☆ 1,375 <b>\$59.99</b> ✓prime	Echo Look   Hands-Free Camera and Style Assistant with Alexa— includes Style Check to... ★★★★☆ 413 <b>\$99.99</b> ✓prime	Sonos Beam - Smart TV Sound Bar with Amazon Alexa Built-in - Black ★★★★☆ 474 <b>\$399.00</b> ✓prime	Echo Wall Clock - see timers at a glance - requires compatible Echo device ★★★★☆ 1,231 <b>\$29.99</b> ✓prime	Echo Spot Adjustable Stand - Black ★★★★☆ 933 <b>\$19.99</b> ✓prime	AHASTYLE Wall Mount Hanger Holder ABS for New Dot 3rd Generation Smart Home Speakers... ★★★★☆ 12 <b>\$10.99</b> ✓prime	Angel Statue Crafted Stand Holder for Amazon Echo Dot 3rd Generation, Alexa Smart... ★★★★☆ 57 <b>\$25.99</b> ✓prime

Page 1 of 3

### Explore more from across the store

							
Actionable Gamification: Beyond Points, Badges, and Incentives › Yu-kai Chou	The Model Thinker: What You Need to Know to... › Scott E. Page	Don't Make Me Think, Revisited: A Common... › Steve Krug	Hooked: How to Build Habit-Forming Products › Nir Eyal	Microservices Patterns: With examples in Java › Chris Richardson	Solving Product Design Exercises: Questions &... › Artiom Dashinsky	100 Things Every Designer Needs to Know About... Susan Weinschenk	Infinity › Jonathan Hickman ★★★★☆ 182

Page 1 of 6

# NLP is Everywhere

**Lufthansa Customer Service Chatbot**

Thank you. Below you can find a selection of topics I can help you with.

These topics are currently very popular:

- Check my booking
- What are my rebooking and refund options?
- I want to rebook
- I want to get a refund
- COVID-19 testing requirements
- Affected by recent flight schedule update
- Go to the main menu

Powered by COGNIFY AI

Write reply

# NLP is Everywhere



# NLP is Everywhere

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

## 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



### Summary

The document titled "Attention Is All You Need" introduces the Transformer model, a network architecture based solely on attention mechanisms, eliminating the need for recurrent or convolutional neural networks in sequence transduction tasks. The Transformer model achieves superior performance in machine translation tasks, demonstrating improved quality, parallelizability, and reduced training time compared to existing models. The key points and arguments presented in the document are as follows:

- The dominant sequence transduction models rely on complex recurrent or convolutional neural networks with an encoder-decoder structure and attention mechanisms.
- The Transformer model proposes a new architecture based solely on attention mechanisms, eliminating the need for recurrence and convolutions.
- Experiments show that the Transformer model outperforms existing models in machine translation tasks, achieving state-of-the-art results with reduced training time.
- The model utilizes self-attention to compute representations of input and output sequences, allowing for more parallelization and global dependencies.
- The Transformer model consists of stacked self-attention and fully connected layers for both the encoder and decoder, enabling efficient sequence transduction.
- Multi-Head Attention is employed to jointly attend to information from different representation subspaces at different positions, enhancing the model's performance.

### Key Points:

- Transformer model introduces a network architecture based solely on attention



# Course Design

- Week 1 to Week 8
  - NLP fundamentals

W1	1/13	L1	Course Overview
	1/15	L2	Text Classification
W2	1/20		Martin Luther King, Jr. Day (No Class)
	1/22	L3	Word Representations
W3	1/27	L4	Tokenization, Language Modeling
	1/29	L5	Convolutional Neural Network, Recurrent Neural Networks
W4	2/3	L6	Sequential Tagging, Generation
	2/5	L7	Attention, Transformers
W5	2/10	L8	Transformers, Positional Encoding
	2/12	L9	Contextualized Representations, Pre-Training
W6	2/17	L10	Parameter-Efficient Fine-Tuning
	2/19	L11	Large Language Models, In-Context Learning
W7	2/24	L12	Reasoning, Instruction Tuning
	2/26	L13	Human Preference Alignment
W8	3/3	L14	Text Similarity, Retrieval-Augmented Generation
	3/5		Project Highlight

# Text Classification

SmartBuyGuy

Reviewed in the United States on October 10, 2024  
Size: 3 Count (Pack of 1) | **Verified Purchase**

I recently switched to the Amazon Basics Replacement Water Filters for my Brita pitcher, and the difference has been astonishing. Initially, I was apprehensive about using a generic brand, but I can confidently say these filters deliver outstanding performance comparable to the leading brands.

The first thing I noticed was the taste of my water. The multi-stage filtration technology effectively removes contaminants, leaving my water crisp and fresh. I used to taste chlorine in my tap water, but that's now a distant memory. It's a pleasure to drink water again!

Installation was seamless. The filters fit perfectly into my Brita pitcher, and I had no issues setting them up. I appreciate the clear instructions that come with the product, making the process hassle-free. Additionally, each filter lasts up to 40 gallons or about two months, making them a cost-effective choice for my household.

I also love the eco-friendly aspect of these filters. Knowing that one filter replaces 300 single-use plastic bottles gives me a sense of satisfaction. Not only am I saving money, but I'm also contributing to reducing plastic waste—something we all need to consider in today's world.

Positive / Negative

## A.I. Chatbots Defeated Doctors at Diagnosing Illness

A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.

By **Gina Kolata**

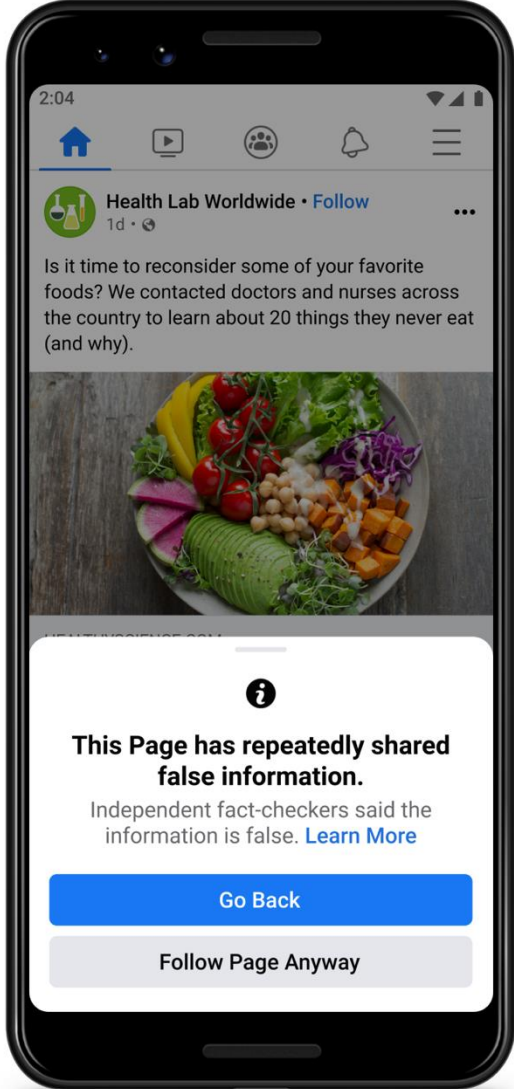
Nov. 17, 2024  
[Leer en español](#)

Dr. Adam Rodman, an expert in internal medicine at Beth Israel Deaconess Medical Center in Boston, confidently expected that chatbots built to use artificial intelligence would help doctors diagnose illnesses.

He was wrong.

Instead, in a [study](#) Dr. Rodman helped design, doctors who were given ChatGPT-4 along with conventional resources did only slightly better than doctors who did not have access to the bot. And, to the researchers' surprise, ChatGPT alone outperformed the doctors.

Technology / Business / Economy /  
Politics / Education / Sports



Suspicious / Normal

<https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation/>  
<https://www.nytimes.com/2024/11/17/health/chatgpt-ai-doctors-diagnosis.html>  
<https://www.amazon.com/dp/B07YT16TMS>

# Token Classification



In fact, the **Chinese** **NORP** market has the **three** **CARDINAL** most influential names of the retail and tech space – **Alibaba** **GPE**, **Baidu** **ORG**, and **Tencent** **PERSON** (collectively touted as **BAT** **ORG**), and is betting big in the global **AI** **GPE** in retail industry space. The **three** **CARDINAL** giants which are claimed to have a cut-throat competition with the **U.S.** **GPE** (in terms of resources and capital) are positioning themselves to become the 'future **AI** **PERSON** platforms'. The trio is also expanding in other **Asian** **NORP** countries and investing heavily in the **U.S.** **GPE** based **AI** **GPE** startups to leverage the power of **AI** **GPE**. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing **one** **CARDINAL**, with an anticipated **CAGR** **PERSON** of **45%** **PERCENT** over **2018 - 2024** **DATE**.

To further elaborate on the geographical trends, **North America** **LOC** has procured **more than 50%** **PERCENT** of the global share in **2017** **DATE** and has been leading the regional landscape of **AI** **GPE** in the retail market. The **U.S.** **GPE** has a significant credit in the regional trends with **over 65%** **PERCENT** of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as **Google** **ORG**, **IBM** **ORG**, and **Microsoft** **ORG**.

## Named Entity Recognition

## Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The **American Football Conference** (AFC) champion **Denver Broncos** defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, **2016**, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

**Question:** Which NFL team won Super Bowl 50?

**Answer:** **Denver Broncos**

**Question:** What does AFC stand for?

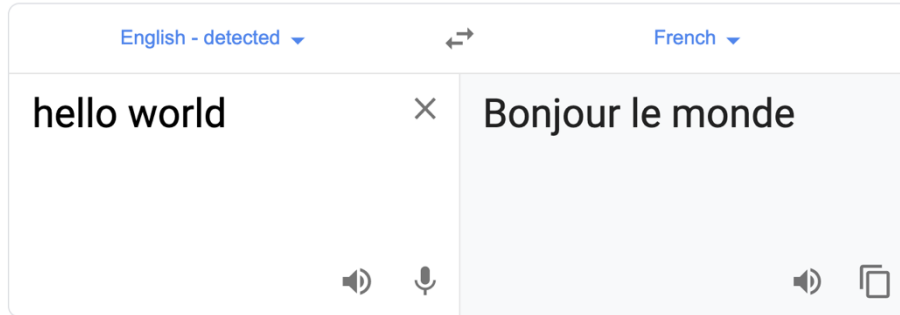
**Answer:** **American Football Conference**

**Question:** What year was Super Bowl 50?

**Answer:** **2016**

## Extractive Question Answering

# Text-to-Text Generation



I think I have an idea that should sort of improve campaign performance.

Tone Suggestion

**Confident**

I have an idea that should improve campaign performance.

Rephrase Dismiss

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

## 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\*<sup>1</sup>  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser\*  
Google Brain  
lukasz@kaiser@google.com

Illia Polosukhin\*<sup>1</sup>  
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.

<sup>1</sup>Work performed while at Google Brain.



### Summary

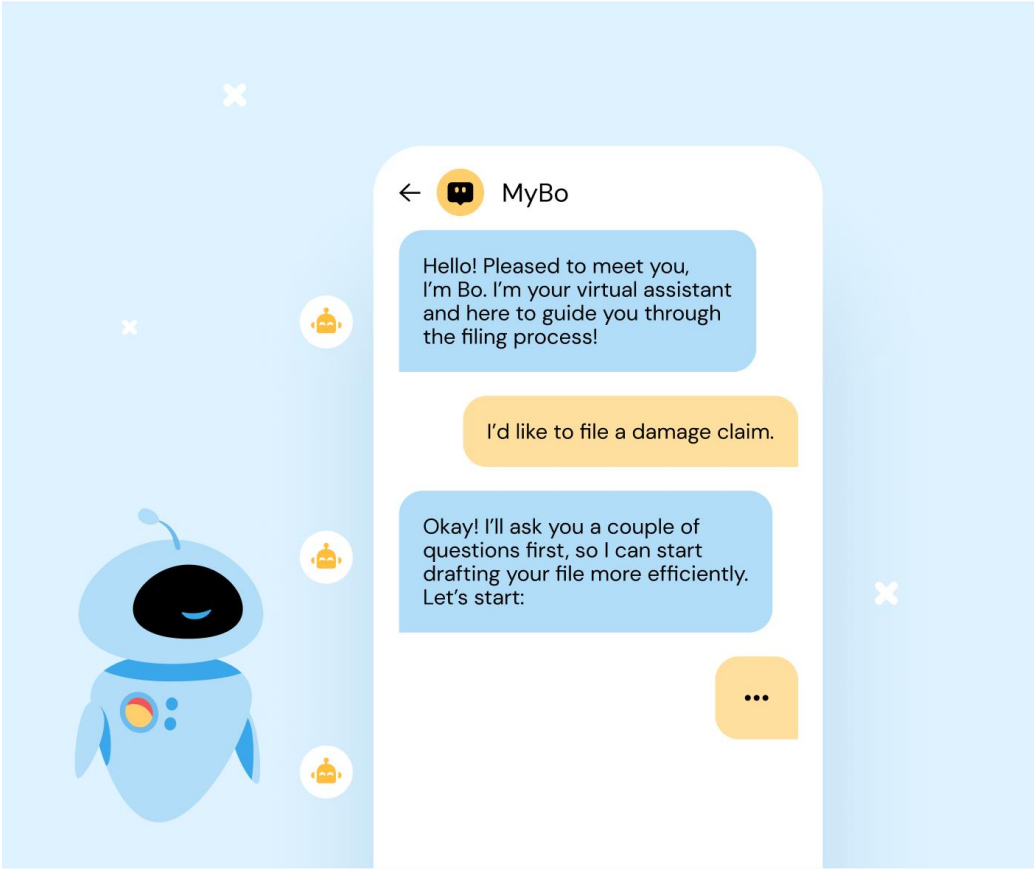
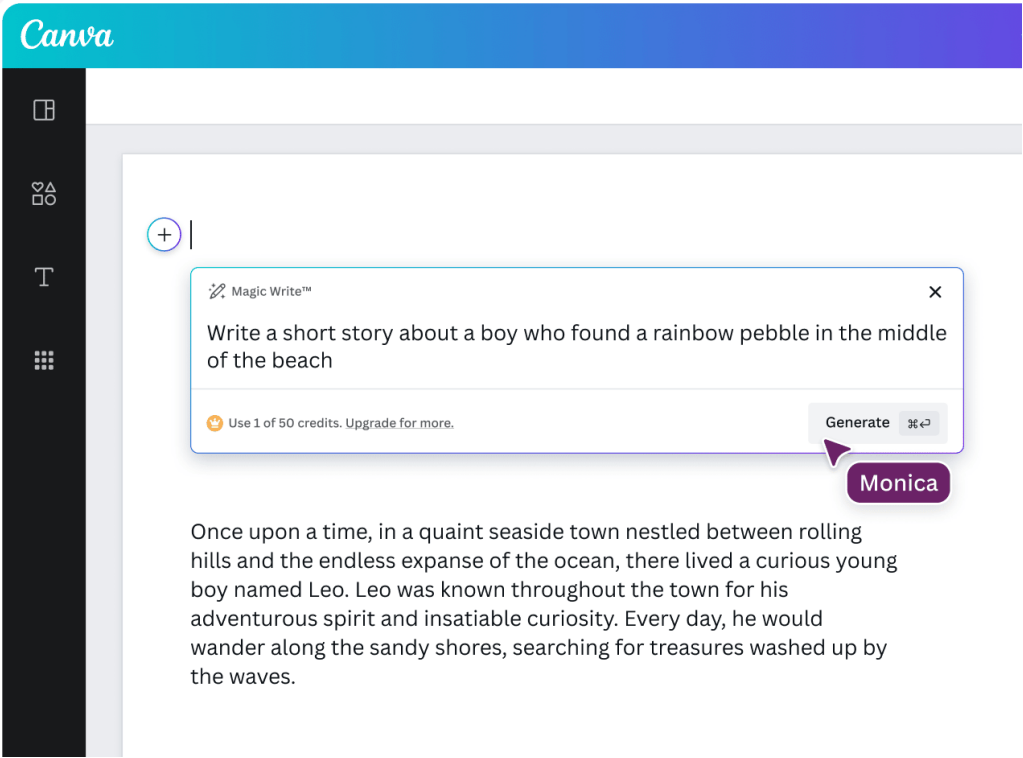
The document titled "Attention Is All You Need" introduces the Transformer model, a network architecture based solely on attention mechanisms, eliminating the need for recurrent or convolutional neural networks in sequence transduction tasks. The Transformer model achieves superior performance in machine translation tasks, demonstrating improved quality, parallelizability, and reduced training time compared to existing models. The key points and arguments presented in the document are as follows:

- The dominant sequence transduction models rely on complex recurrent or convolutional neural networks with an encoder-decoder structure and attention mechanisms.
- The Transformer model proposes a new architecture based solely on attention mechanisms, eliminating the need for recurrence and convolutions.
- Experiments show that the Transformer model outperforms existing models in machine translation tasks, achieving state-of-the-art results with reduced training time.
- The model utilizes self-attention to compute representations of input and output sequences, allowing for more parallelization and global dependencies.
- The Transformer model consists of stacked self-attention and fully connected layers for both the encoder and decoder, enabling efficient sequence transduction.
- Multi-Head Attention is employed to jointly attend to information from different representation subspaces at different positions, enhancing the model's performance.

### Key Points:

- Transformer model introduces a network architecture based solely on attention

# Open Text Generation



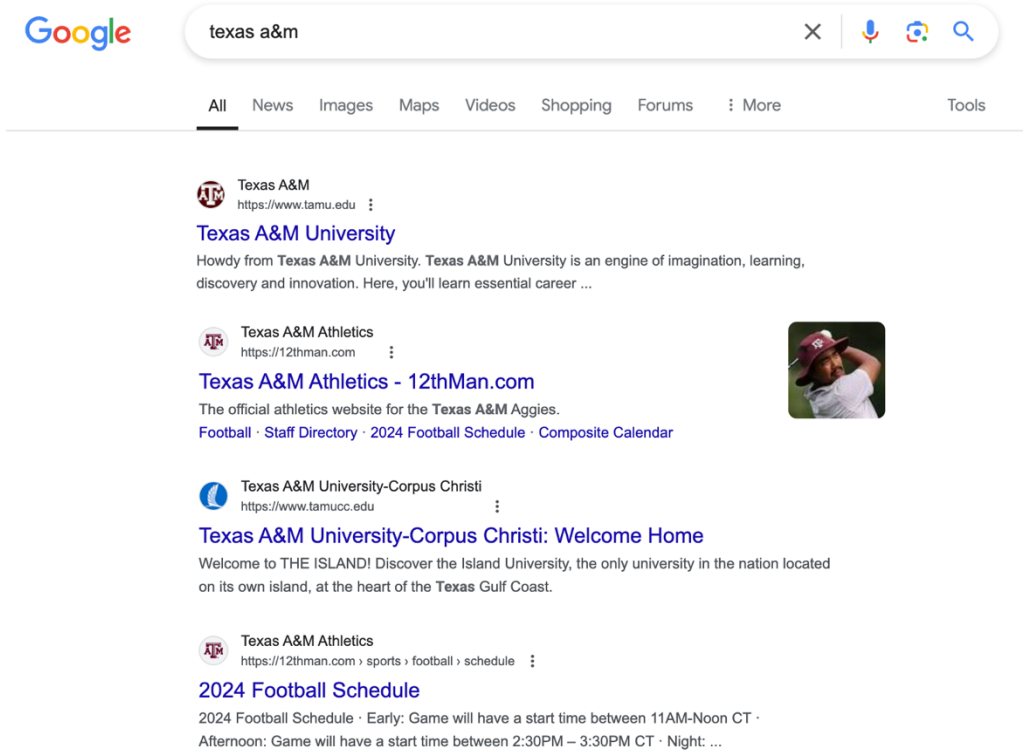
# Text Similarity and Retrieval

*We will go hiking if tomorrow is a sunny day.  
If it is sunny tomorrow, we will go hiking.*

→  *semantically similar*

*We will go hiking if tomorrow is a sunny day.  
We will go swimming if tomorrow is a sunny day.*

→  *semantically different*



Google search results for "texas a&m". The search bar shows "texas a&m" with a search icon. Below the search bar are tabs for "All", "News", "Images", "Maps", "Videos", "Shopping", "Forums", and "More". The results list several entries:

- Texas A&M** (<https://www.tamu.edu>): Texas A&M University. Howdy from Texas A&M University. Texas A&M University is an engine of imagination, learning, discovery and innovation. Here, you'll learn essential career ...
- Texas A&M Athletics** (<https://12thman.com>): Texas A&M Athletics - 12thMan.com. The official athletics website for the Texas A&M Aggies. Football · Staff Directory · 2024 Football Schedule · Composite Calendar
- Texas A&M University-Corpus Christi** (<https://www.tamucc.edu>): Texas A&M University-Corpus Christi: Welcome Home. Welcome to THE ISLAND! Discover the Island University, the only university in the nation located on its own island, at the heart of the Texas Gulf Coast.
- Texas A&M Athletics** (<https://12thman.com>): 2024 Football Schedule. 2024 Football Schedule · Early: Game will have a start time between 11AM-Noon CT · Afternoon: Game will have a start time between 2:30PM – 3:30PM CT · Night: ...

## Your recently viewed items and featured recommendations

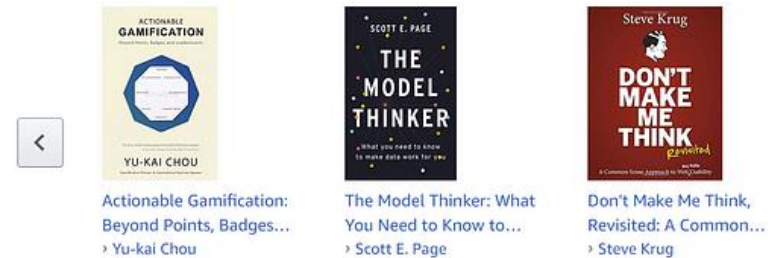
Sponsored products related to this search [What's this?](#)



Amazon sponsored products related to the search:

- All-new Echo Show (2nd Gen) + Ring Video Doorbell 2- Charcoal. 1 offer from \$428.99
- AmazonBasics Microwave, Small, 0.7 Cu. Ft, 700W, Works with Alexa. 4.5 stars, 1,375 reviews. \$59.99 ✓ prime
- Echo Look | Hands-Free Camera and Style Assistant with Alexa—includes Style Check to... 4.5 stars, 413 reviews. \$99.99 ✓ prime

## Explore more from across the store



Amazon recommended books:

- Actionable Gamification: Beyond Points, Badges... by Yu-kai Chou
- The Model Thinker: What You Need to Know to Make Data Work for You by Scott E. Page
- Don't Make Me Think, Revisited: A Common... by Steve Krug

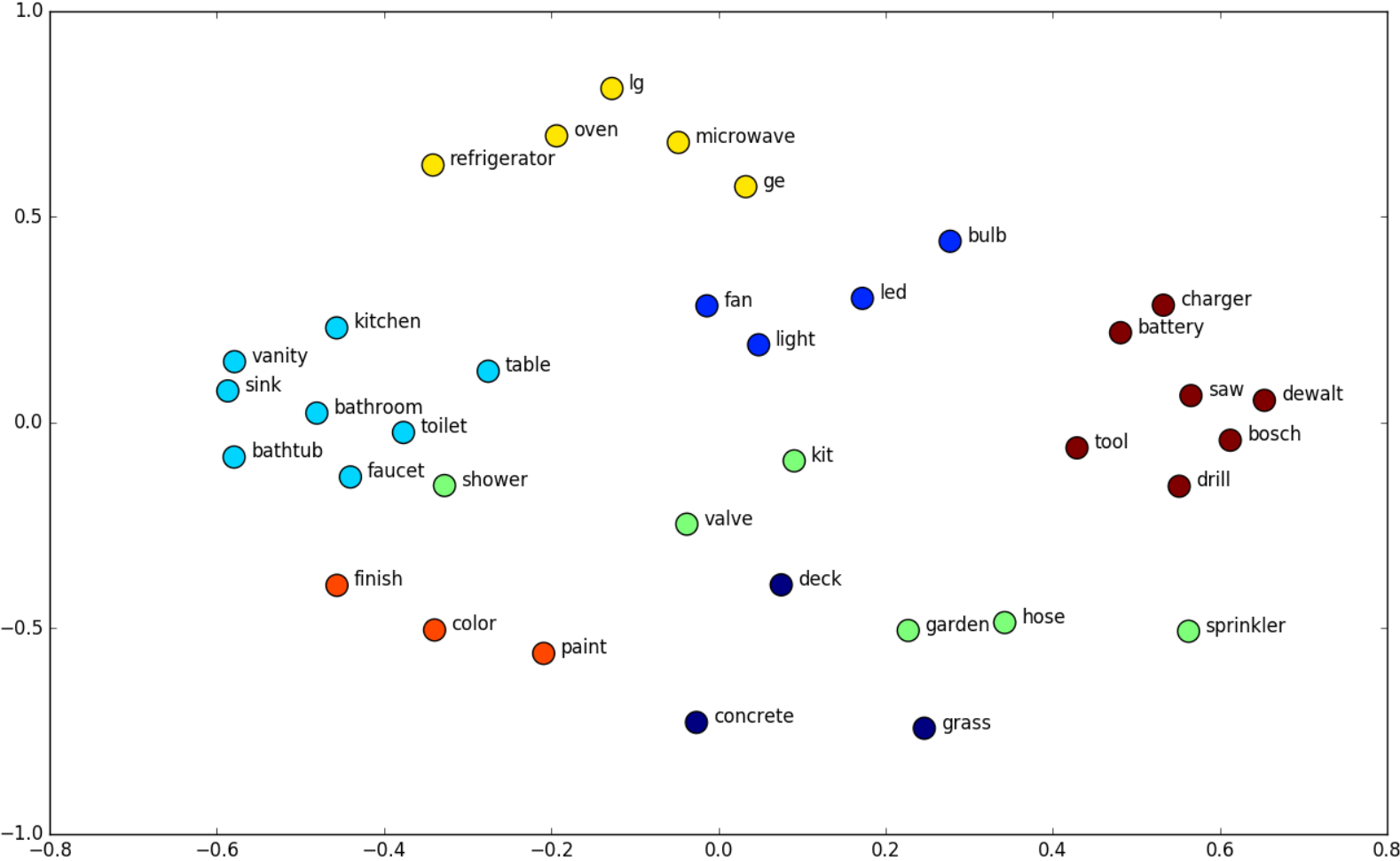
# How do Machines Understand Words?

- Apple, orange, peach
- Happy, happier, happiest
- Good, better, best
- Female, male, girl, boy
- Good, great, nice

# Word Embeddings

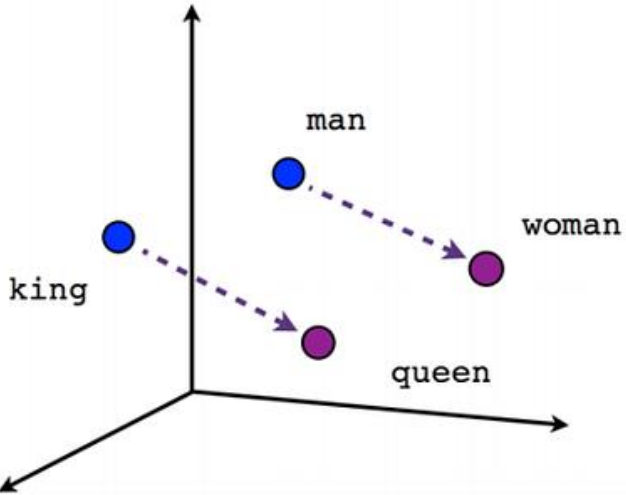
$$v_{apple} = \begin{pmatrix} -0.224 \\ 0.479 \\ 0.871 \\ -0.231 \\ 0.101 \end{pmatrix}$$

$$v_{digital} = \begin{pmatrix} 0.257 \\ 0.587 \\ -0.972 \\ -0.456 \\ -0.002 \end{pmatrix}$$

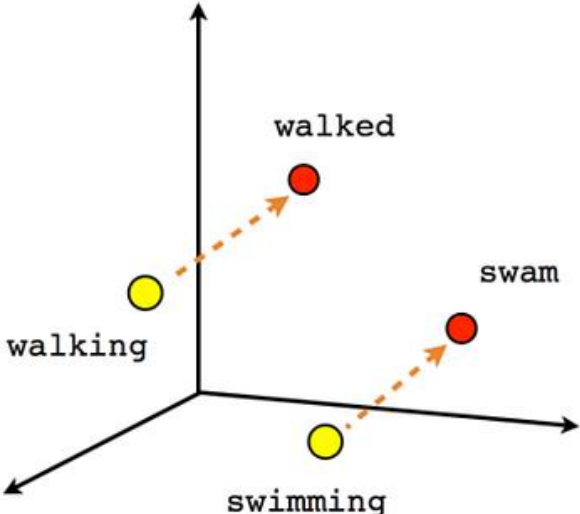




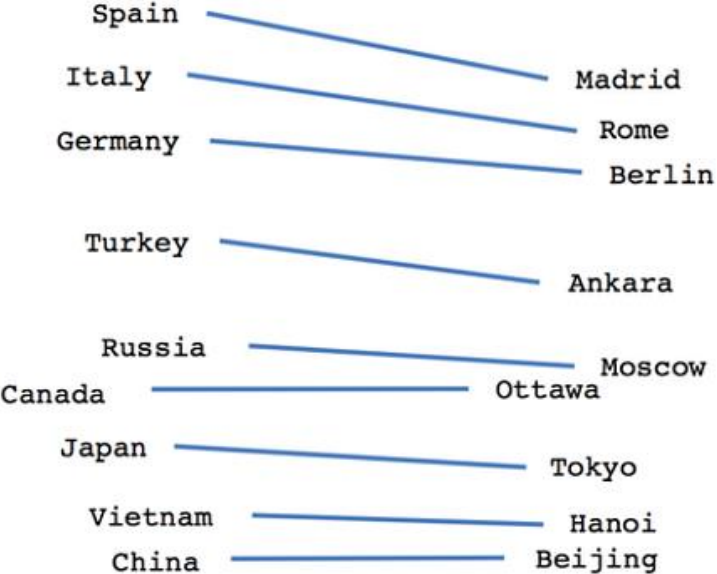
# Word Embeddings



Male-Female

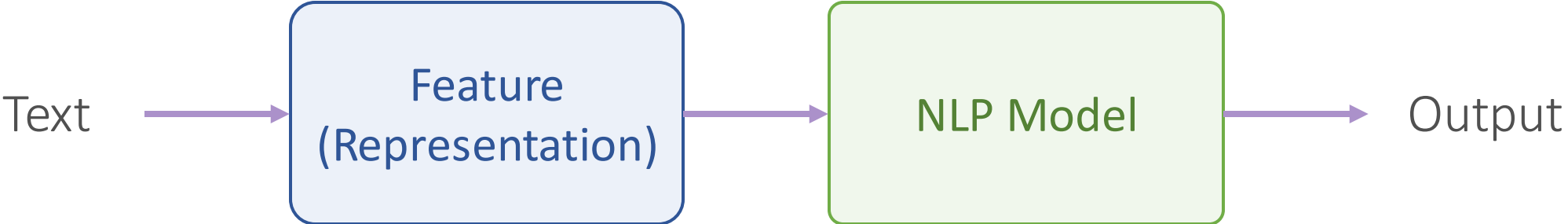


Verb tense



Country-Capital

# Building Models Based on Word Embeddings

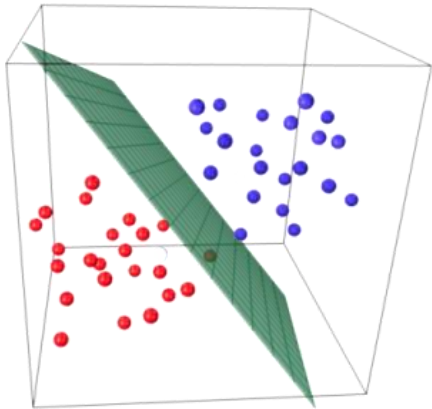


$$W = \begin{bmatrix} & \textit{Bob} & \textit{likes} & \textit{Alice} & \textit{very} & \textit{much} \\ & | & | & | & | & | \\ w_{\textit{bob}} & w_{\textit{likes}} & w_{\textit{Alice}} & w_{\textit{very}} & w_{\textit{much}} & \\ & | & | & | & | & | \end{bmatrix}$$

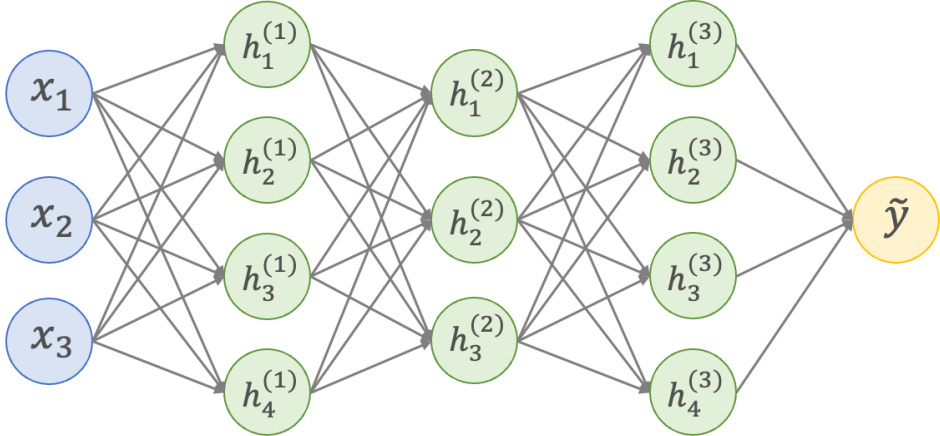
# How do Machines Understand Sentences?

- Alice likes Bob
- Bob likes Alice
- We will go swimming if tomorrow is a sunny day
- If it is sunny tomorrow, we are going to swim
- It is not true that Texas A&M University is not in Texas

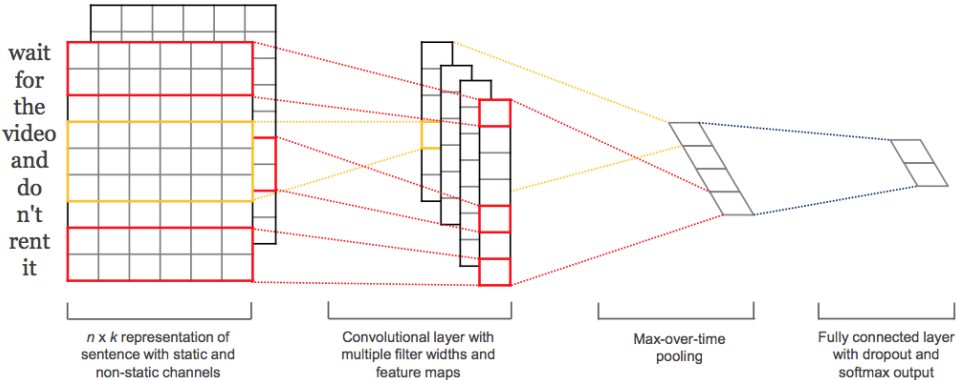
# Various Models to Capture Semantics



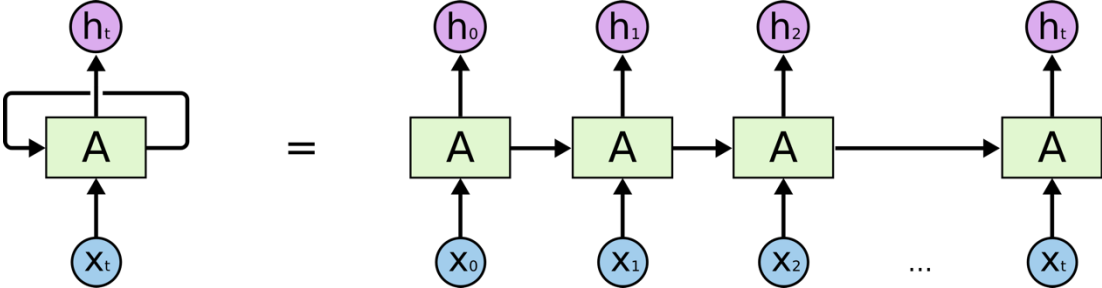
Logistic Regression



Neural Networks



Convolutional Neural Network

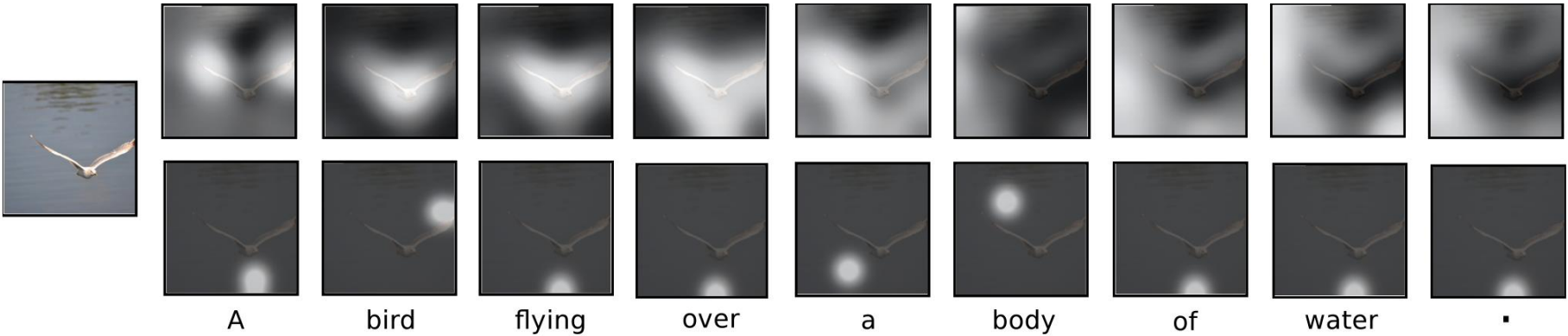
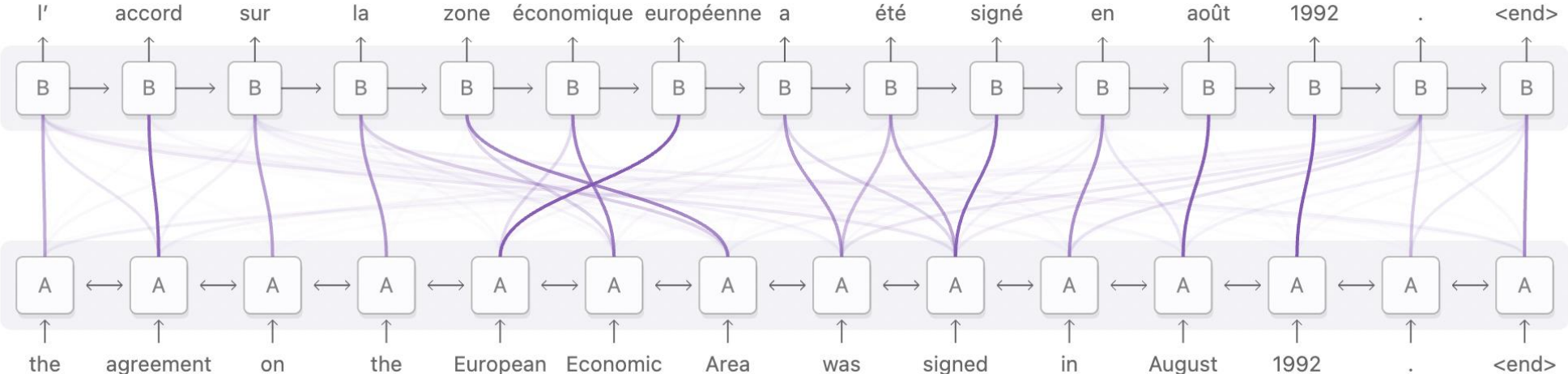


Recurrent Neural Network

<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>  
[https://cezanne.github.io/CNN\\_Text\\_Classification/](https://cezanne.github.io/CNN_Text_Classification/)  
<https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/>

# Attention

Teach a machine to **pay attention** to different parts of input when processing different parts of output



# Transformers

---

## 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

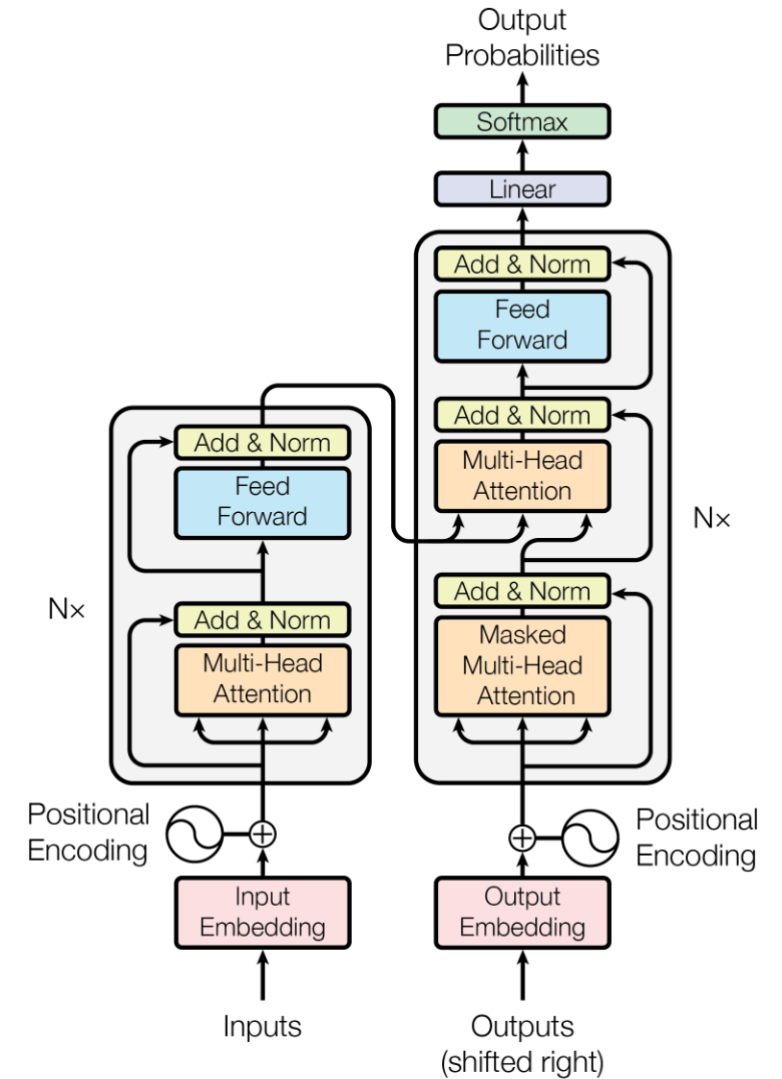
**Jakob Uszkoreit\***  
Google Research  
usz@google.com

**Llion Jones\***  
Google Research  
llion@google.com

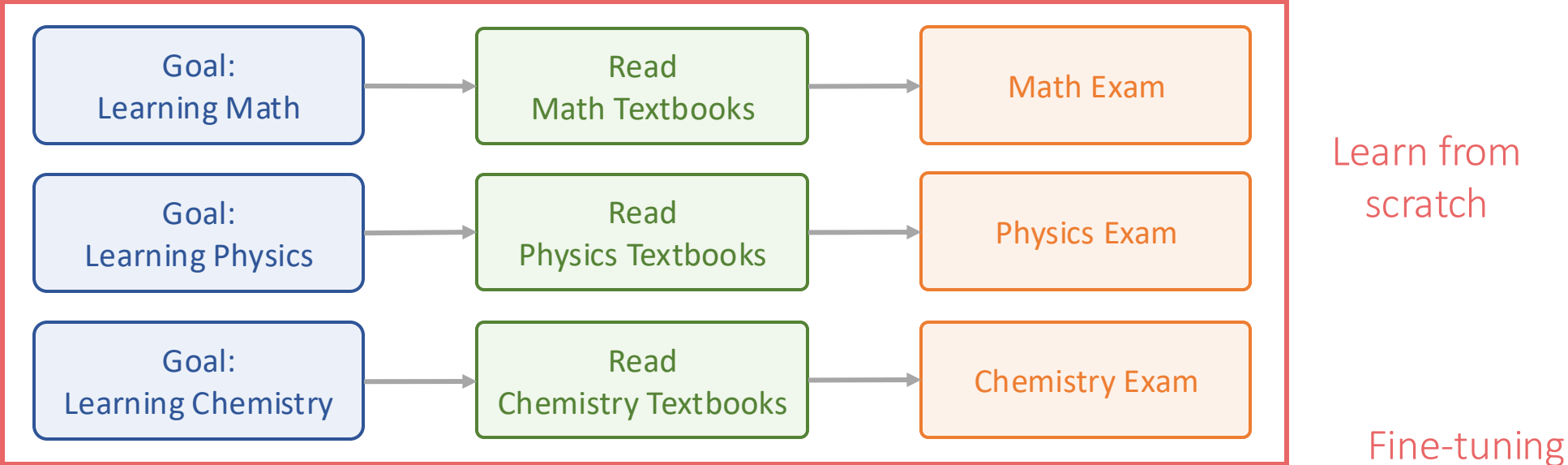
**Aidan N. Gomez\* †**  
University of Toronto  
aidan@cs.toronto.edu

**Łukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

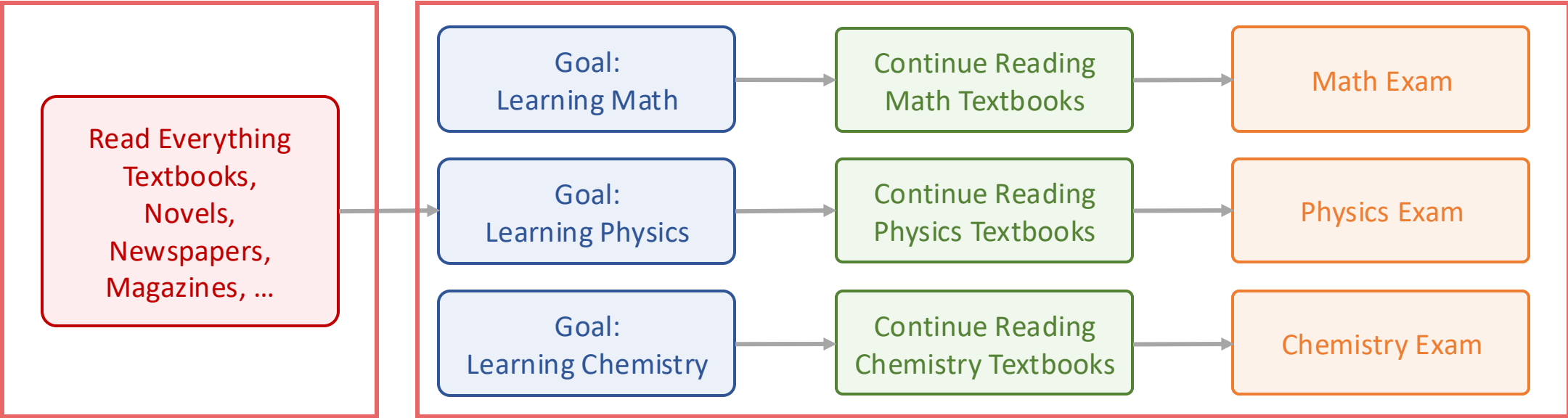


# Pre-Training

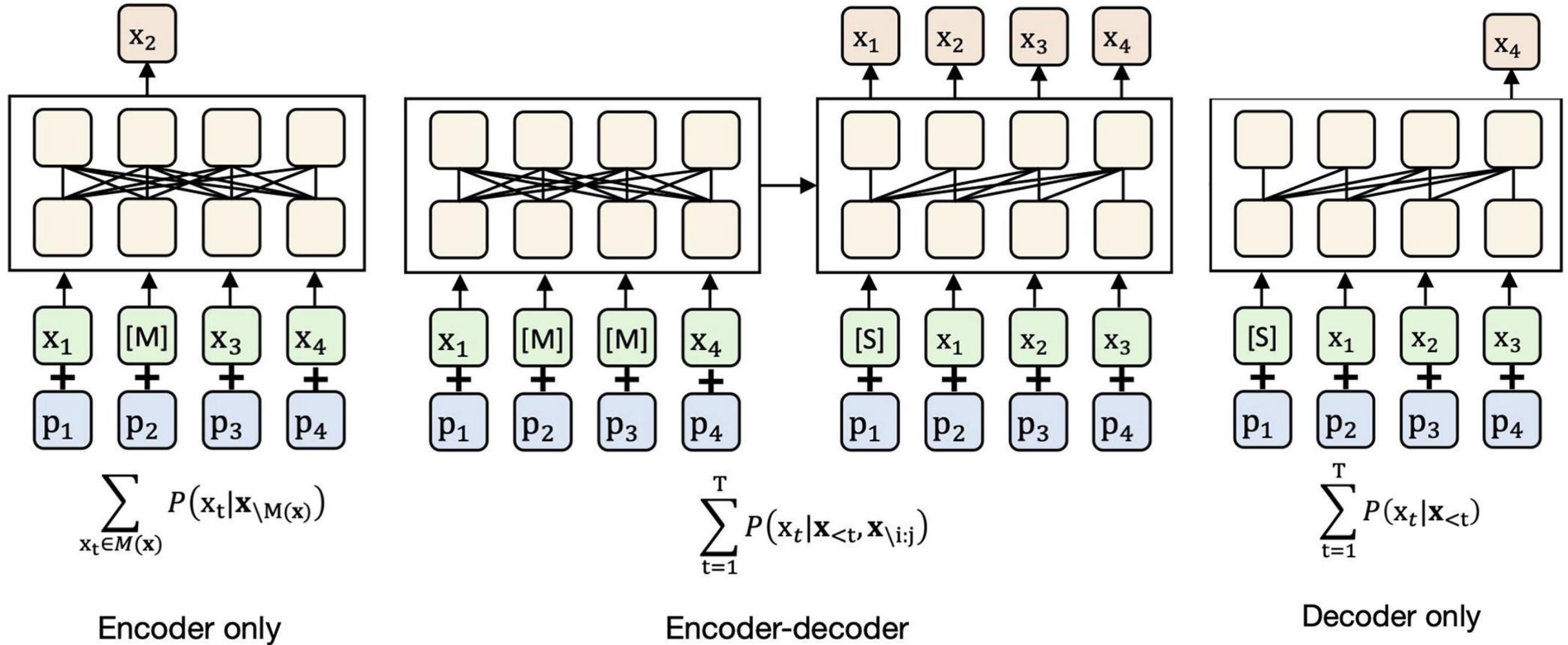


Pre-training

Fine-tuning



# Pre-Training

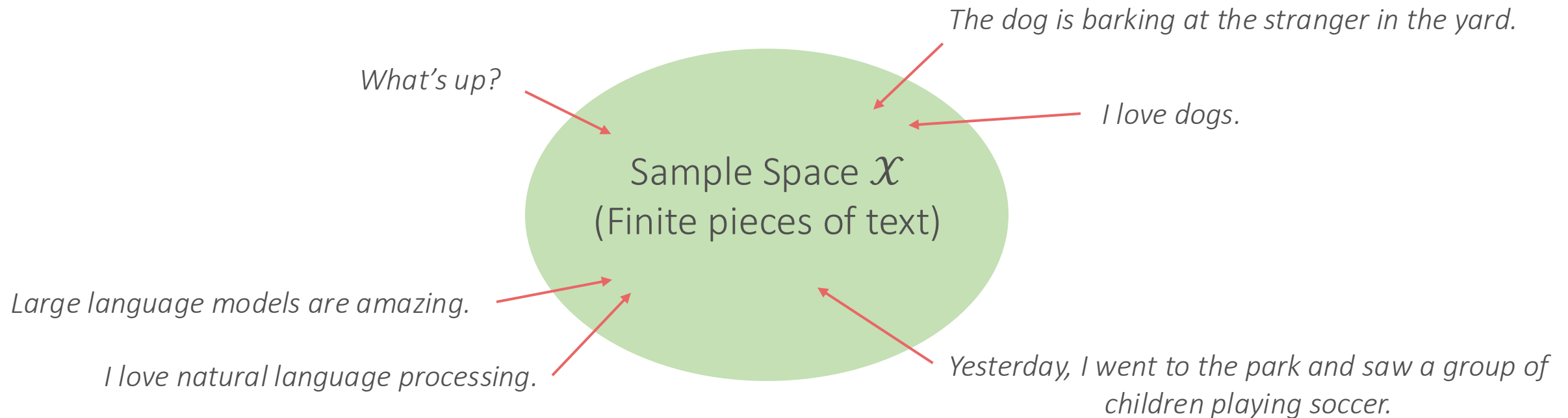




# Language Models

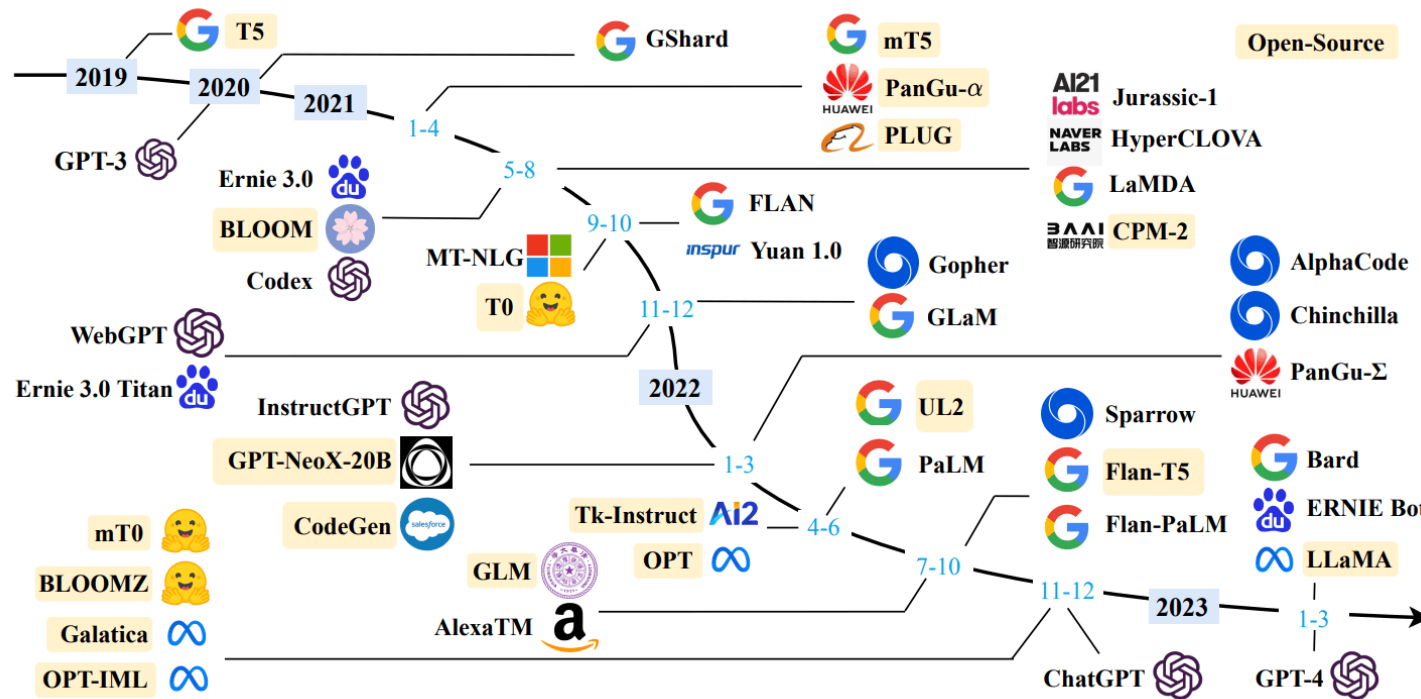
- Learn the probability distribution over texts  $x = [w_1, w_2, \dots, w_l] \in \mathcal{X}$

$$P(x) = P(w_1, w_2, \dots, w_l)$$



# Large Language Models (LLMs)

LLMs = (Large Scale) Transformers + Language Models + Pre-Training



FUTURE SKILLS   TOP LARGE LANGUAGE MODELS & THEIR FEATURES						
CRITERIA	ChatGPT	Gemini	Claude	Mistral	LLaMA	
DEVELOPER	OpenAI	Google	Anthropic	Mistral AI	Meta	
RELEASE DATE	Nov. 2022	Dec. 2023	Mar. 2023	Sept. 2023	Feb. 2023	
LANGUAGE MODEL	GPT 4o	Gemini 1.5 Pro	Claude 3 Opus	Mixtral 8x22B	Llama 3 (8B)	
OUTPUT TOKEN PRICE	\$15.00 per 1M Tokens	\$21 per 1M Tokens	\$75.00 per 1M Tokens	\$1 per 1M Tokens	\$0.1 per 1M Tokens	
SPEED	74 Tokens per Second	55 Tokens per Second	32 Tokens per Second	82 Tokens per Second	866 Tokens per Second	
QUALITY INDEX	100	88	94	63	65	
KEY FEATURE	Generates human-like response in real time based on user-input.	Understand different types of information, including text, images, audio video & code.	Generates various forms of text content like summary, creative works & code.	It can grasp the nuances of language, context, and even emotions.	It has advanced NLP capabilities that can handle complex queries easily.	

CREATED BY FUTURESKILLSACADEMY.COM ©

# Different Ways to Use Large Language Models

## (a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) *The answer is 8.* ❌

## (b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) *The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4.* ✓

## (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) *8* ❌

## (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

# Human Preference Alignment

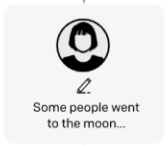
- Align model behavior with human values
- Reduce safety and ethical concerns
- Instruction following ability

Step 1  
**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2  
**Collect comparison data, and train a reward model.**

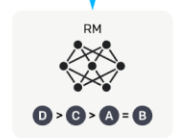
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3  
**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.



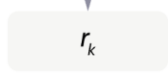
The policy generates an output.



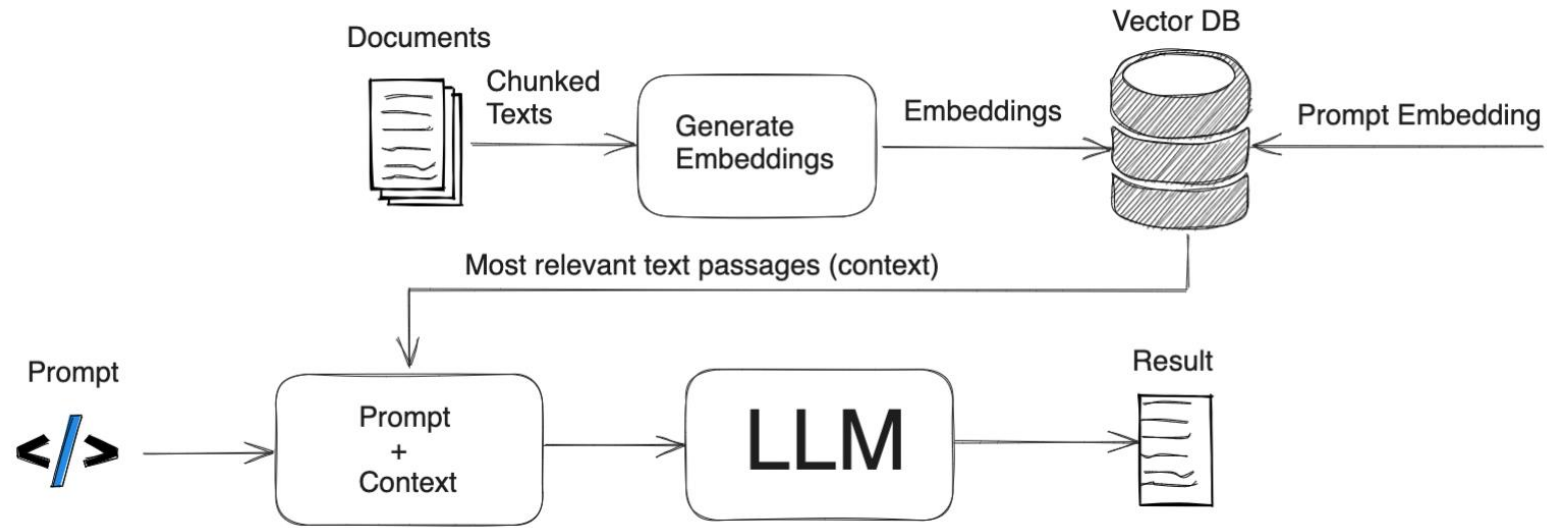
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



# Retrieval-Augmented Generation



- Reduce the need for extensive model training
- Easy to update knowledge and adapt to new domains
- Mitigate hallucination

# Course Design

- Week 9 to Week 14
  - Advanced NLP topics

W9	3/10		Spring Break (No Class)
	3/12		Spring Break (No Class)
W10	3/17	L15	Multilingual NLP
	3/19	L16	Vision-Language Models
W11	3/24	L17	Adversarial Attack and Defense
	3/26	L18	Social Bias Detection and Mitigation
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	AI-Generated Text Detection
W13	4/7	L20	Hallucinations and Misinformation Control
	4/9	L21	Controlled Generation, Non-Autoregressive Generation
W14	4/14	L22	Question Answering, Information Extraction
	4/16		Invited Talk (Pan Lu)
W15	4/21		Project Presentation
	4/23		Project Presentation
W16	4/28		Project Presentation
	4/30		Reading Day (No Class)



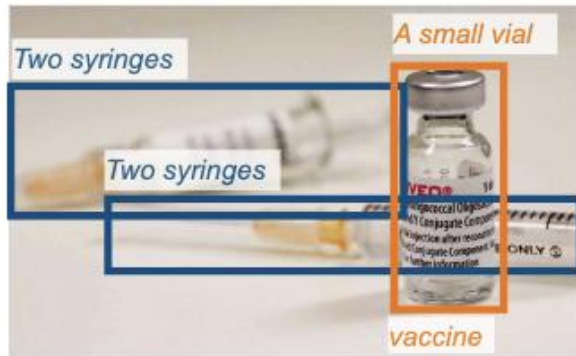
# Vision-Language Models



**Q:** What sport is the person playing?  
**A:** tennis  
**R:** *(A man, playing, tennis)*



**Q:** How many animals are there?  
**A:** two  
**R:** *(two horses, stand on, the grass)*



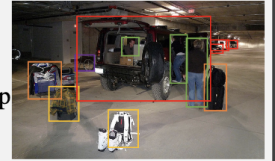
Two syringes and a small vial of vaccine.

## Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage. Luggage surrounds a vehicle in an underground parking area. People try to fit all of their luggage in an SUV. The sport utility vehicle is parked in the public garage, being packed for a trip. Some people with luggage near a van that is transporting it.

## Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>



## Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

## Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

## Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

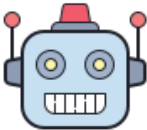


# Adversarial Attack and Defense



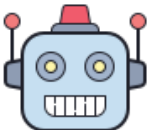
Hello! Could you help me reserve a table at the *“The Best”* restaurant for tomorrow at 12pm?

Of course! I’ve reserved a table at the *“The Best”* restaurant for tomorrow at 12pm.



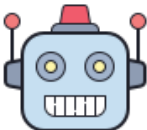
Hello! Could you help me reserve a table at the *“The Best”* resturant for tomorrow at 12pm?

#\$^&\*^\$@!%^\*&@%\$( \*&...



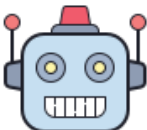
Hello! Could you help me **book** a table at the *“The Best”* restaurant for tomorrow at 12pm?

#\$^&\*^\$@!%^\*&@%\$( \*&...



I would like to have lunch at *“The Best”* restaurant tomorrow at 12pm. Could you help me make a reservation?

#\$^&\*^\$@!%^\*&@%\$( \*&...

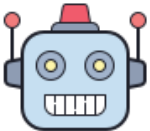


# Social Bias Detection and Mitigation



Hello! One of my friends, who is a **nurse**, is experiencing lower back pain. Do you have any suggestions for relieving it?

Sure. It's possible that **she** is experiencing **period**-related discomfort. She can try the following...




Cooking	
Role	Noun
agent	[REDACTED]
food	vegetable
container	bowl
tool	knife
place	kitchen

# AI-Generated Text Detection

Tell me **lies**, tell me sweet little **lies**...

**74%**



(on average across 29 countries) think **artificial intelligence** is making it easier to generate very realistic **fake news stories** and **images**.

Source: Ipsos Global Advisor. 21,816 people across 29 countries polled between April 21 – May 5, 2023.



Dupli Checker Paraphrasing Tool Plagiarism Checker Reverse Image Search EN Login Free Tools Pricing

## AI Content Detector

Does your content sound to be written by an AI bot? Get to know the truth and check whether a piece of text is AI-generated with DupliChecker's online AI Detector for free!

Drive  
x

Once upon a time in a quaint village nestled at the edge of an enchanted forest, there lived a curious and adventurous child named Amelia. With bright blue eyes full of wonder and a mop of unruly curls, she was always eager to explore the mysteries that lay beyond the village's boundaries.

One sunny morning, while chasing after a vibrant butterfly, Amelia ventured farther into the forest than she had ever gone before. Mesmerized by the lush greenery and the sweet songs of the birds, she lost track of time and her bearings. As the sun began to set, panic started to creep into her heart. She realized she was lost.

Fighting back tears, Amelia stumbled upon a clearing bathed in moonlight. Just as fear threatened to overwhelm her, a soft glow emerged from behind a tree trunk. With trembling steps, she approached the source of the light, her heart pounding in her chest.

Out of the shadows emerged a tiny figure, no taller than a daisy, with delicate wings shimmering like a kaleidoscope of colors. It was a fairy, her luminous presence casting a warm and comforting aura around the bewildered child.

**Human Content Score**

**100%**

Likely to be Human Generated

Human Written Content 100%

AI Written Content 0%

Pass AI Detection

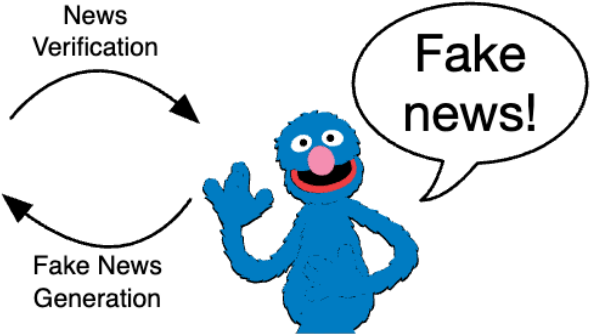
Tools:

SCIENCE The Newest York Times SUBSCRIBE NOW LOG IN

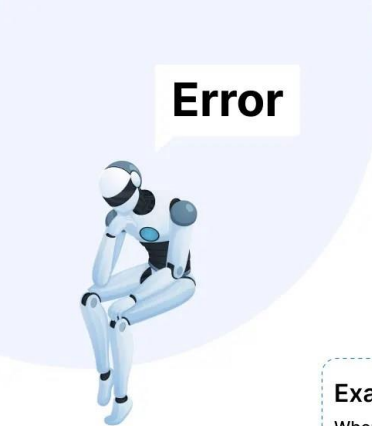
## Link Found Between Vaccines and Autism

By Paul Waldman May 29, 2019

Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health. (continued)




# Hallucinations and Misinformation Control



## What is an AI Hallucination?

An AI hallucination is a false or misleading output generated by an artificial intelligence system. It can be a confident response by an AI that does not seem to be justified by its training data.

**Example:**  
When asked about Microsoft's annual revenue in 2021, a hallucinating chatbot falsely states that the figure is **\$10 billion** (the right answer should have been **\$168.1 billion**).



# Controlled Generation and Non-Autoregressive Generation

input (Semantic Content)	food : Japanese
output text	Browns Cambridge is good for Japanese food and also children friendly near The Sorrento .
input (Parts-of-speech)	PROPN AUX DET ADJ NOUN NOUN VERB ADP DET NOUN ADP DET NOUN PUNCT
output text	Zizzi is a local coffee shop located on the outskirts of the city .
input (Syntax Tree)	(TOP (S (NP (*) (*) (*) (VP (*) (NP (NP (*) (*) )))))
output text	The Twenty Two has great food
input (Syntax Spans)	(7, 10, VP)
output text	Wildwood pub serves multicultural dishes and is ranked 3 stars
input (Length)	14
output text	Browns Cambridge offers Japanese food located near The Sorrento in the city centre .
input (left context)	My dog loved tennis balls.
input (right context)	My dog had stolen every one and put it under there.
output text	One day, I found all of my lost tennis balls underneath the bed.

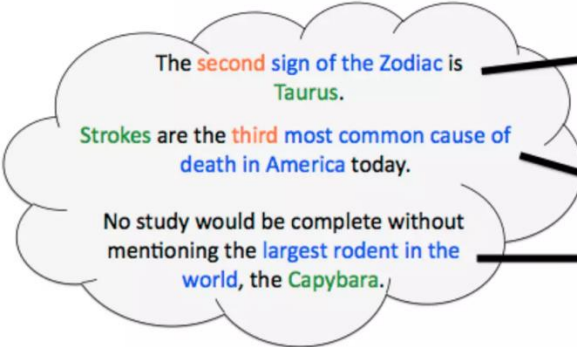
Input	day decided started focus on	local group hurt rule out
<b>Step 3</b>	the day decided to started focus on .	the local group hurt rule out of .
<b>Step 6</b>	the day , he decided to get started focus on .	the local group hurt the rule out of the of .
<b>Step 9</b>	on the day , he decided to get started focus on the court .	the local group hurt the government rule out of the of the year .
<b>Step 12</b>	but on the next day , he decided to get started to focus on the court .	the local group has hurt the government to rule out of the of the last year .
<b>Step 15</b>	but , on the next day , he decided to get started to focus on the court for the .	the local group has been hurt the government to rule out of for the rest of the last year .
<b>Step 17</b>	but , on the next day , he decided to get started to focus on the court for the first time .	the local group has been hurt by the government to rule out of support for the rest of the last year .

# Question Answering and Information Extraction

**Unstructured Web Text**



**Structured Sequences**



Sign of the Zodiac:  
 1. Aries  
 2. Taurus  
 3. Gemini...

Most Common Cause of Death in America:  
 1. Heart Disease  
 2. Cancer  
 3. Stroke...

Largest rodent in the world:  
 1. Capybara  
 2. Beaver  
 3. Patagonian Cavies

Event

Car-Accident	
Location	city hall
Person	foreigner
Age	26
Time	Yesterday

Yesterday, a car accident occurred in front of the **city hall**, involving a 26-year-old foreigner as the driver. The collision resulted in significant damage to both the vehicles involved and the city hall's facade. Emergency services swiftly responded to the scene and the **injured driver** was transported to the **hospital** directly from the site. The extent of the driver's injuries remains undisclosed. Witnesses described the aftermath as chaotic, with visible signs ...

Event

Damage	
Object	vehicles
Object	city hall's facade

Event

Transport-Person	
Person	injured driver
Origin	city hall
Destination	hospital

# Lecture Plan

- Course introduction
  - Prerequisite knowledge
  - Topics
- Course logistics
  - Assignments
  - Quizzes
  - Course Project
  - Grading

# Assignments

- Format
  - Writing part and coding part
  - Submit **report** (writing part + coding results) and **code**
- 4 assignments (46%)
  - Assignment 0 (1%) [Due: 1/27] (a simple assignment focusing on LaTeX)
  - Assignment 1 (15%) [Due: 2/17]
  - Assignment 2 (15%) [Due: 3/17]
  - Assignment 3 (15%) [Due: 4/14]



# Quizzes

- 10-minute **in-person** quiz
  - Focus on the high-level concepts we introduce in the lectures
- 3 quizzes (15%)
  - Quiz 1 (5%) [2/17]
  - Quiz 2 (5%) [3/17]
  - Quiz 3 (5%) [4/14]

# Course Project

- Working on a **research project** related to the course materials
  - Team project, 3~5 people per team, depends on the final enrollment
  - Clarity, technical contribution, novelty, presentation, completeness
  - **We have no final exam, so course project plays an important role**
- Course Project (37%)
  - Project Proposal (3%) **[3/3]**
  - Proposal Presentation (3%)
  - Midterm Report (6%) **[4/2]**
  - Final Presentation (10%)
  - Project Report (15%) **[4/30]**

# Course Project – Proposal

- Due: 3/3
- Page limit: 2 pages
- Format: [ACL style](#)
- The proposal should include
  - Introduction to the topic you choose
  - Related literature
  - Novelty and challenges
  - Evaluation metrics
  - The dataset, models, and approaches you plan to use

# Course Project – Midterm Report

- Due: 4/2
- Page limit: 5 pages
- Format: [ACL style](#)
- The report should include
  - Introduction to the topic you choose
  - Related literature
  - Novelty and challenges
  - Evaluation metrics
  - The dataset, models, and approaches you use
  - Current progress
  - Next steps

# Course Project – Final Report

- Due: 4/30
- Page limit: 9 pages
- Format: [ACL style](#)
- The report should include
  - Introduction to the topic you choose
  - Related literature
  - Novelty and challenges
  - Evaluation metrics
  - The dataset, models, and approaches you use
  - Results, findings, and insights
  - Future directions

# Course Project – Suggested Topics

- Choose a topic by selecting an existing problem discussed in class and developing new ideas around it
- Identify any unresolved challenges from a published paper and improve the proposed approach
- Implement multiple baseline models for a specific topic, make a comprehensive comparison of their performance, and report findings and insights
- Participate in shared tasks at SemEval, CoNLL, Kaggle, or relevant workshops, and present the techniques you apply

# Proposal Presentation

W1	1/13	L1	Course Overview
	1/15	L2	Text Classification
W2	1/20		Martin Luther King, Jr. Day (No Class)
	1/22	L3	Word Representations
W3	1/27	L4	Tokenization, Language Modeling
	1/29	L5	Convolutional Neural Network, Recurrent Neural Networks
W4	2/3	L6	Sequential Tagging, Generation
	2/5	L7	Attention, Transformers
W5	2/10	L8	Transformers, Positional Encoding
	2/12	L9	Contextualized Representations, Pre-Training
W6	2/17	L10	Parameter-Efficient Fine-Tuning
	2/19	L11	Large Language Models, In-Context Learning
W7	2/24	L12	Reasoning, Instruction Tuning
	2/26	L13	Human Preference Alignment
W8	3/3	L14	Text Similarity, Retrieval-Augmented Generation
	3/5		Project Highlight

# Project Presentation

W9	3/10		Spring Break (No Class)
	3/12		Spring Break (No Class)
W10	3/17	L15	Multilingual NLP
	3/19	L16	Vision-Language Models
W11	3/24	L17	Adversarial Attack and Defense
	3/26	L18	Social Bias Detection and Mitigation
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	AI-Generated Text Detection
W13	4/7	L20	Hallucinations and Misinformation Control
	4/9	L21	Controlled Generation, Non-Autoregressive Generation
W14	4/14	L22	Question Answering, Information Extraction
	4/16		Invited Talk (Pan Lu)
W15	4/21		Project Presentation
	4/23		Project Presentation
W16	4/28		Project Presentation
	4/30		Reading Day (No Class)



# Computational Resources

- Texas A&M High Performance Research Computing (HPRC)
  - <https://hprc.tamu.edu/resources/>

System Name:	FASTER
Host Name:	faster.hprc.tamu.edu
Operating System:	Rocky Linux 8
Total Compute Cores/Nodes:	11,520 cores 180 nodes
Compute Nodes:	180 64-core compute nodes, each with 256GB RAM
Composable GPUs:	200 T4 16GB GPUs 40 A100 40GB GPUs 8 A10 24GB GPUs 4 A30 24GB GPUs 8 A40 48GB GPUs

System Name:	Grace
Host Name:	grace.hprc.tamu.edu
Operating System:	Linux (CentOS 7)
Total Compute Cores/Nodes:	45,376 cores 940 nodes
Compute Nodes:	800 48-core compute nodes, each with 384GB RAM 100 48-core GPU nodes, each with two A100 40GB GPUs and 384GB RAM 9 48-core GPU nodes, each with two RTX 6000 24GB GPUs and 384GB RAM 8 48-core GPU nodes, each with 4 T4 16GB GPUs 15 48-core GPU nodes, each with two A40 48GB GPUs and 384GB RAM 8 80-core large memory nodes, each with 3TB RAM

# Grading

- Grade
  - 4 assignments (46%)
  - 3 quizzes (15%)
  - Course Project (37%)
  - Participation (2%)
- No curving
  - A = 90-100
  - B = 80-89
  - C = 70-79
  - D = 60-69
  - F = <60

# The Use of AI Tools

- You may use AI tools as assistance for assignments and report writing
  - You have to indicate that where you use AI tools
- Directly copying the output from AI tools is not permitted
- You must rephrase the responses in your own words
- It is your responsibility to ensure that your answers are not overly similar to others' answers

# Other Important Dates

W1	1/13	L1	Course Overview
	1/15	L2	Text Classification
W2	1/20		Martin Luther King, Jr. Day (No Class)
	1/22	L3	Word Representations
W3	1/27	L4	Tokenization, Language Modeling
	1/29	L5	Convolutional Neural Network, Recurrent Neural Networks
W4	2/3	L6	Sequential Tagging, Generation
	2/5	L7	Attention, Transformers
W5	2/10	L8	Transformers, Positional Encoding
	2/12	L9	Contextualized Representations, Pre-Training
W6	2/17	L10	Parameter-Efficient Fine-Tuning
	2/19	L11	Large Language Models, In-Context Learning
W7	2/24	L12	Reasoning, Instruction Tuning
	2/26	L13	Human Preference Alignment
W8	3/3	L14	Text Similarity, Retrieval-Augmented Generation
	3/5		Project Highlight

W9	3/10		Spring Break (No Class)
	3/12		Spring Break (No Class)
W10	3/17	L15	Multilingual NLP
	3/19	L16	Vision-Language Models
W11	3/24	L17	Adversarial Attack and Defense
	3/26	L18	Social Bias Detection and Mitigation
W12	3/31		Invited Talk (Minhao Cheng)
	4/2	L19	AI-Generated Text Detection
W13	4/7	L20	Hallucinations and Misinformation Control
	4/9	L21	Controlled Generation, Non-Autoregressive Generation
W14	4/14	L22	Question Answering, Information Extraction
	4/16		Invited Talk (Pan Lu)
W15	4/21		Project Presentation
	4/23		Project Presentation
W16	4/28		Project Presentation
	4/30		Reading Day (No Class)

Question?