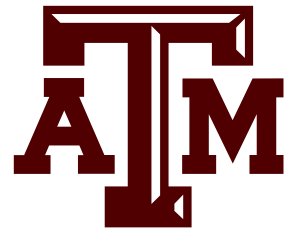


# CSCE 689: Special Topics in Trustworthy NLP

## Lecture 7: Text Similarity, Retrieval-Augmented Generation, Vision-Language Models

Kuan-Hao Huang  
khhuang@tamu.edu



(Some slides adapted from Chris Manning, Karthik Narasimhan, and Danqi Chen)

# Paper Summary

- Paper Summary (10%)
  - Starting from week 5, a paper summary of **two** papers will be due **each Monday**
  - Page limit: 1 page

W5	9/22	Human Preference Alignment	<div>Training language models to follow instructions with human feedback, NeurIPS 2022 Direct Preference Optimization: Your Language Model is Secretly a Reward Model, NeurIPS 2023 SimPO: Simple Preference Optimization with a Reference-Free Reward, NeurIPS 2024 Understanding R1-Zero-Like Training: A Critical Perspective, arXiv 2025</div>
	9/24	Bias Detection and Mitigation	<div>Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, NeurIPS 2016 Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, EMNLP 2017 BLIND: Bias Removal With No Demographics, ACL 2023 On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning, ACL 2023</div>

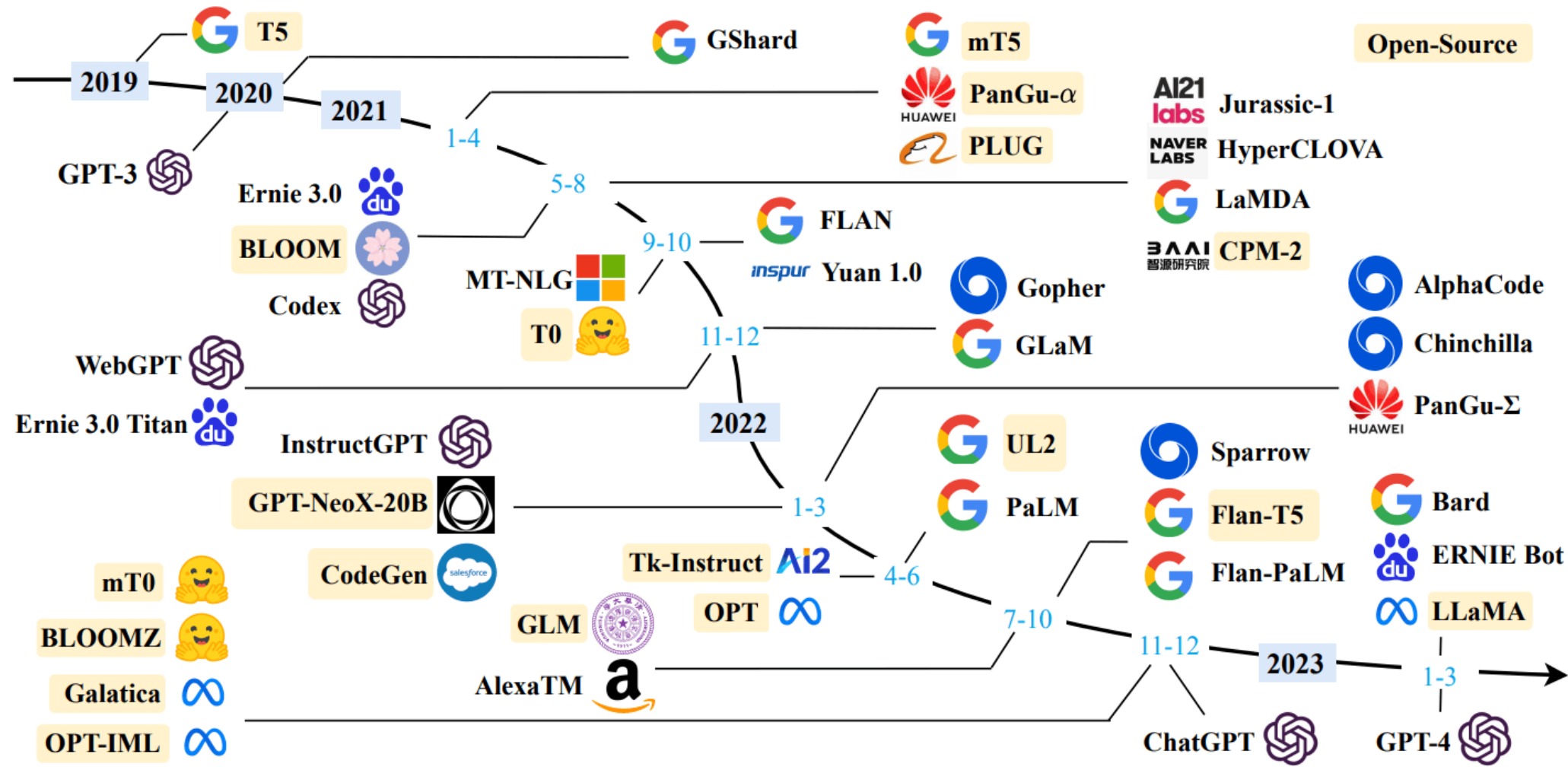
Choose 1 paper here

Choose 1 paper here

# Paper Summary

- A brief overview of the main objectives and contributions of the paper
- Key methodologies and approaches used in the study
- Significant findings and results
- Strengths and weaknesses of the paper

# Large Language Models





# Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

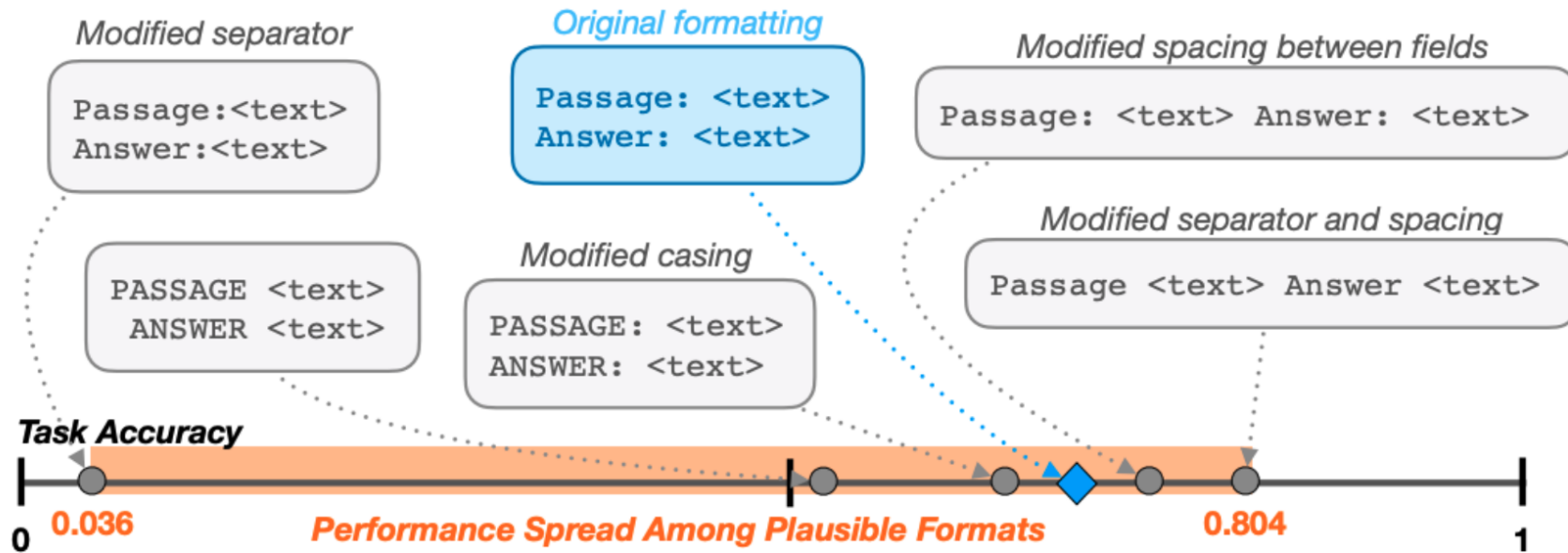
Completion

# A New Way to Use NLP Models

- Task-specific features + task-specific model
- General embeddings + task-specific model
- General embeddings + general model + task-specific fine-tuning
- General embeddings + general model + task-specific prompting

# Prompt Engineering

- Craft inputs to guide LLMs models effectively



# Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

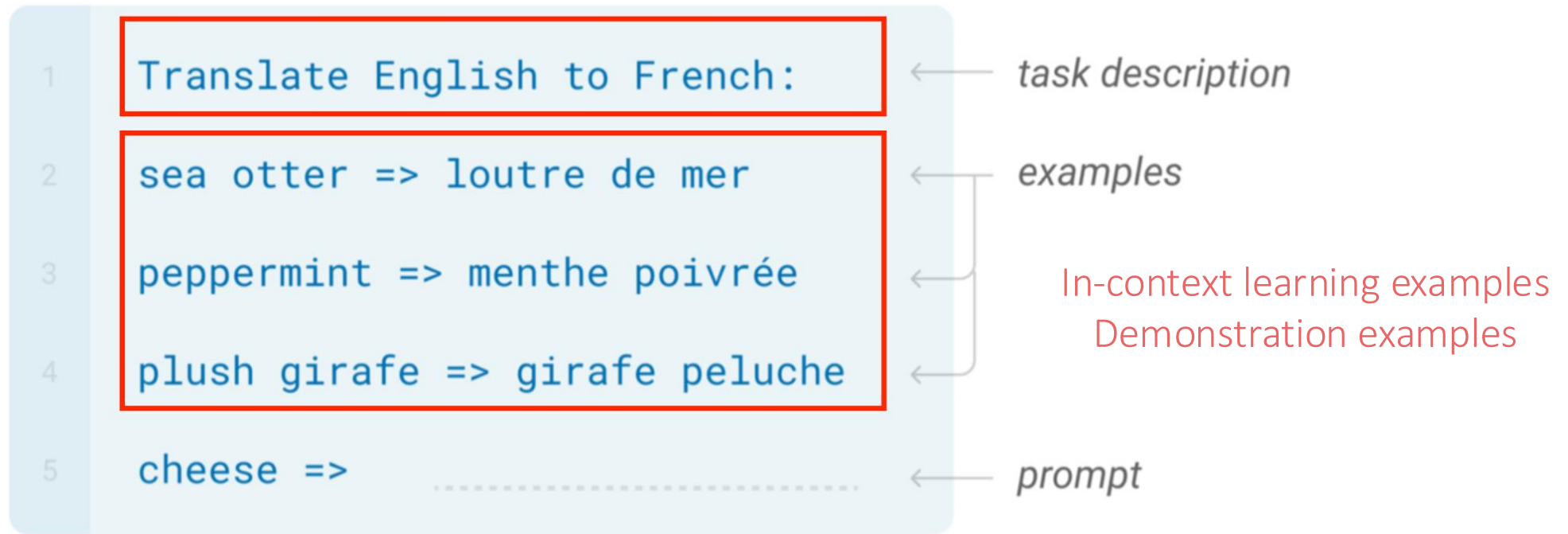
Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

Completion

# Few-Shot Prompting / In-Context Learning



# Few-Shot Prompting / In-Context Learning

Input: 2014-06-01

Output: !06!01!2014!

Input: 2007-12-13

Output: !12!13!2007!

Input: 2010-09-23

Output: !09!23!2010!

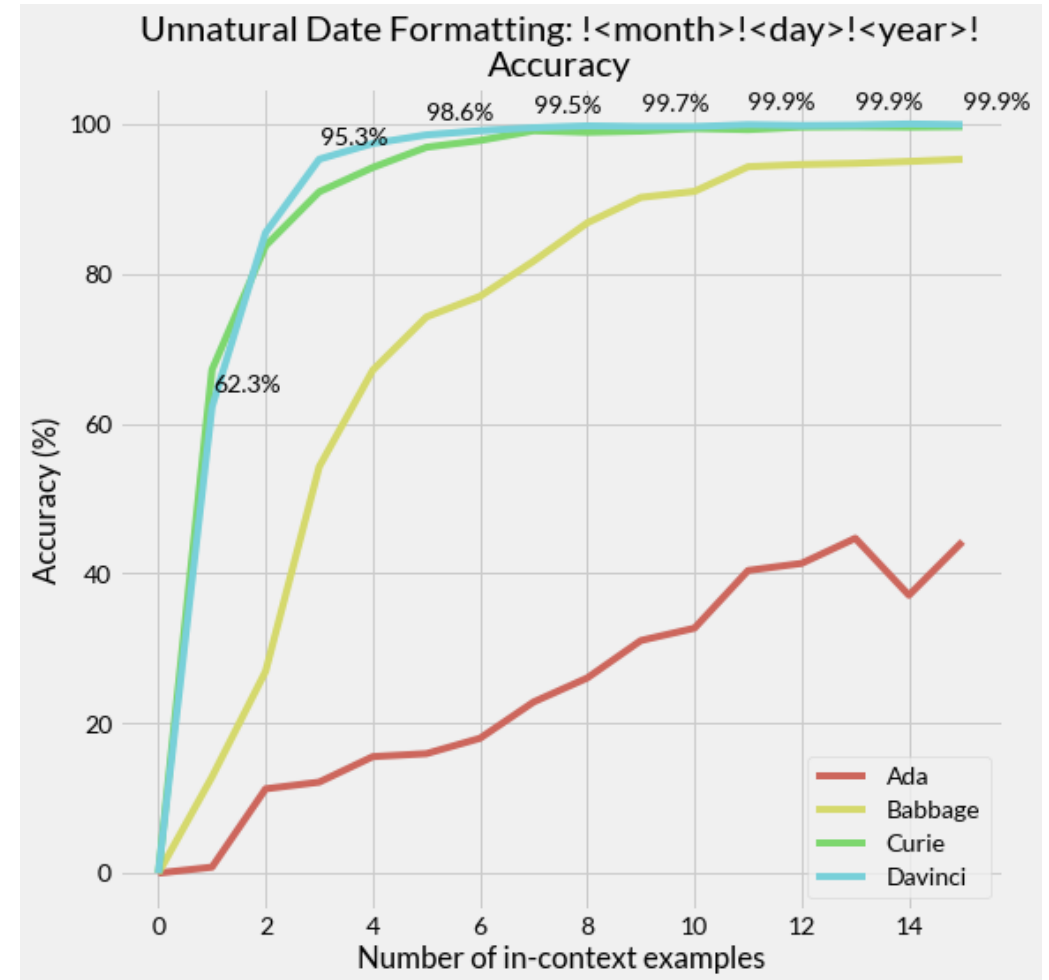
Input: **2005-07-23**

Output: **!07!23!2005!**

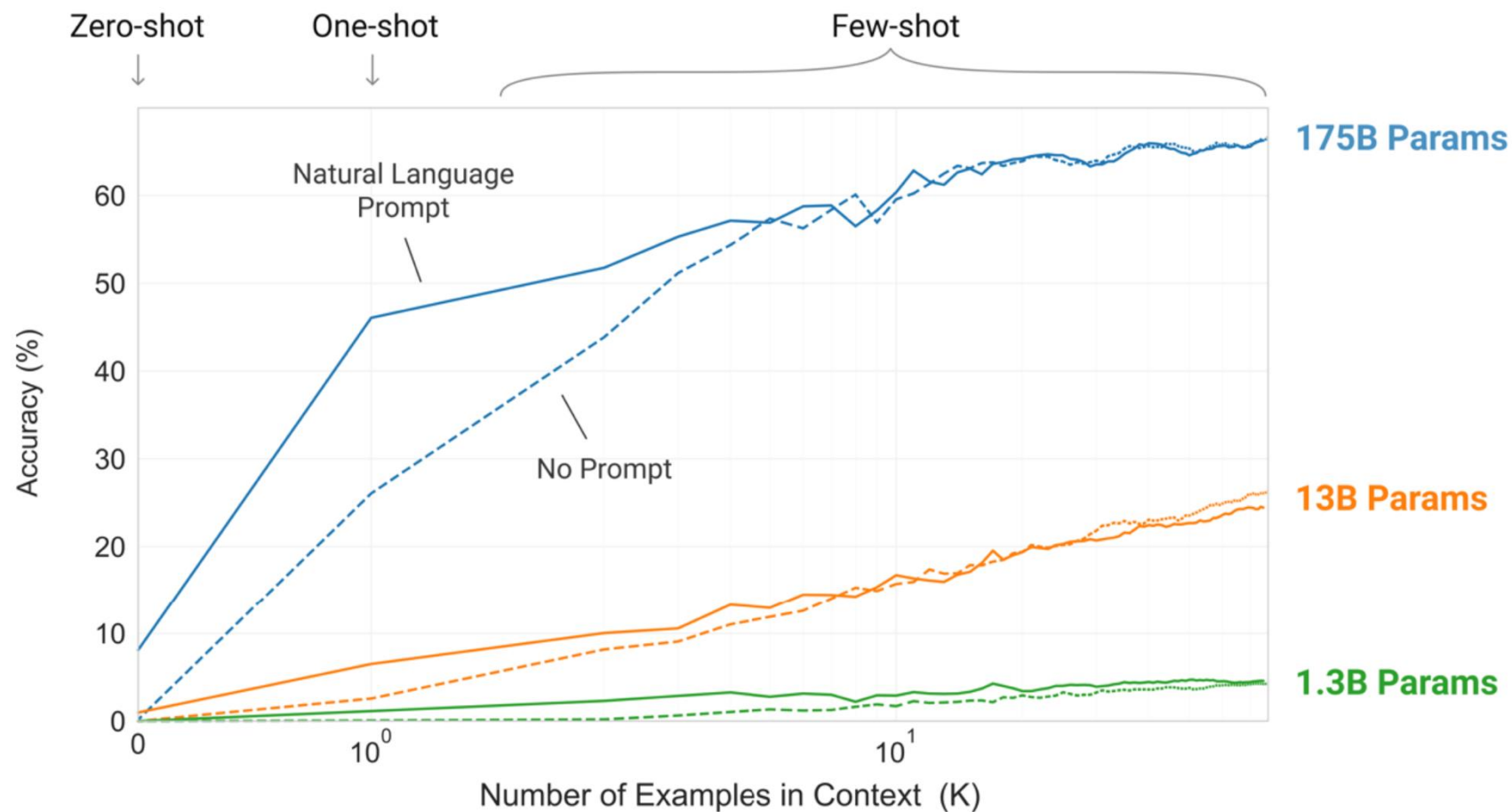
*in-context  
examples*

*test example*

*! - - - model completion*



# Few-Shot Prompting / In-Context Learning



# Chain-of-Thought (CoT) Prompting

- Ask the model to explain its reasoning before making an answer

## Standard Prompting

### Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain-of-Thought Prompting

### Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

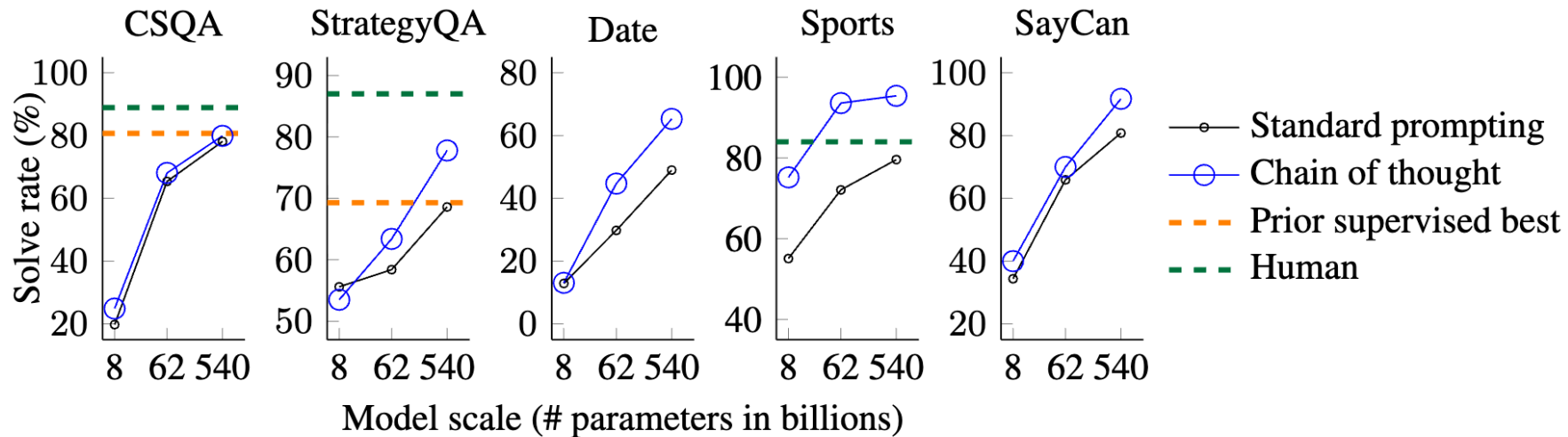
### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



# Chain-of-Thought (CoT) Prompting

- Ask the model to explain its reasoning before making an answer



# Zero-Shot Chain-of-Thought Prompting

(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. ✓




# Zero-Shot Chain-of-Thought Prompting

	MultiArith	GSM8K
<b>Zero-Shot</b>	<b>17.7</b>	<b>10.4</b>
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
<b>Zero-Shot-CoT</b>	<b>78.7</b>	<b>40.7</b>
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
<b>Zero-Plus-Few-Shot-CoT (8 samples) (*2)</b>	<b>92.8</b>	<b>51.5</b>
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
<b>PaLM 540B: Zero-Shot</b>	<b>25.5</b>	<b>12.5</b>
<b>PaLM 540B: Zero-Shot-CoT</b>	<b>66.1</b>	<b>43.0</b>
<b>PaLM 540B: Zero-Shot-CoT + self consistency</b>	<b>89.0</b>	<b>70.1</b>
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4


# Reasoning Models

Question: If  $a > 1$ , then the sum of the real solutions of  $\sqrt{a - \sqrt{a + x}} = x$  is equal to


Response: <think>  
To solve the equation  $\sqrt{a - \sqrt{a + x}} = x$ , let's start by squaring both ...  
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$ .  
Rearrange to isolate the inner square root term:  
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$   
...  
**Wait, wait. Wait. That's an aha moment I can flag here.**  
Let's reevaluate this step-by-step to identify if the correct sum can be ...  
We started with the equation:  
 $\sqrt{a - \sqrt{a + x}} = x$   
First, let's square both sides:  
 $a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$   
Next, I could square both sides again, treating the equation: ...  
...

GPT-4o	o1	o3
		
Fast, intelligent, flexible GPT model	Previous full o-series reasoning model	Reasoning model for complex tasks, succeeded by GPT-5
Learn more	Learn more	Learn more
Playground	Playground	Playground
<div>Intelligence<div>●●●</div></div>	<div>Reasoning<div>●●●●</div></div>	<div>Reasoning<div>●●●●●</div></div>
<div>Speed<div>⚡⚡⚡</div></div>	<div>Speed<div>⚡</div></div>	<div>Speed<div>⚡</div></div>
<div>Input<div>🗨️🗣️🔍</div></div>	<div>Input<div>🗨️🗣️🔍</div></div>	<div>Input<div>🗨️🗣️🔍</div></div>
<div>Output<div>🗨️🔍🔍</div></div>	<div>Output<div>🗨️🔍🔍</div></div>	<div>Output<div>🗨️🔍🔍</div></div>
<div>Reasoning tokens<div>⊗</div></div>	<div>Reasoning tokens<div>✔️</div></div>	<div>Reasoning tokens<div>✔️</div></div>


# LLaMA Series

- Creator:  **Meta** <https://ai.meta.com/blog/meta-llama-3/>
- **Goal:** Strong and safe open language model
- **Unique features:** Open models with strong safeguards and chat tuning, good performance


# Mistral/Mixtral

- Creator:  **MISTRAL  
AI\_** <https://mistral.ai/en/news/mixtral-of-experts>
- **Goal:** Strong and somewhat multilingual open language model
- **Unique features:** Speed optimizations, including GQA and Mixture of Experts

# Qwen Series


- Creator:  <https://qwen.ai/>
- **Goal:** Strong multilingual (esp. English and Chinese) language model
- **Unique features:** Large vocabulary for multilingual support, strong performance

# DeepSeek Series

- Creator:  **deepseek** <https://www.deepseek.com/>
- **Goal:** Strongest open-weight language model so far
- **Unique features:** Relatively low-cost reinforcement-learning-based alignment for reasoning



# OLMo

- Creator:  <https://allenai.org/olmo>
- **Goal:** Better science of state-of-the-art LMs
- **Unique features:** fully open-source and fully documented model, instruction tuned etc.

# Proprietary LLMs

GPT Series



<https://openai.com/>

Gemini Series



<https://gemini.google.com/>

Claude Series



<https://claude.ai/>

Grok Series



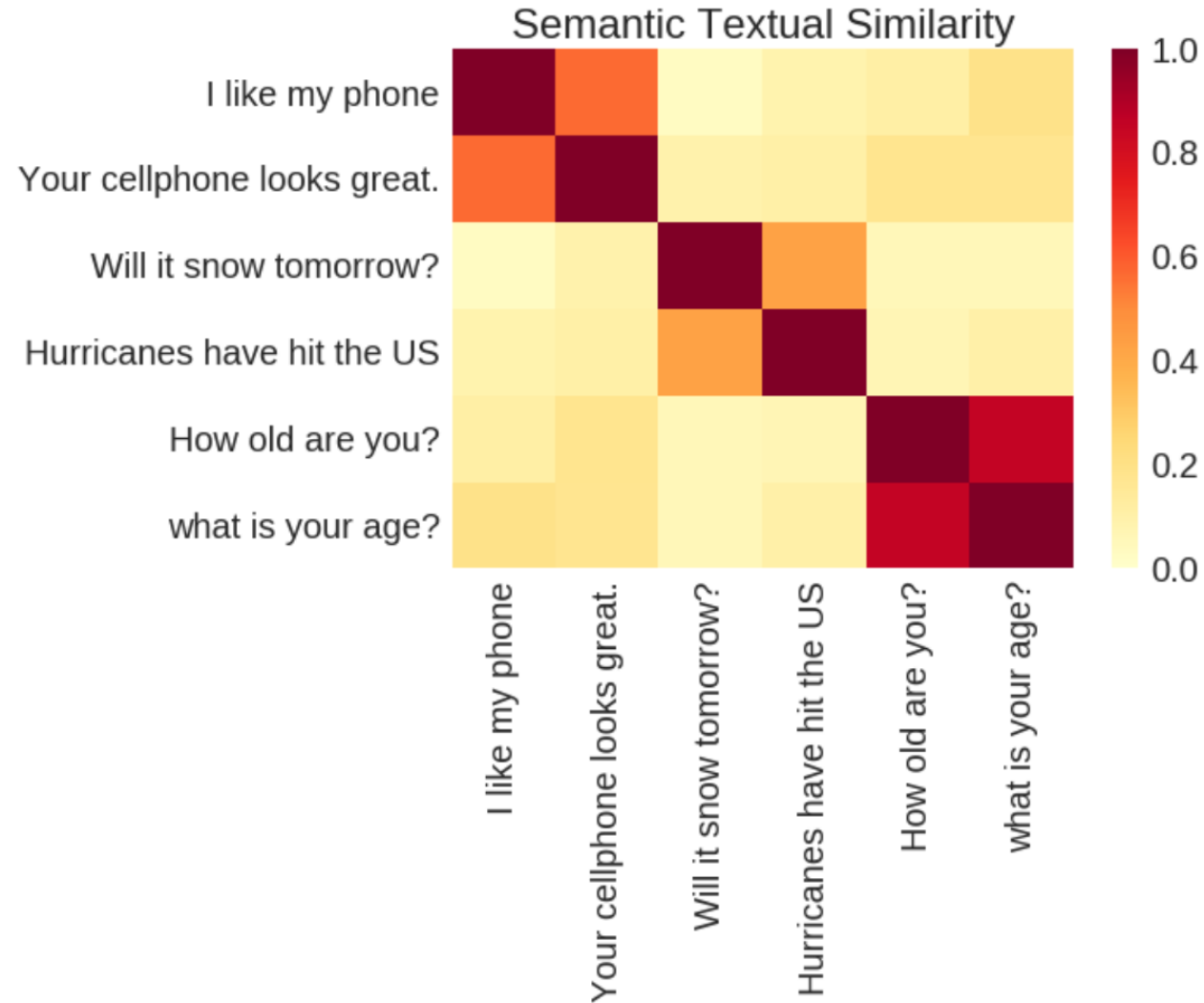
<https://x.ai/>

# LM Arena

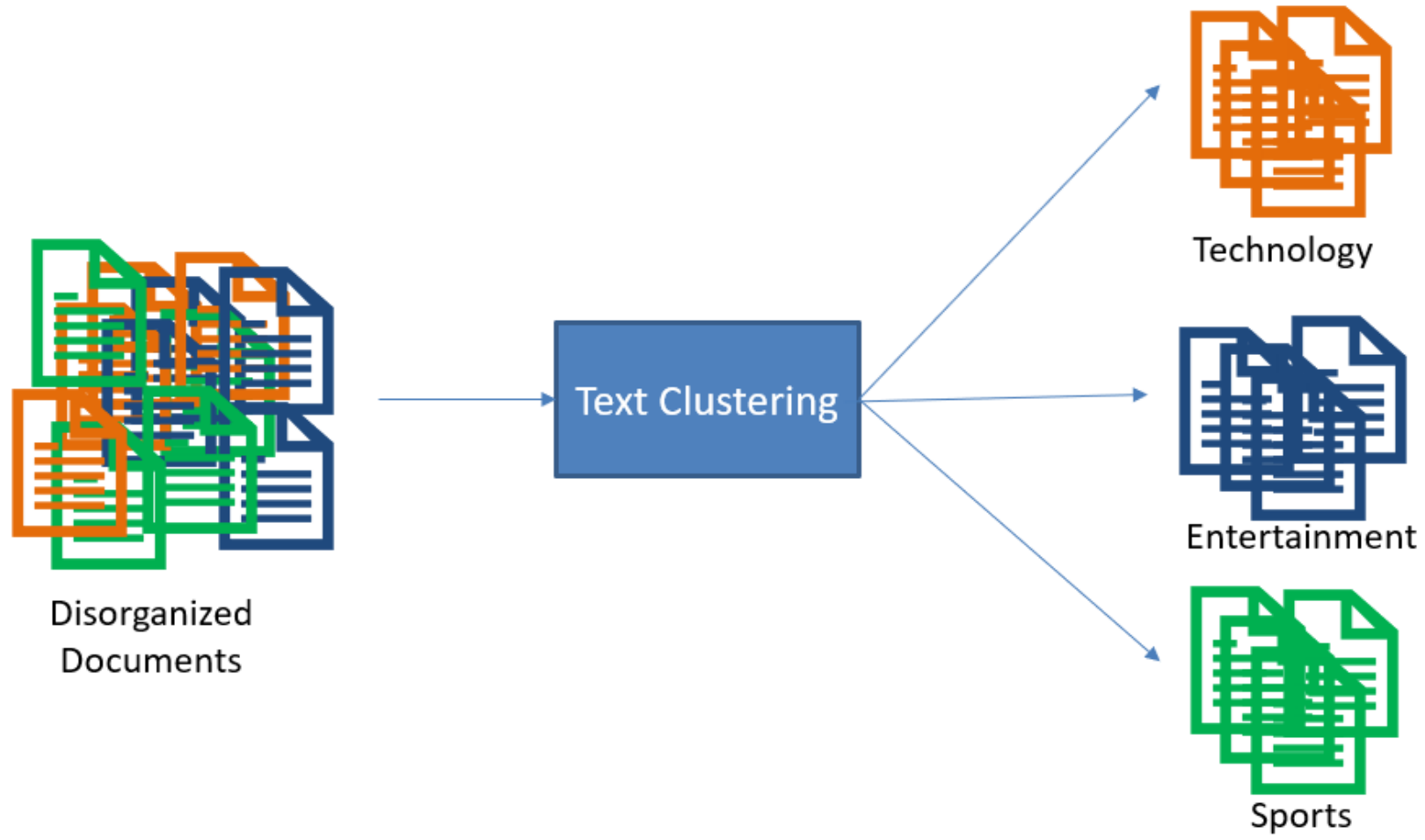
Rank (UB) ↑	Model ↑↓	Score ↑↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	License ↑↓
1	 gemini-2.5-pro	1456	±5	35,405	Google	Proprietary
1	 gpt-5-high	1447	±7	11,405	OpenAI	Proprietary
1	 claude-opus-4-1-20250805-thinking-16k	1447	±7	8,615	Anthropic	Proprietary
2	 o3-2025-04-16	1444	±4	40,935	OpenAI	Proprietary
2	 chatgpt-4o-latest-20250326	1443	±4	36,773	OpenAI	Proprietary
2	 gpt-4.5-preview-2025-02-27	1439	±6	15,271	OpenAI	Proprietary
2	 claude-opus-4-1-20250805	1436	±6	11,548	Anthropic	Proprietary
7	 gpt-5-chat	1426	±7	8,585	OpenAI	Proprietary
8	 grok-4-0709	1422	±6	18,239	xAI	Proprietary
8	 kimi-k2-0711-preview	1421	±5	18,588	Moonshot	Modified MIT

<https://lmarena.ai/>


# Text Similarity



# Document Clustering



# Information Retrieval



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: ...

# Recommendation Systems

Your recently viewed items and featured recommendations

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

Page 1 of 3





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  
★★★★☆ 1,375  
**\$59.99** ✓prime



Echo Look | Hands-Free Camera and Style Assistant with Alexa—includes Style Check to...  
★★★★☆ 413  
**\$99.99** ✓prime



Sonos Beam - Smart TV Sound Bar with Amazon Alexa Built-in - Black  
★★★★☆ 474  
**\$399.00** ✓prime



Echo Wall Clock - see timers at a glance - requires compatible Echo device  
★★★★☆ 1,231  
**\$29.99** ✓prime



Echo Spot Adjustable Stand - Black  
★★★★☆ 933  
**\$19.99** ✓prime



AHASTYLE Wall Mount Hanger Holder ABS for New Dot 3rd Generation Smart Home Speakers...  
★★★★☆ 12  
**\$10.99** ✓prime




Angel Statue Crafted Stand Holder for Amazon Echo Dot 3rd Generation, Alexa Smart...  
★★★★☆ 57  
**\$25.99** ✓prime



Explore more from across the store

Page 1 of 6





Actionable Gamification: Beyond Points, Badges, and Trophies  
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



# Semantic Quality Control

- Paraphrase generation

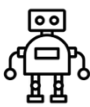
*We will go hiking if tomorrow is a sunny day.*

*If it is sunny tomorrow, we will go hiking.*

- Style transfer
- Plagiarism detection



# Semantic Textual Similarity Benchmark



*A soccer player is kicking the soccer ball into the goal from a long way down the field.*

*A soccer player kicks the ball into the goal.*

3.25

3.94

*Earlier this month, RIM had said it expected to report second-quarter earnings of between 7 cents and 11 cents a share.*

*Excluding legal fees and other charges it expected a loss of between 1 and 4 cents a share.*

1.2

0.5

...

...

...

...

*David Beckham Announces Retirement From Soccer.*

*David Beckham retires from football.*

4.4

3.8

# Pearson's Correlation Coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

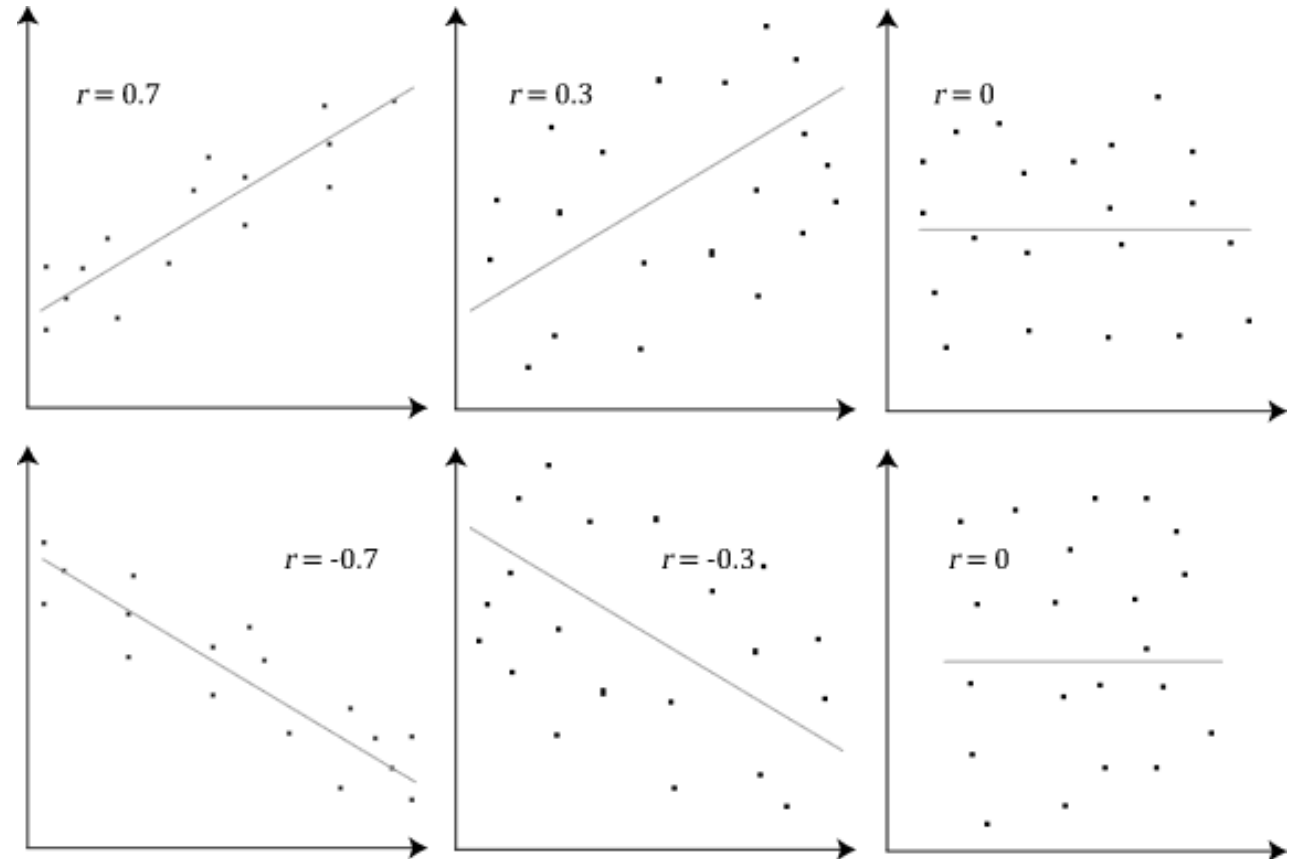
$r$  = correlation coefficient

$x_i$  = values of the x-variable in a sample

$\bar{x}$  = mean of the values of the x-variable

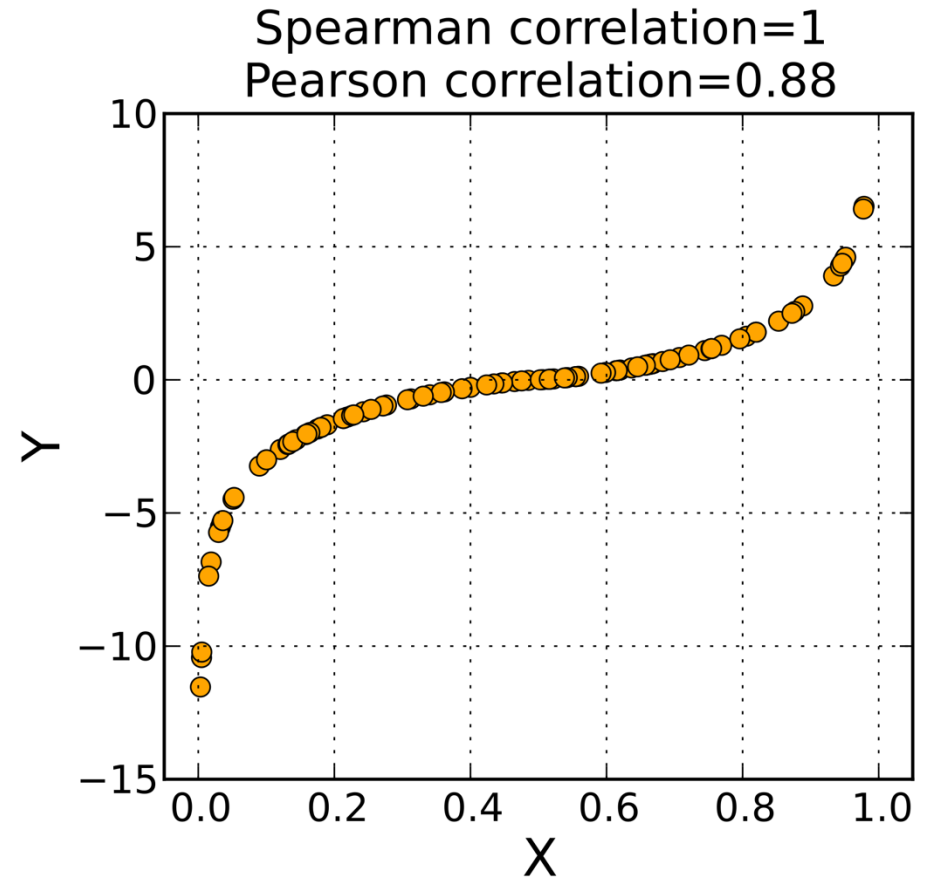
$y_i$  = values of the y-variable in a sample

$\bar{y}$  = mean of the values of the y-variable

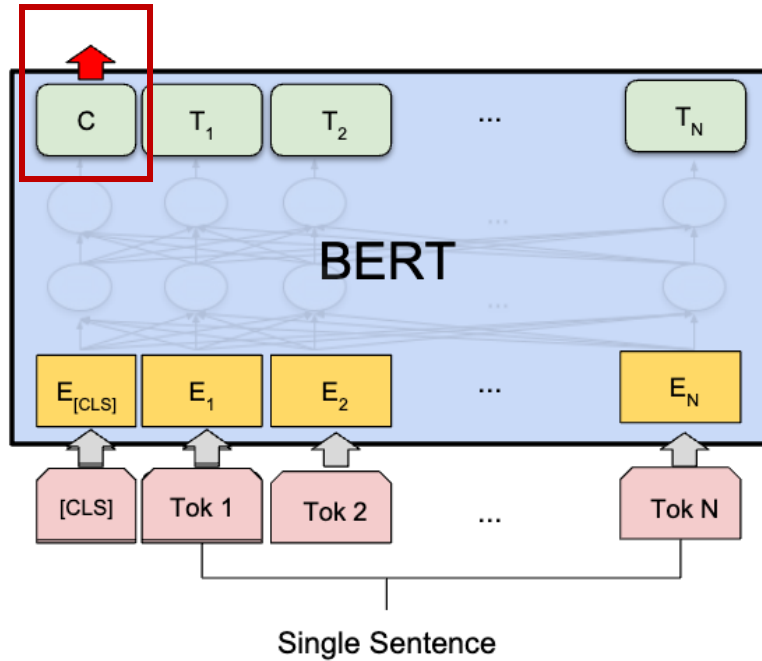


# Spearman's Correlation Coefficient

- Pearson's correlation coefficient on **rank**
- Score
  - Human: [1.2, 3.4, 2.5, 0.7, 4.0]
  - Machine: [0.5, 3.3, 1.0, 1.2, 3.4]
- Rank
  - Human: [4, 2, 3, 5, 1]
  - Machine: [5, 2, 4, 3, 1]
- Assesses monotonic relationships
  - whether linear or not

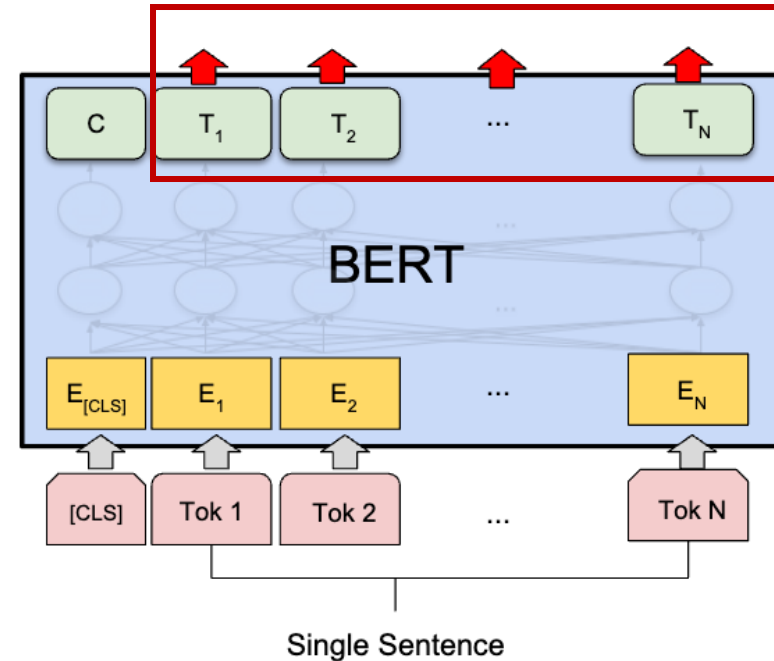


# A Simple Approach: Text Encoder + Cosine Similarity



$$E_1 = \text{Encoder}(S_1)$$

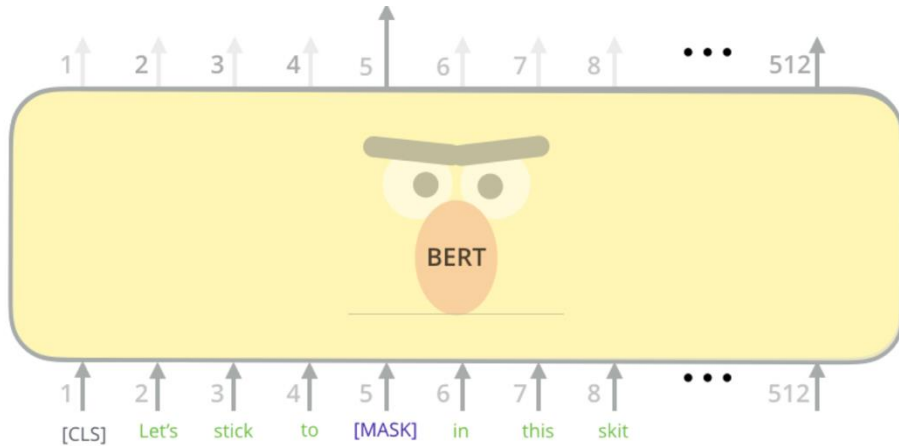
$$E_2 = \text{Encoder}(S_2)$$



$$\text{Similarity}(S_1, S_2) = \frac{E_1 \cdot E_2}{\|E_1\| \|E_2\|}$$

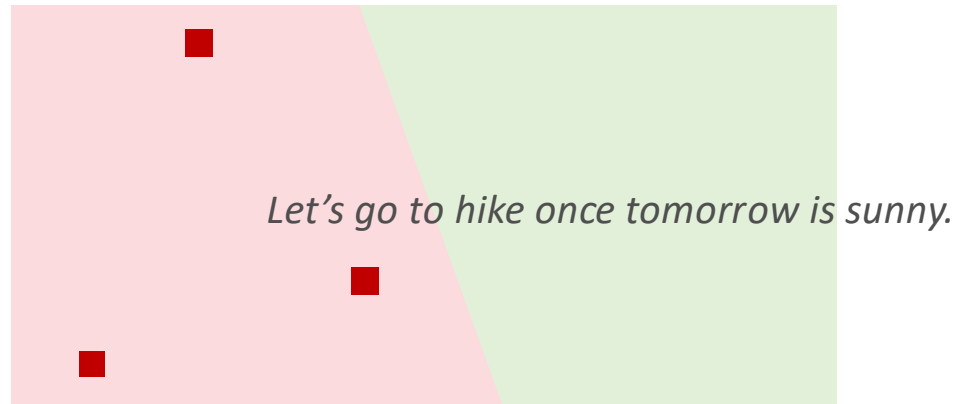
Unfortunately, the performance is bad (why?)

# A Simple Approach: Text Encoder + Cosine Similarity



Pre-trained BERT embeddings are more about lexical information

*If it is sunny tomorrow, we will go hiking.*



Good classification performance  $\neq$  Good similarity

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

# Sentence-BERT

- Consider SNLI dataset
  - Stanford Natural Language Inference

*A boy is jumping on skateboard in the middle of a red bridge.*

*The boy skates down the sidewalk.*

Contradiction

*A boy is jumping on skateboard in the middle of a red bridge.*

*The boy is wearing safety equipment.*

Neutral

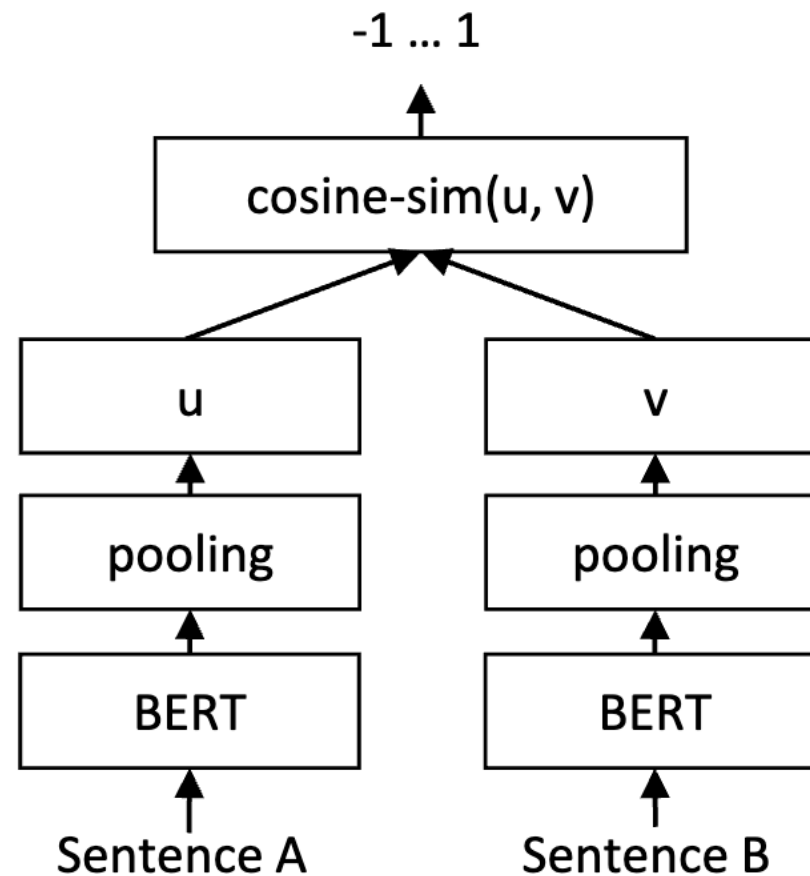
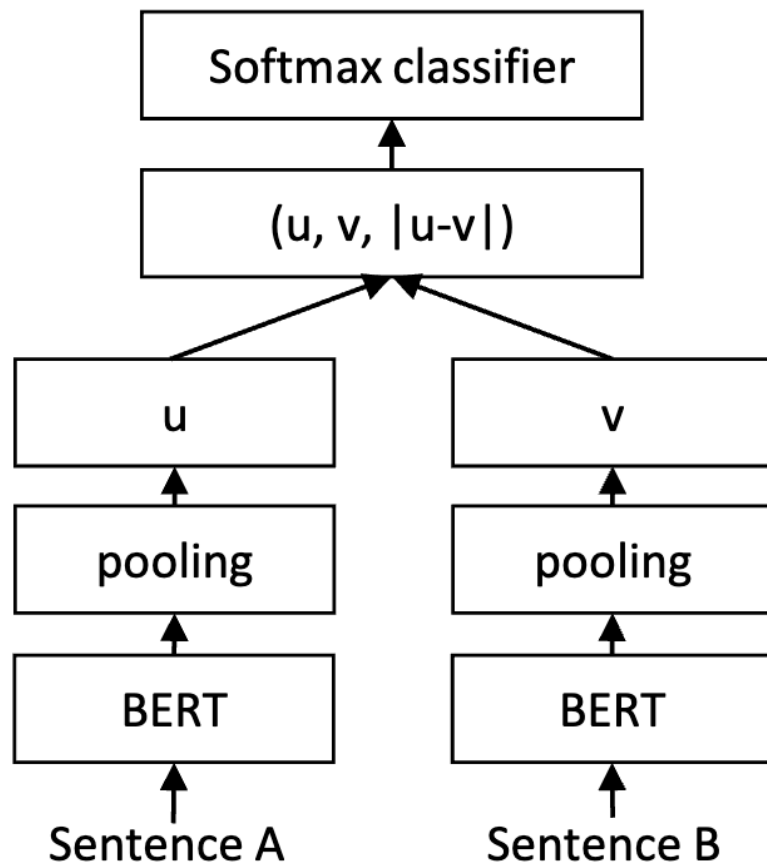
*A boy is jumping on skateboard in the middle of a red bridge.*

*The boy does a skateboarding trick.*

Entailment

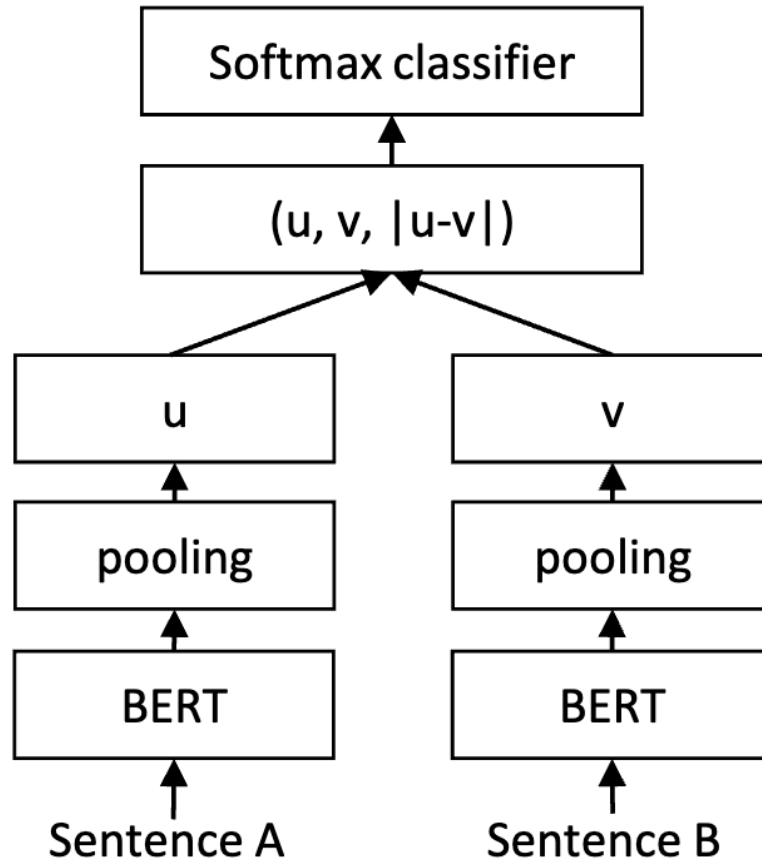
# Sentence-BERT

Contradiction    Neutral    Entailment



# Sentence-BERT

Contradiction    Neutral    Entailment



Cross Entropy Loss

$$o = \text{softmax}(W_t(u, v, |u - v|))$$

Triplet Loss

$$\max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$$

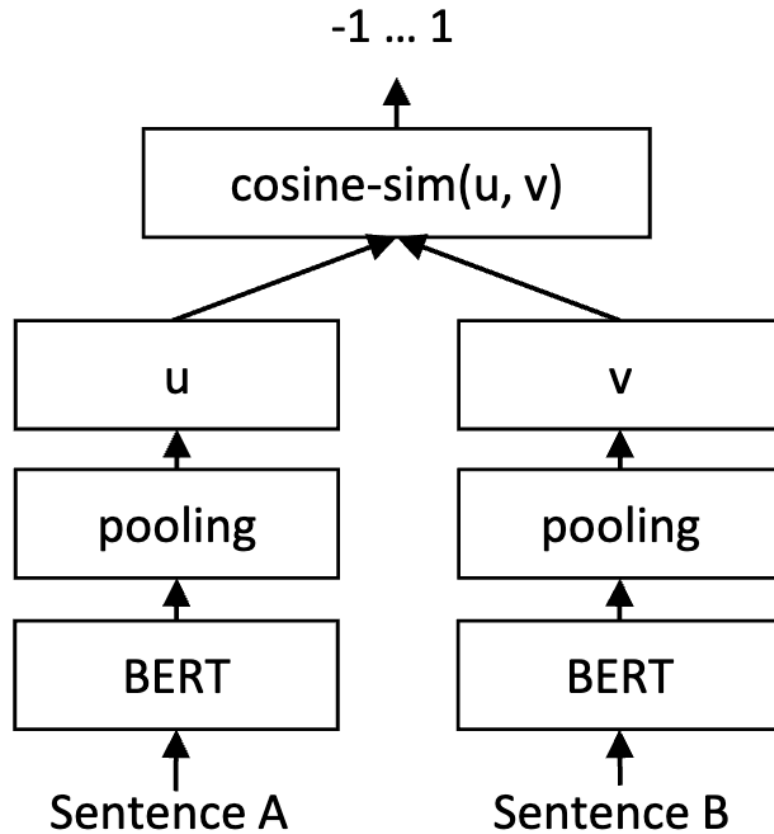


# Sentence-BERT: Performance

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	<b>76.69</b>	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	<b>78.46</b>	<b>74.90</b>	80.99	76.25	<b>79.23</b>	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	<b>74.53</b>	77.00	73.18	<b>81.85</b>	<b>76.82</b>	79.10	74.29	<b>76.68</b>

# SimCSE

- Simple Contrastive Learning of Sentence Embeddings



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

# Contrastive Learning

*Sentence 1A*

*Sentence 1B*

*Sentence 2A*

*Sentence 2B*

*Sentence 3A*

*Sentence 3B*

*Sentence 4A*

*Sentence 4B*

*Sentence 5A*

*Sentence 5B*

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+) / \tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+) / \tau}}$$

# Contrastive Learning

Sentence 1A	Sentence 1B
Sentence 2A	Sentence 2B
Sentence 3A	Sentence 3B
Sentence 4A	Sentence 4B
Sentence 5A	Sentence 5B

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

# Contrastive Learning

*Sentence 1A*

*Sentence 1B*

*Sentence 2A*

*Sentence 2B*

*Sentence 3A*

*Sentence 3B*

*Sentence 4A*

*Sentence 4B*

*Sentence 5A*

*Sentence 5B*

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+) / \tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+) / \tau}}$$

# Contrastive Learning

Sentence 1A	Sentence 1B
Sentence 2A	Sentence 2B
Sentence 3A	Sentence 3B
Sentence 4A	Sentence 4B
Sentence 5A	Sentence 5B

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

# Unsupervised Contrastive Learning

*Sentence 1*

*Sentence 1'*

*Sentence 2*

*Sentence 3*

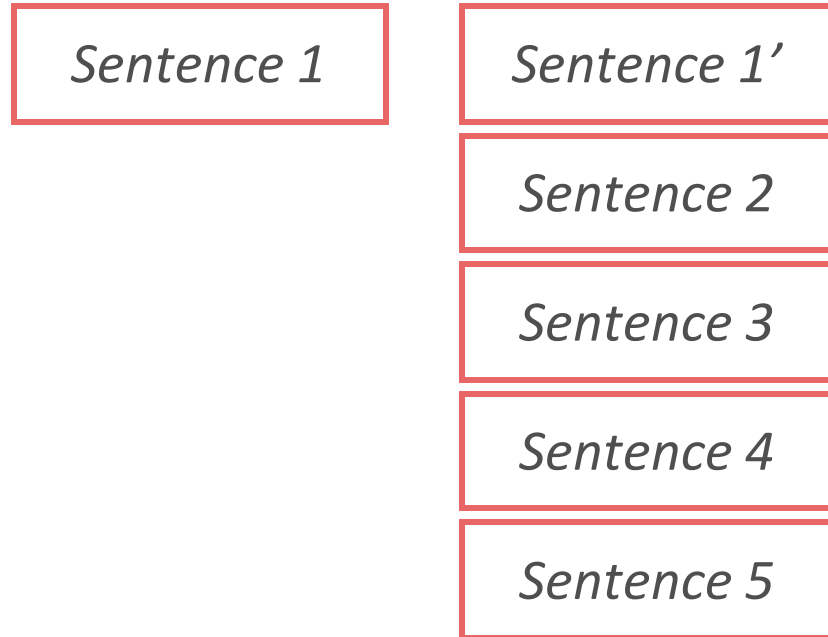
*Sentence 4*

*Sentence 5*

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

# Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

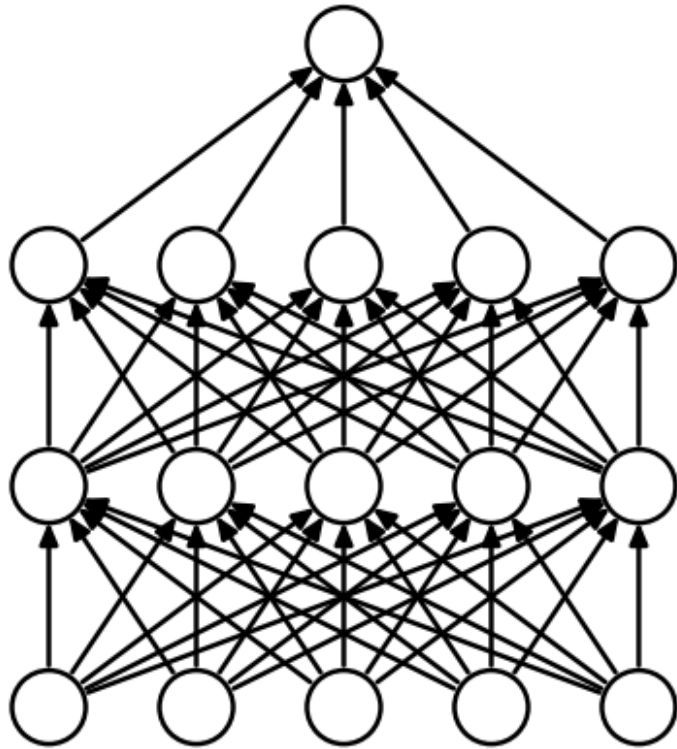
Generate positive example with masking

*If it is sunny tomorrow, we will go hiking.*

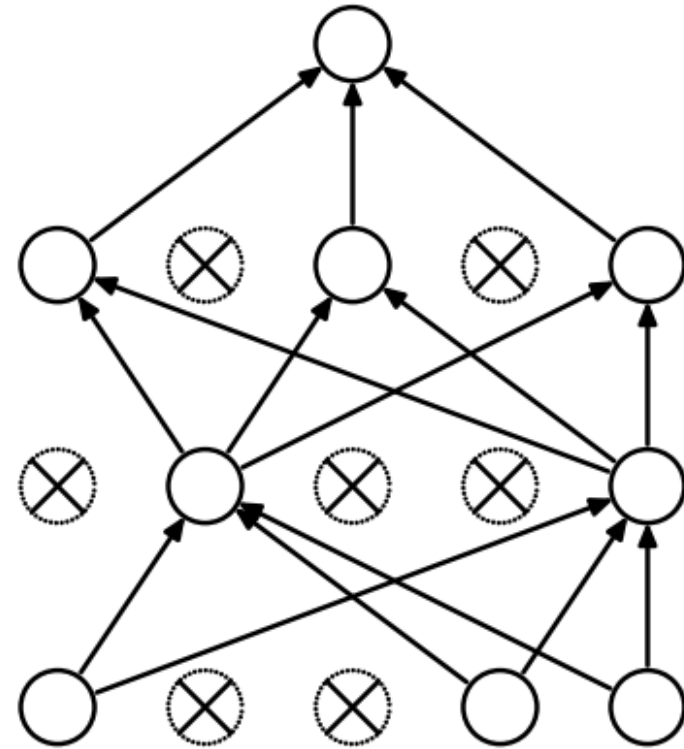
*If [mask] is sunny tomorrow, we [mask] go hiking.*



# Dropout



(a) Standard Neural Net



(b) After applying dropout.

Generate positive example with neuron masking

# Unsupervised Contrastive Learning

*Sentence 1*

*Sentence 1'*

*Sentence 2*

*Sentence 3*

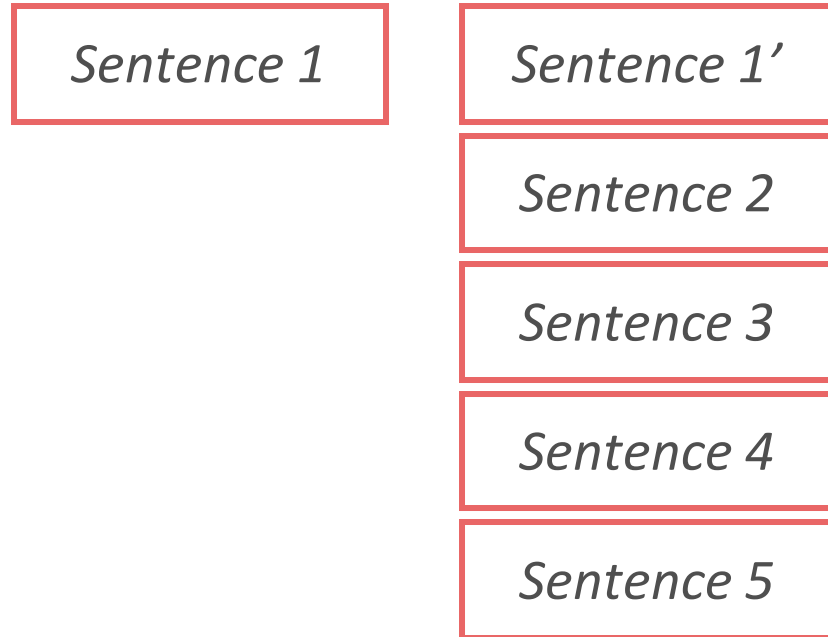
*Sentence 4*

*Sentence 5*

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

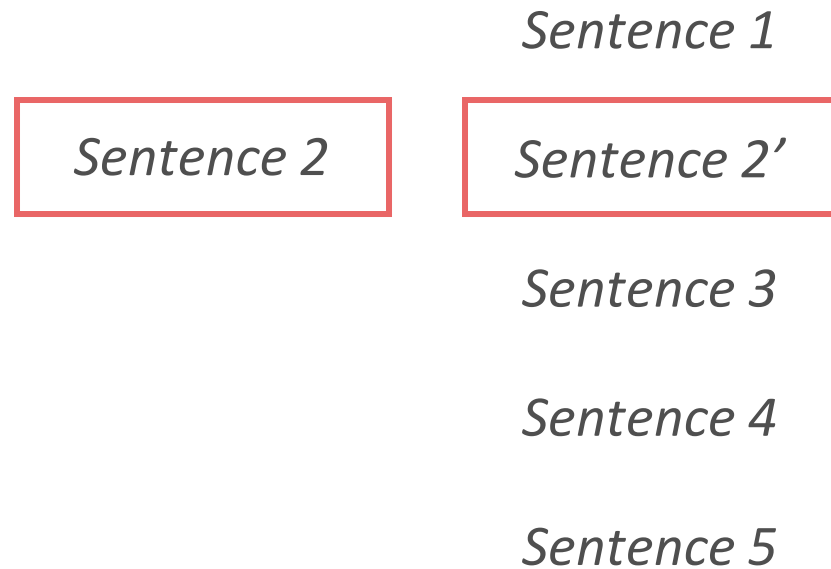
# Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

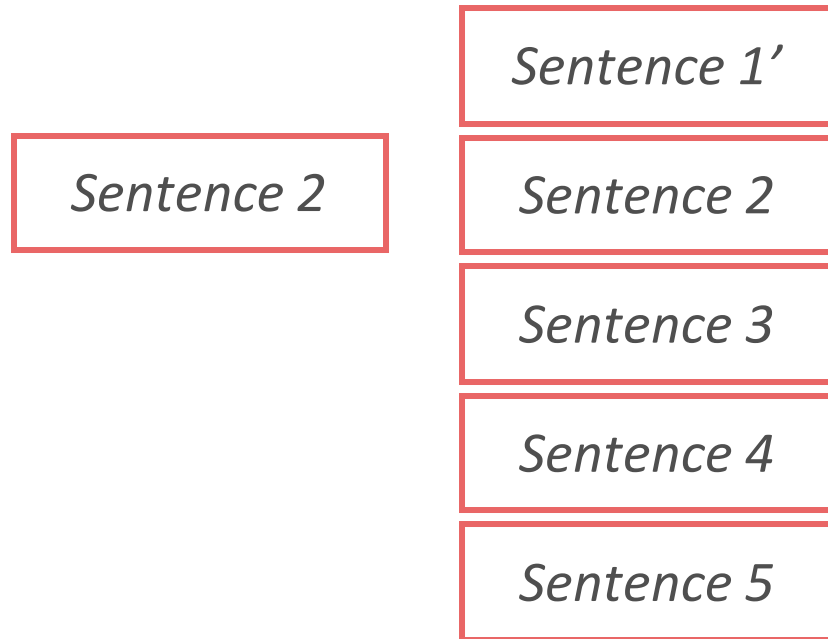
# Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

# Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}$$

# SimCSE: Performance

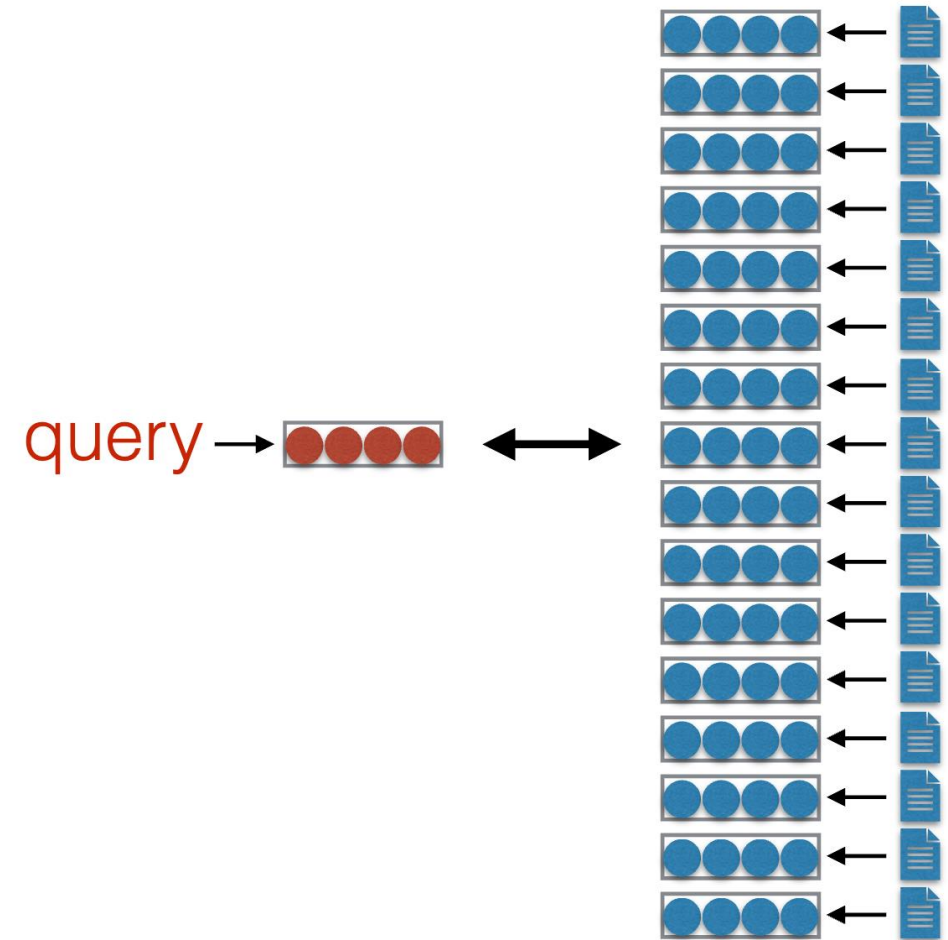
Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
<i>Unsupervised models</i>								
GloVe embeddings (avg.) <sup>♣</sup>	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT <sub>base</sub> (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT <sub>base</sub> -flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT <sub>base</sub> -whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT <sub>base</sub> <sup>♡</sup>	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT <sub>base</sub>	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
* SimCSE-BERT <sub>base</sub>	<b>68.40</b>	<b>82.41</b>	<b>74.38</b>	<b>80.91</b>	<b>78.56</b>	<b>76.85</b>	<b>72.23</b>	<b>76.25</b>
RoBERTa <sub>base</sub> (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
RoBERTa <sub>base</sub> -whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-RoBERTa <sub>base</sub>	52.41	75.19	65.52	77.12	78.63	72.41	<b>68.62</b>	69.99
* SimCSE-RoBERTa <sub>base</sub>	<b>70.16</b>	<b>81.77</b>	<b>73.24</b>	<b>81.36</b>	<b>80.65</b>	<b>80.22</b>	68.56	<b>76.57</b>
* SimCSE-RoBERTa <sub>large</sub>	<b>72.86</b>	<b>83.99</b>	<b>75.62</b>	<b>84.77</b>	<b>81.80</b>	<b>81.98</b>	<b>71.26</b>	<b>78.90</b>
<i>Supervised models</i>								
InferSent-GloVe <sup>♣</sup>	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder <sup>♣</sup>	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT <sub>base</sub> <sup>♣</sup>	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT <sub>base</sub> -flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT <sub>base</sub> -whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT <sub>base</sub>	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
* SimCSE-BERT <sub>base</sub>	<b>75.30</b>	<b>84.67</b>	<b>80.19</b>	<b>85.40</b>	<b>80.82</b>	<b>84.25</b>	<b>80.39</b>	<b>81.57</b>
SRoBERTa <sub>base</sub> <sup>♣</sup>	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa <sub>base</sub> -whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
* SimCSE-RoBERTa <sub>base</sub>	<b>76.53</b>	<b>85.21</b>	<b>80.95</b>	<b>86.03</b>	<b>82.57</b>	<b>85.83</b>	<b>80.50</b>	<b>82.52</b>
* SimCSE-RoBERTa <sub>large</sub>	<b>77.46</b>	<b>87.27</b>	<b>82.36</b>	<b>86.66</b>	<b>83.93</b>	<b>86.70</b>	<b>81.95</b>	<b>83.76</b>

# Dense Passage Retrieval

Similarity between query and documents

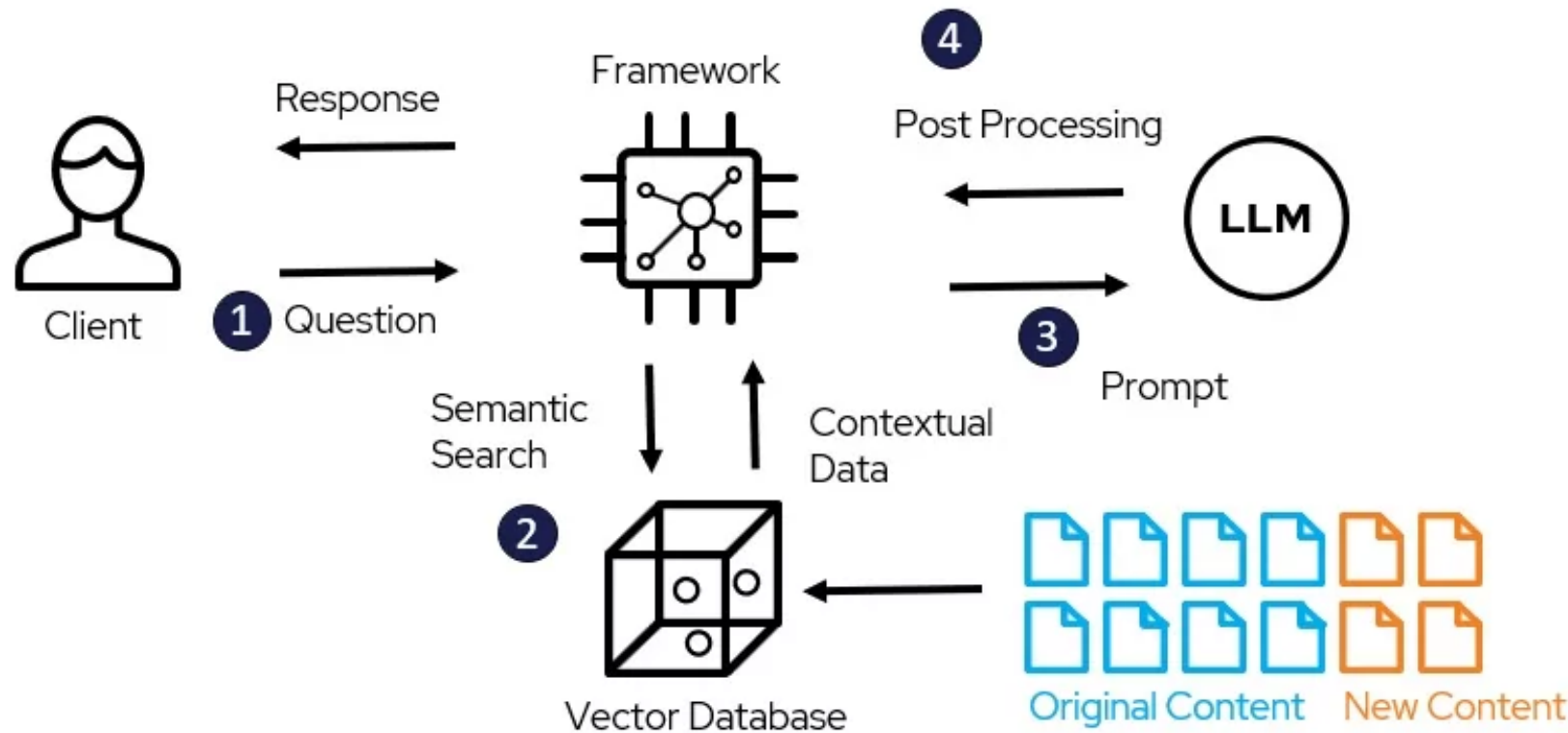
Similarity between two sentences

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



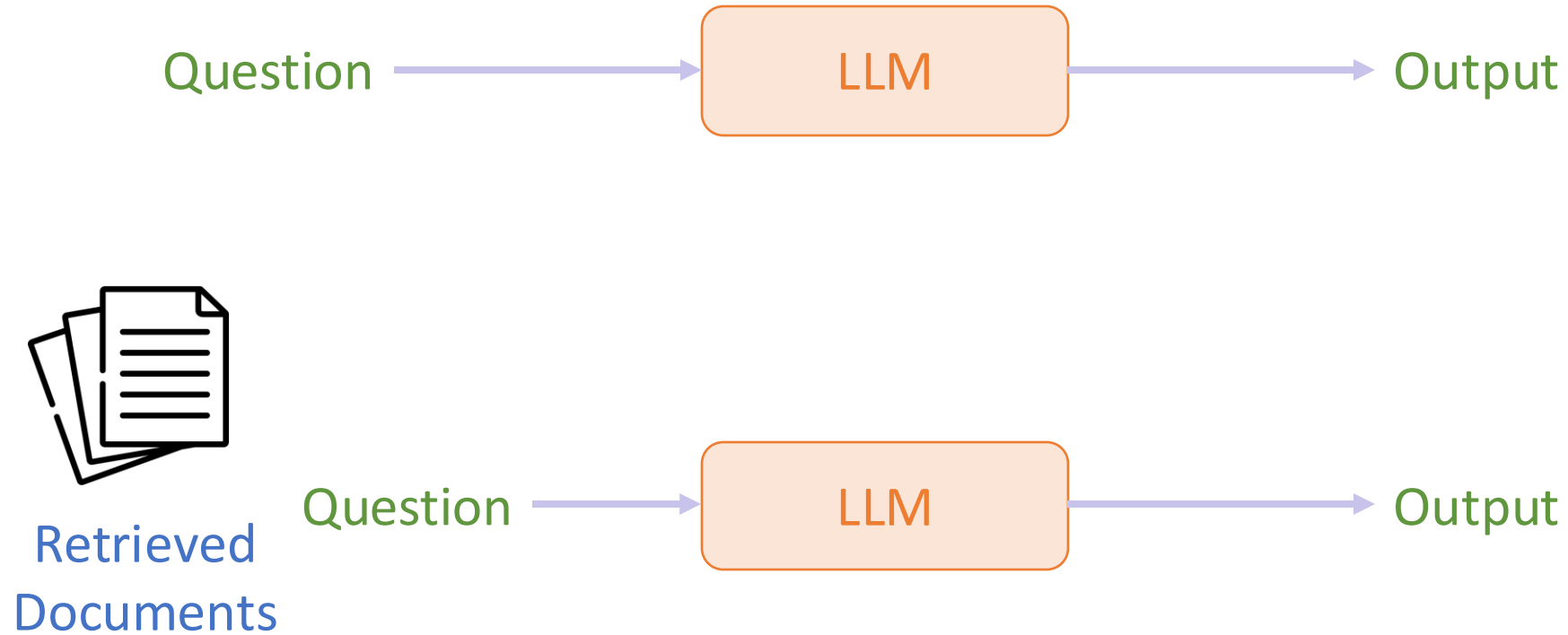
# Retrieval-Augmented Generation (RAG)

RAG Architecture Model





# Retrieval-Augmented Generation (RAG)



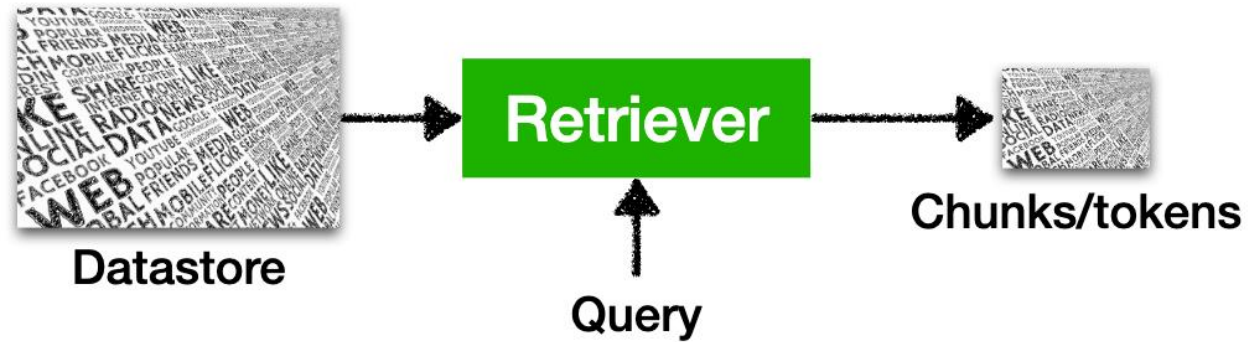
# Retrieval-Augmented Generation (RAG)

**Retrieval models** and **language models** are trained **independently**

- Training language models

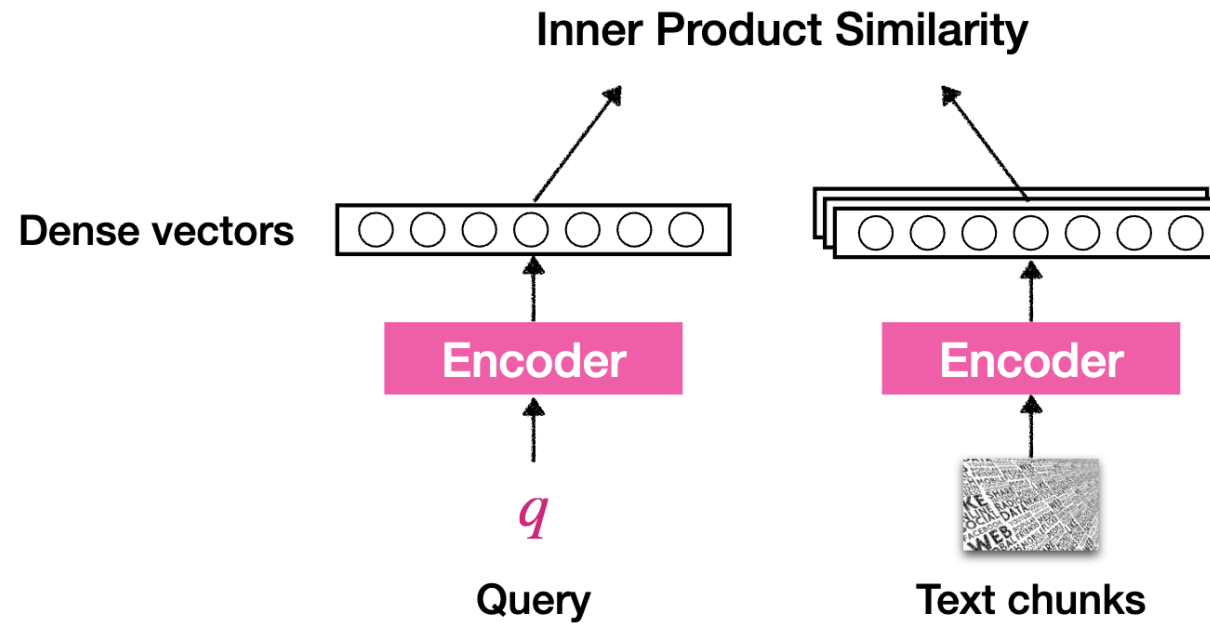


- Training retrieval models

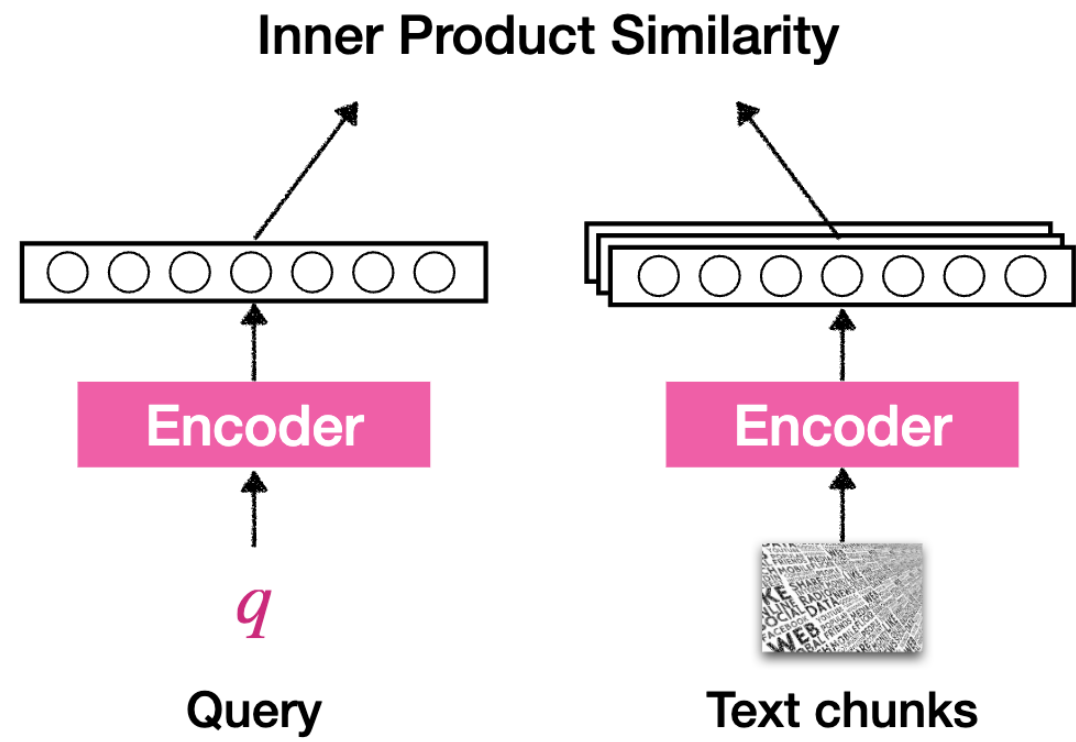


# How to Train A Retriever?

## Dense retrieval models: DPR (Karpukhin et al. 2020)



# How to Train A Retriever?

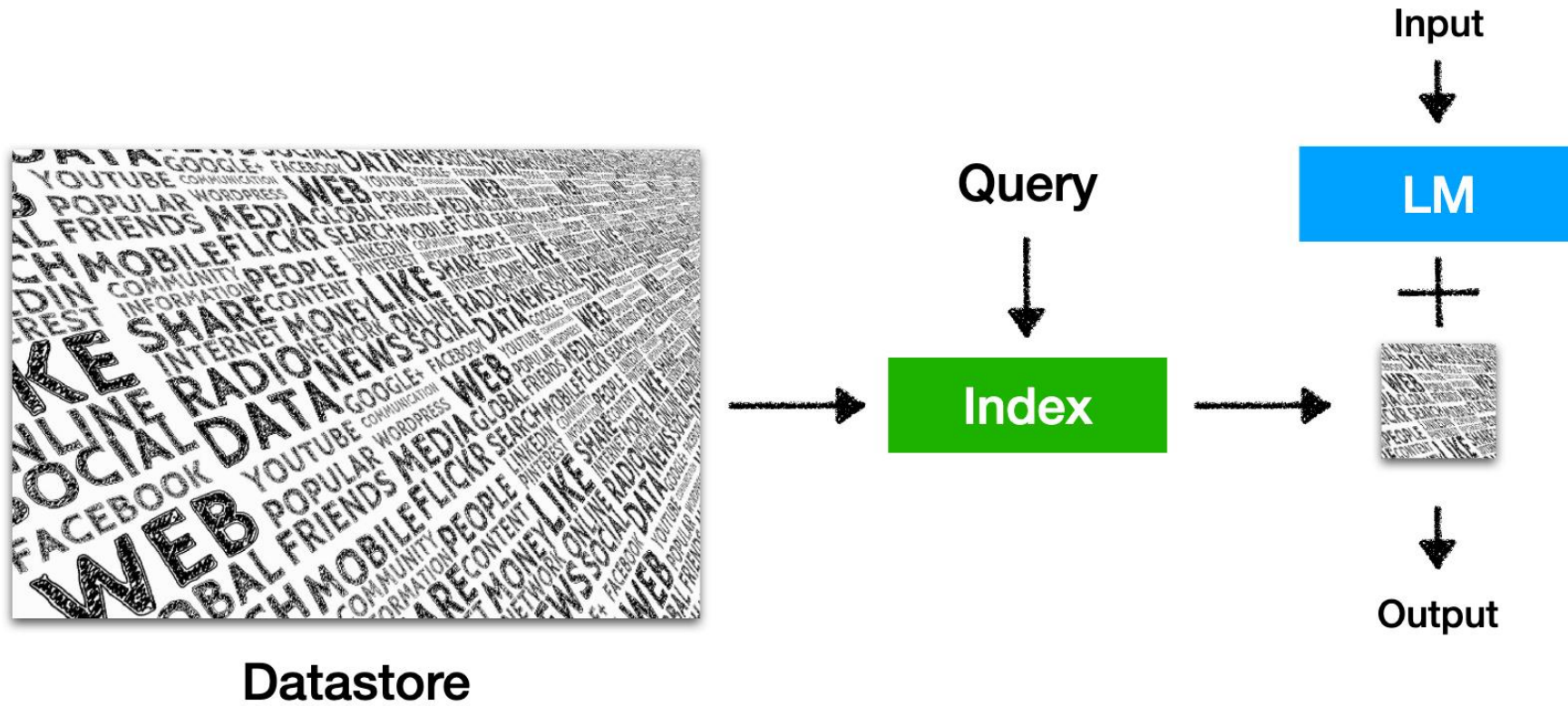


$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-)$$
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Contrastive learning



# Retrieval-Augmented Generation (RAG)



# Why RAG?

- LLMs can't memorize all (long-tail) knowledge in their parameters



List 5 important papers authored by Geoffrey Hinton



Geoffrey Hinton is a renowned computer scientist ...  
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, .. ✓
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, .. ✓
- ...
4. "**Deep Learning**" (with Y. Bengio and A. Courville) - Published as a book in **2016**,... ✗
5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,... ✗

# Why RAG?

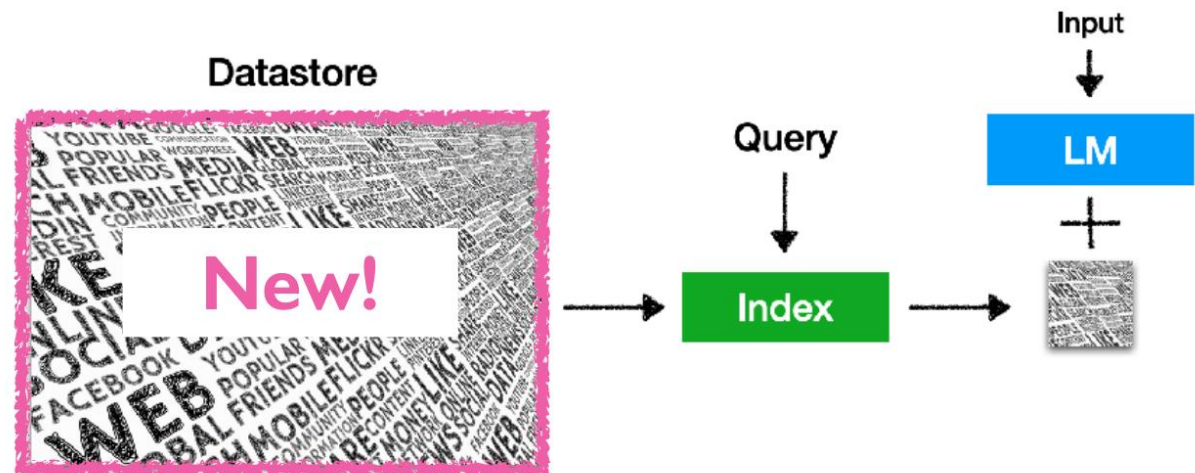
- LLMs' knowledge is easily outdated and hard to update



# Who is the CEO of Twitter?



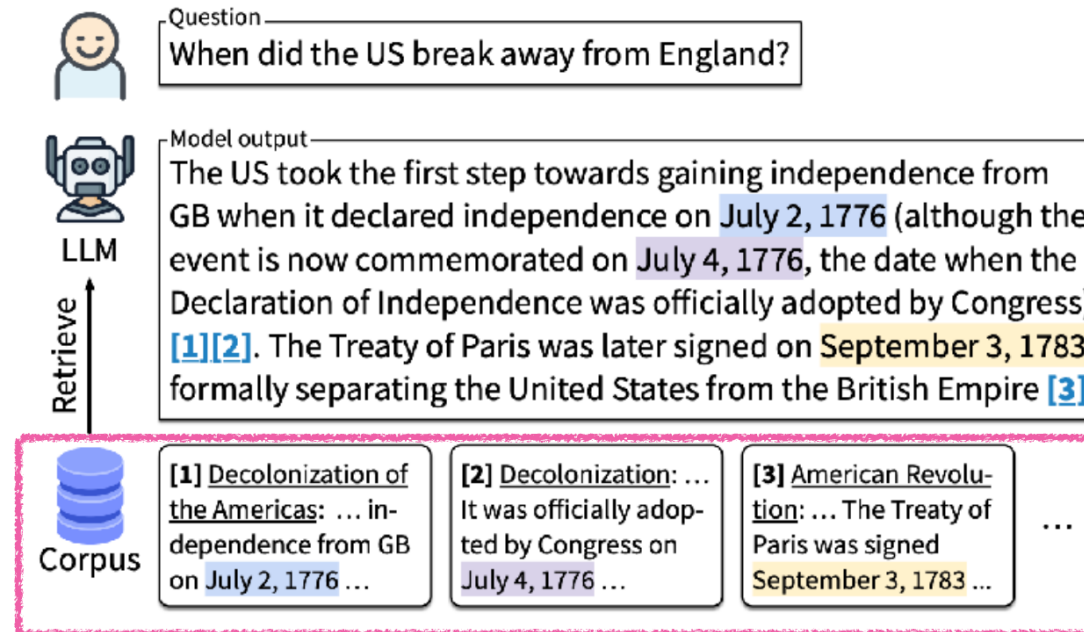
As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....



# Why RAG?

- LLMs' output is challenging to interpret and verify

## Generating text with citations



(Nakano et al. 2021; Menick et al., 2022; Gao et al., 2023)



# Why RAG?

- LLMs are shown to easily leak private training data

### Extracting Training Data from Large Language Models

Nicholas Carlini <sup>1</sup>	Florian Tramèr <sup>2</sup>	Eric Wallace <sup>3</sup>	Matthew Jagielski <sup>4</sup>
Ariel Herbert-Voss <sup>5,6</sup>	Katherine Lee <sup>1</sup>	Adam Roberts <sup>1</sup>	Tom Brown <sup>5</sup>
Dawn Song <sup>3</sup>	Úlfar Erlingsson <sup>7</sup>	Alina Oprea <sup>4</sup>	Colin Raffel <sup>1</sup>
<sup>1</sup> Google <sup>2</sup> Stanford <sup>3</sup> UC Berkeley <sup>4</sup> Northeastern University <sup>5</sup> OpenAI <sup>6</sup> Harvard <sup>7</sup> Apple			

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
<b>Named individuals (non-news samples only)</b>	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
<b>Contact info (address, email, phone, twitter, etc.)</b>	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

# Vision + Language

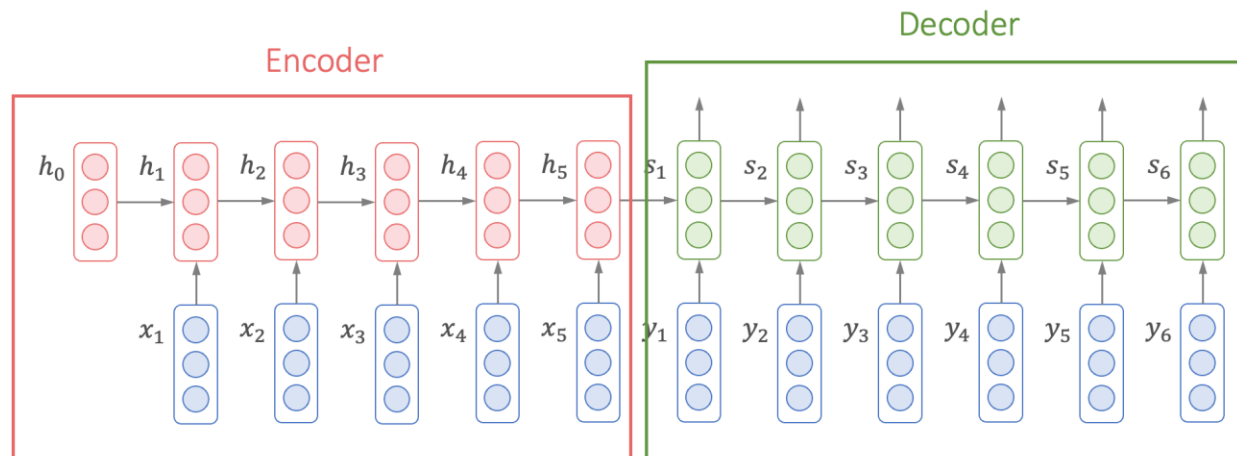
- Image captioning

<p>A young boy is playing basketball.</p> 	<p>Two dogs play in the grass.</p> 	<p>A dog swims in the water.</p> 	<p>A little girl in a pink shirt is swinging.</p> 
<p>A group of people walking down a street.</p> 	<p>A group of women dressed in formal attire.</p> 	<p>Two children play in the water.</p> 	<p>A dog jumps over a hurdle.</p> 

# Image Captioning with Encoder-Decoder Models



Replace the text encoder  
as an image encoder



Encoder-Decoder Model

# Recap: Convolutional Neural Network (For Text)

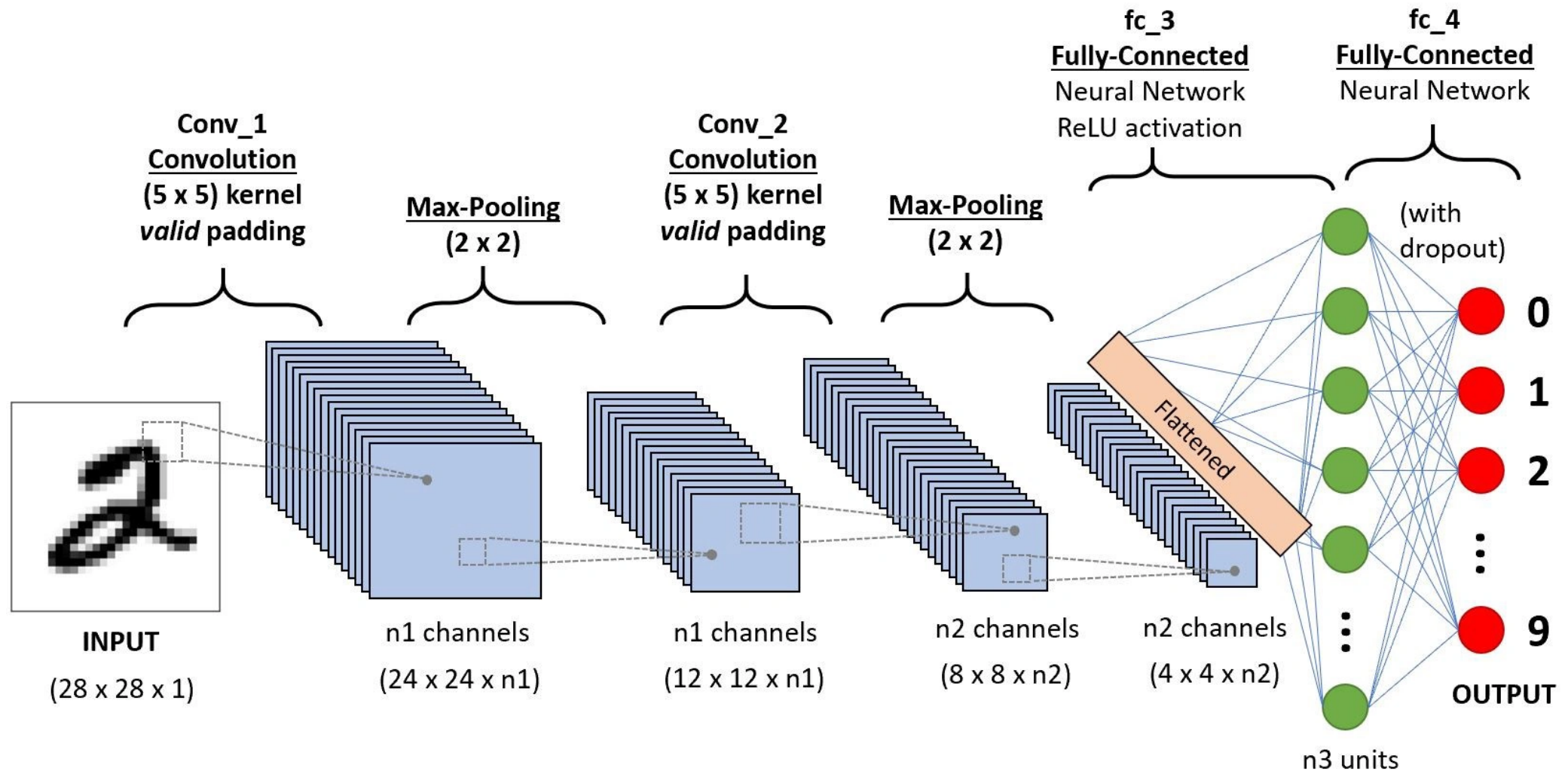
Learnable Weight (Filter)  
Filter Size = 3

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ \dots & \dots & \dots \\ w_{4,1} & w_{4,2} & w_{4,3} \end{bmatrix}$$

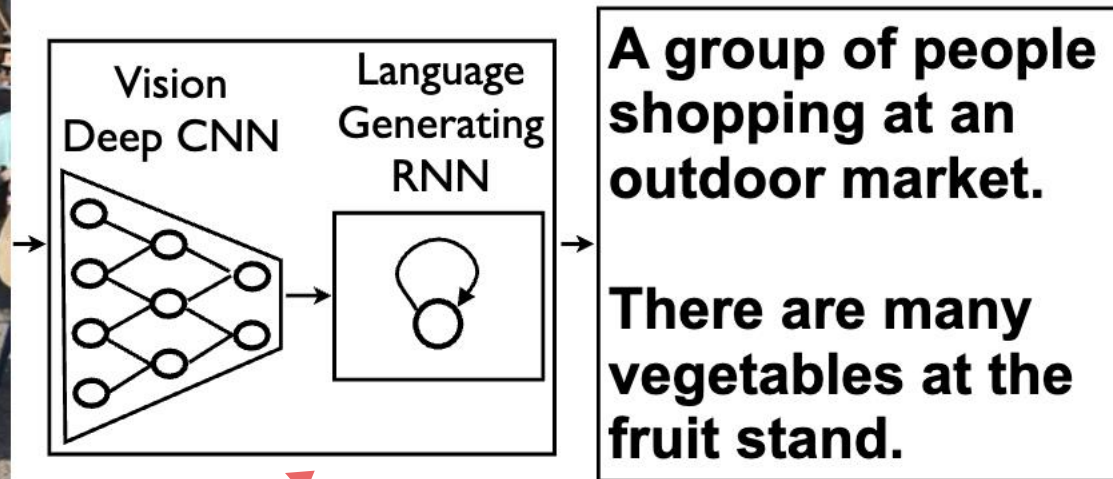




# Convolutional Neural Network (For Image)

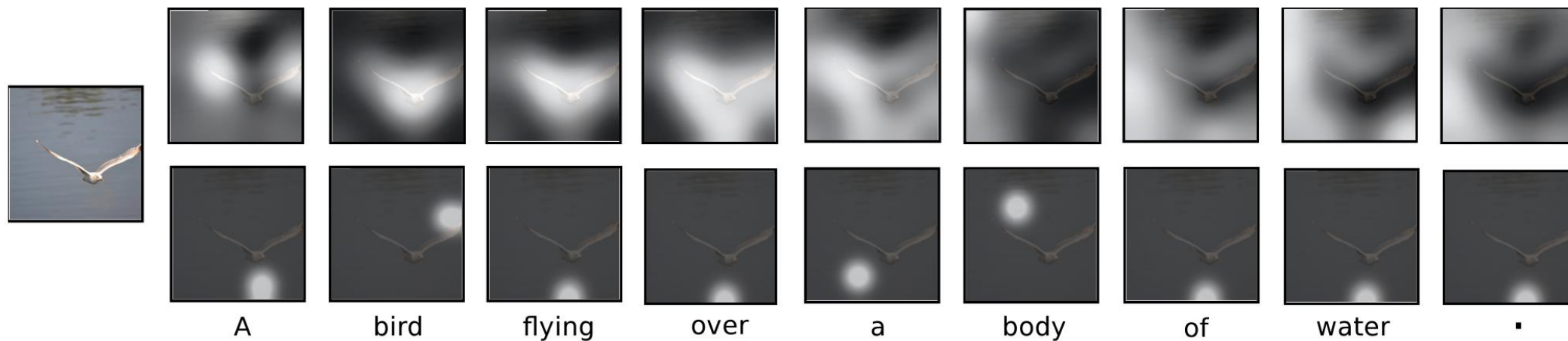


# Encoder-Decoder: CNN-RNN

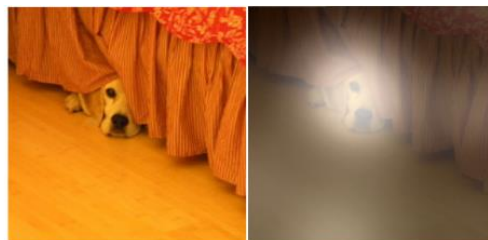


Text embedding space and image embedding space can be aligned!

# CNN + Attention LSTM



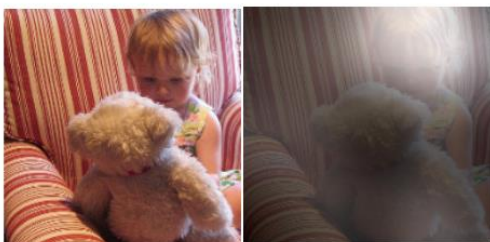
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



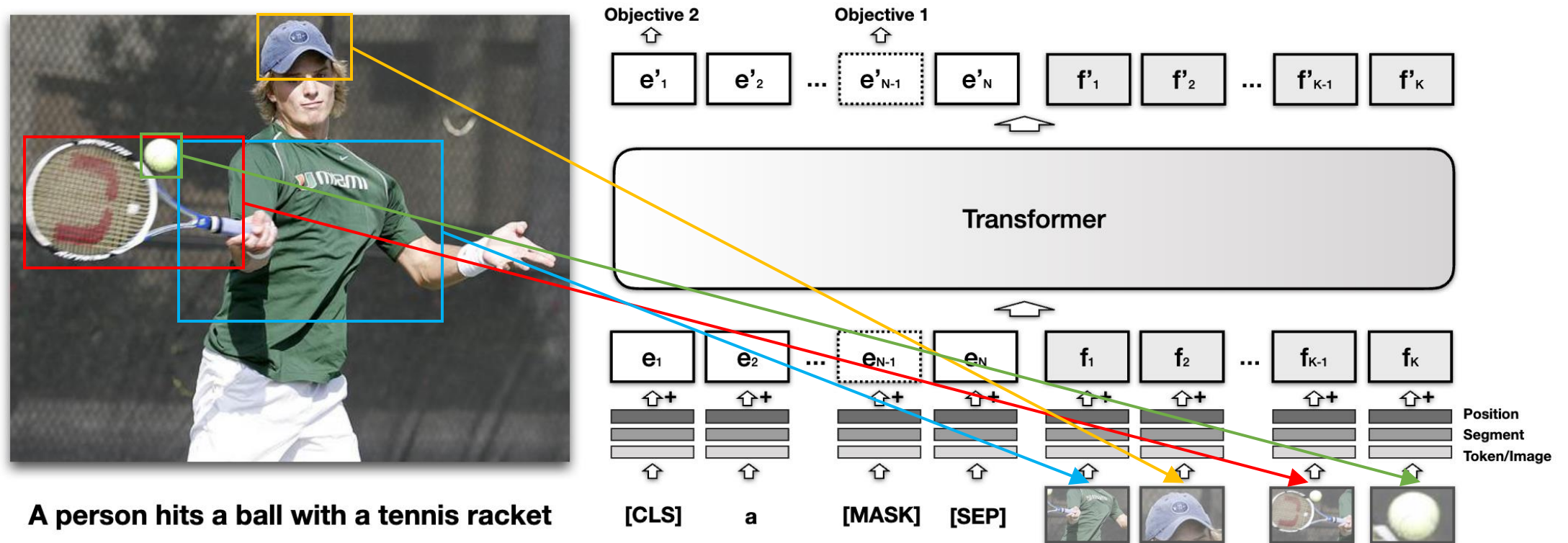
A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.



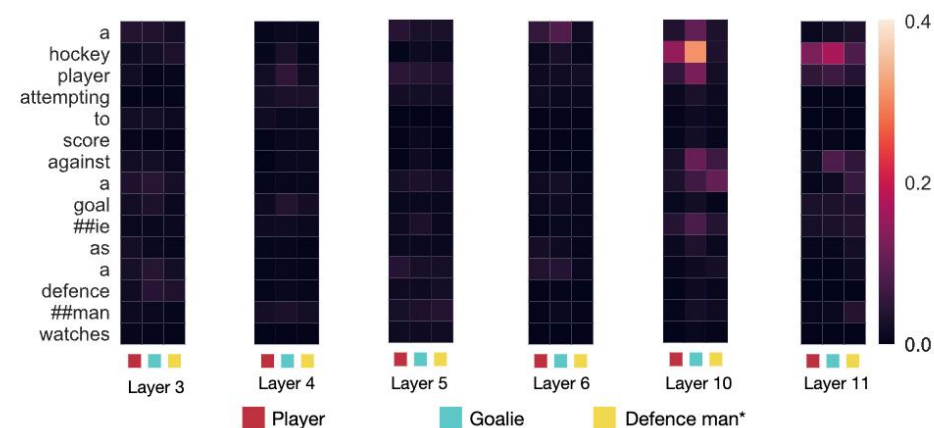
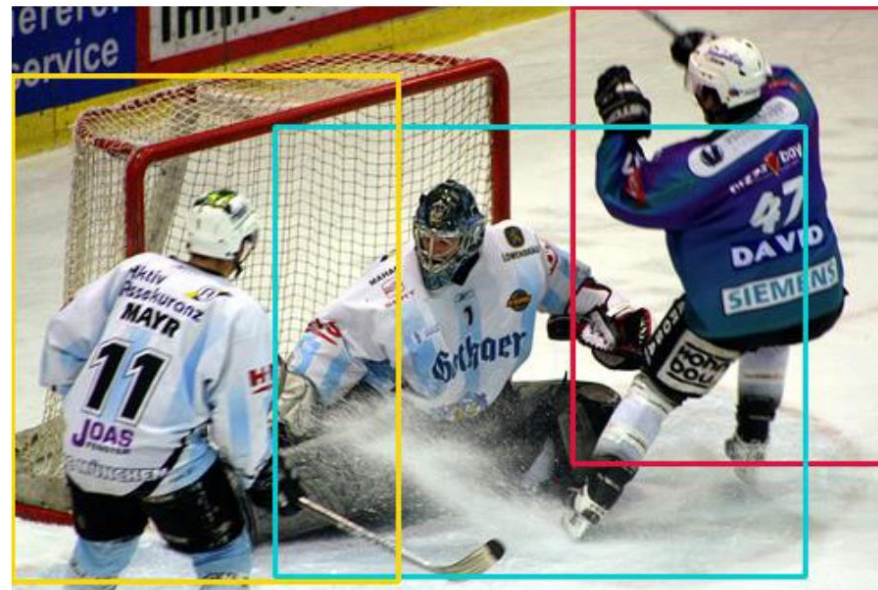
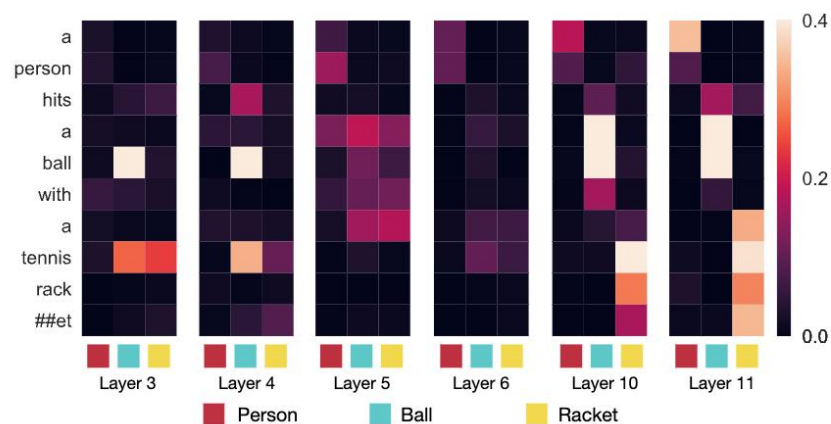
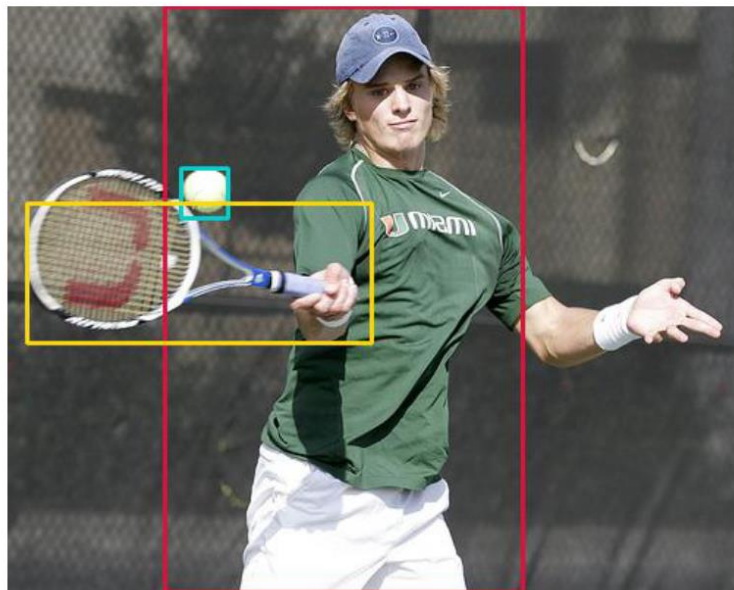
# Joint Visual and Textual Embeddings: VisualBERT



Require an object detection model



# Joint Visual and Textual Embeddings: VisualBERT



# Visual Question Answering

Who is wearing glasses?

man



woman



Where is the child sitting?

fridge



arms



Is the umbrella upside down?

yes



no



How many children are in the bed?

2

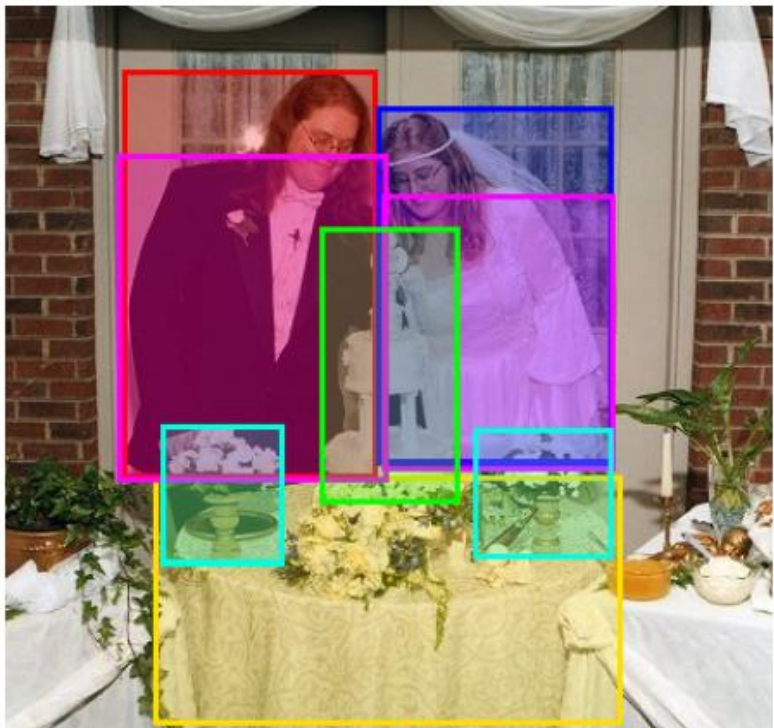


1



Model	Test-Dev	Test-Std
Pythia v0.1 (Jiang et al., 2018)	68.49	-
Pythia v0.3 (Singh et al., 2019)	68.71	-
VisualBERT w/o Early Fusion	68.18	-
VisualBERT w/o COCO Pre-training	70.18	-
VisualBERT	70.80	71.00
Pythia v0.1 + VG + Other Data Augmentation (Jiang et al., 2018)	70.01	70.24
MCAN + VG (Yu et al., 2019b)	70.63	70.90
MCAN + VG + Multiple Detectors (Yu et al., 2019b)	72.55	-
MCAN + VG + Multiple Detectors + BERT (Yu et al., 2019b)	72.80	-
MCAN + VG + Multiple Detectors + BERT + Ensemble (Yu et al., 2019b)	75.00	75.23

# Language Grounding

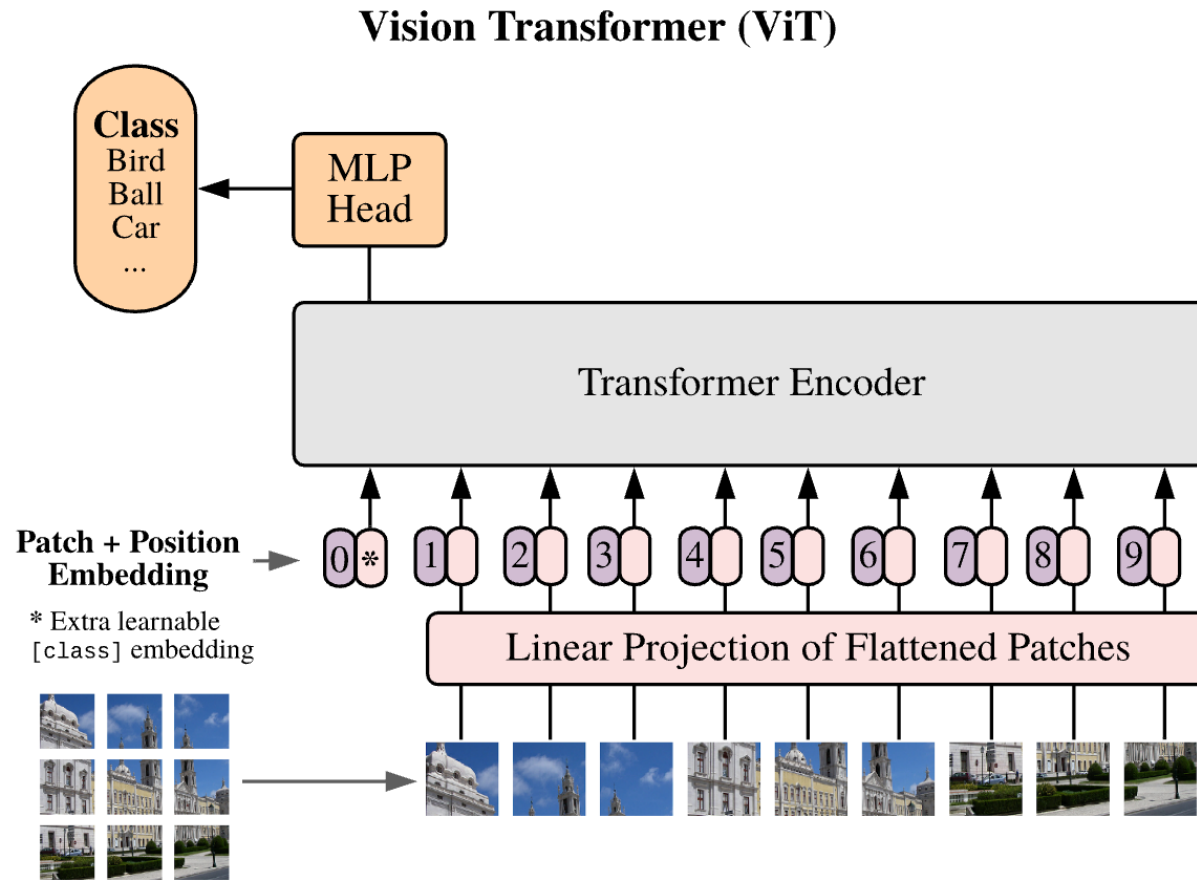


A couple in **their wedding attire** stand behind **a table** with **a wedding cake** and **flowers**.  
**A bride** and **groom** are standing in front of **their wedding cake** at their reception.  
**A bride** and **groom** smile as **they** view **their wedding cake** at a reception.  
**A couple** stands behind **their wedding cake**.  
**Man** and **woman** cutting **wedding cake**.

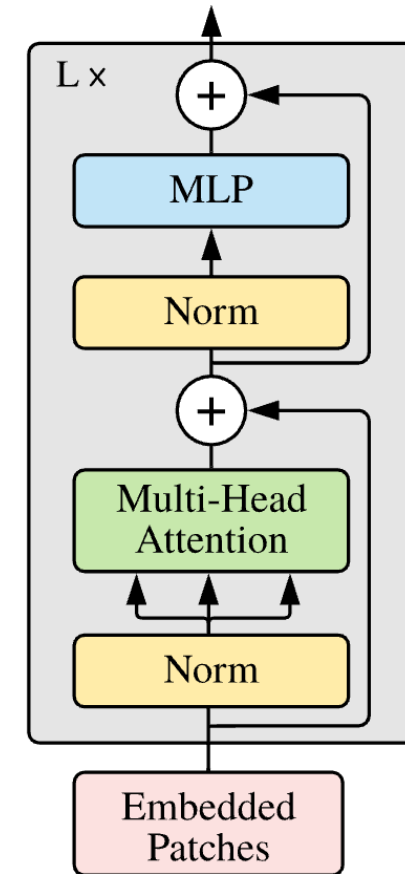
Model	R@1		R@5		R@10		Upper Bound	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
BAN (Kim et al., 2018)	-	69.69	-	84.22	-	86.35	86.97	87.45
VisualBERT w/o Early Fusion	70.33	-	84.53	-	86.39	-	86.97	87.45
VisualBERT w/o COCO Pre-training	68.07	-	83.98	-	86.24	-		
VisualBERT	70.40	71.33	84.49	84.98	86.31	86.51		



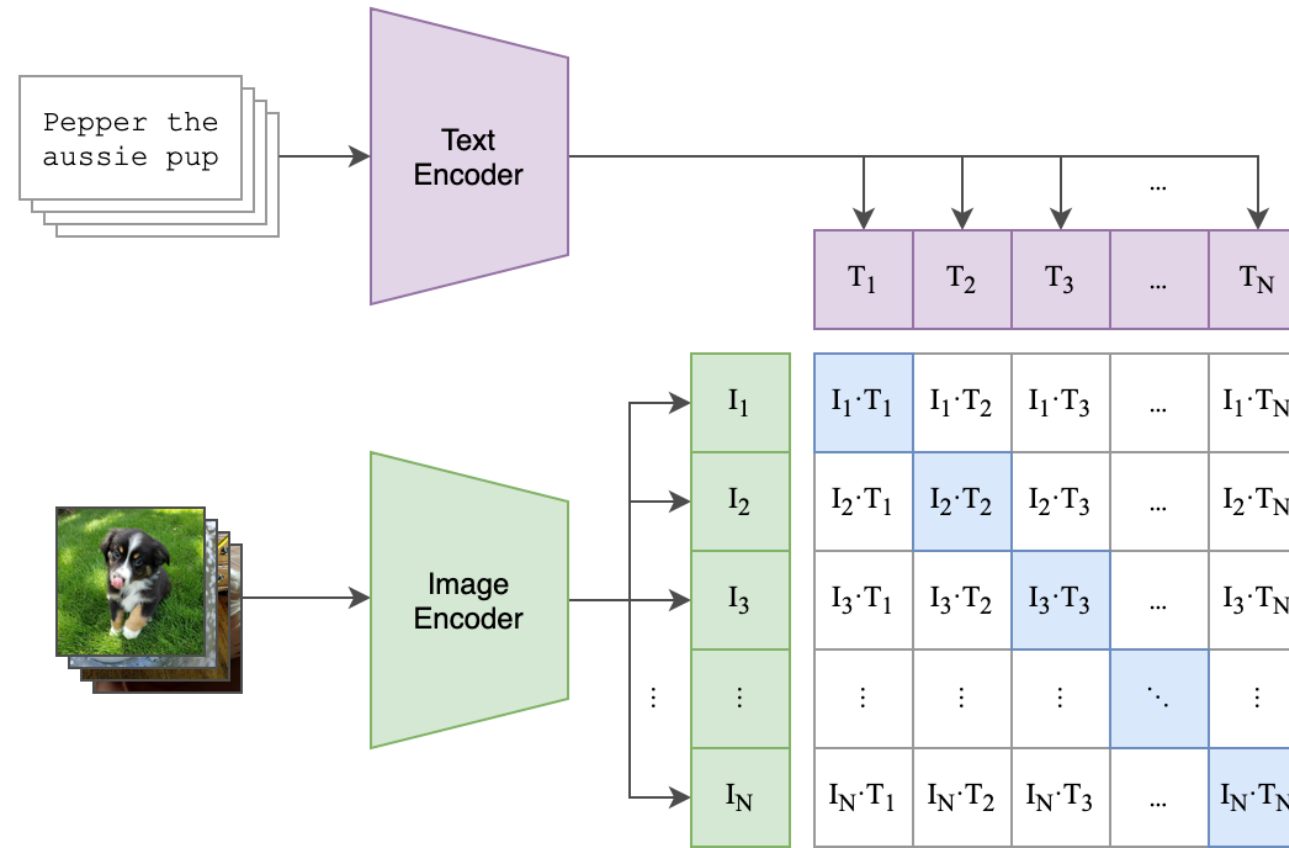
# Vision Transformer



## Transformer Encoder



# CLIP: Contrastive Language-Image Pre-Training

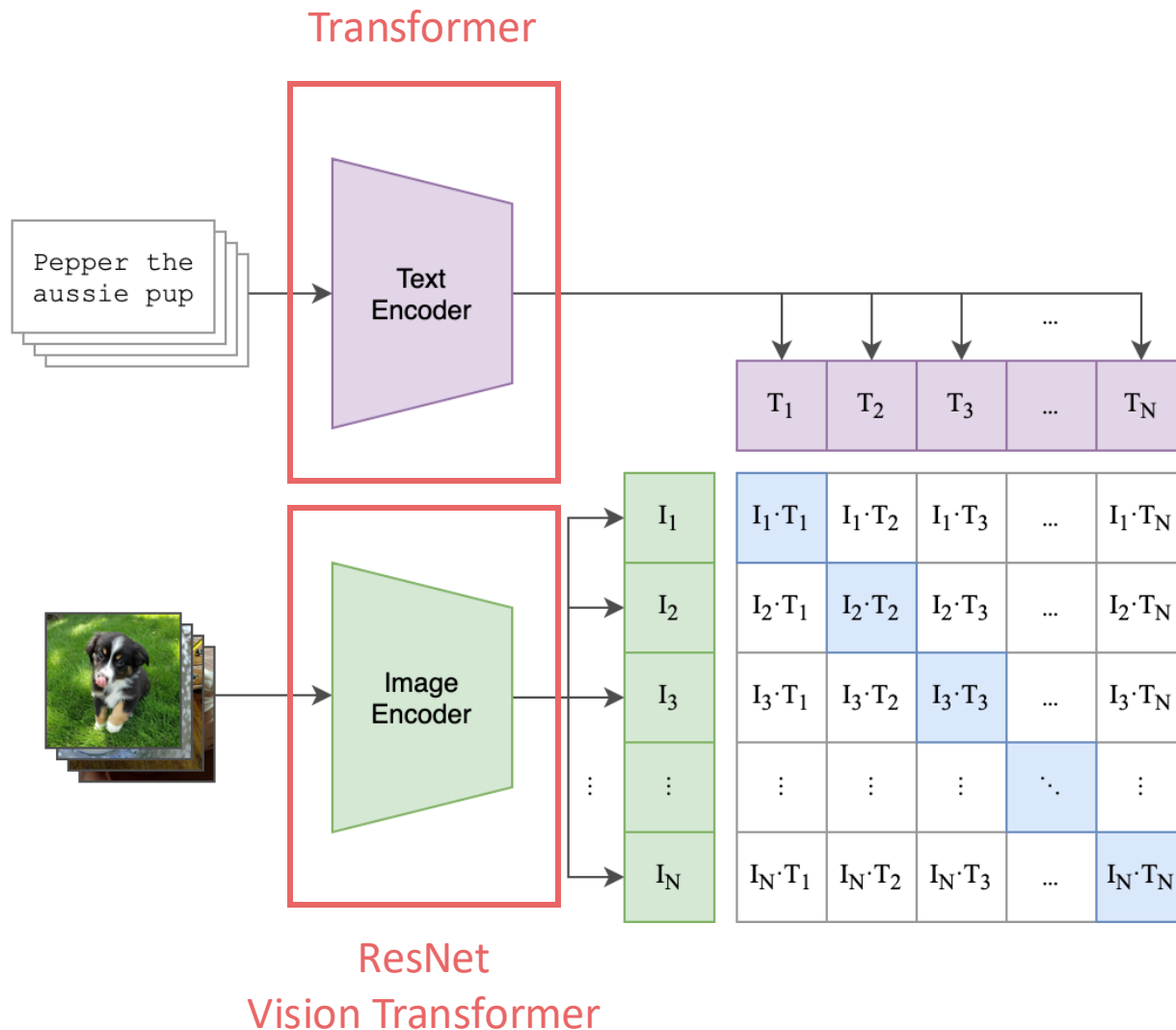


# Training with Image-Caption Pairs

Cosine similarity between text and image features



# Training Details



```
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]
```

```
# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
```

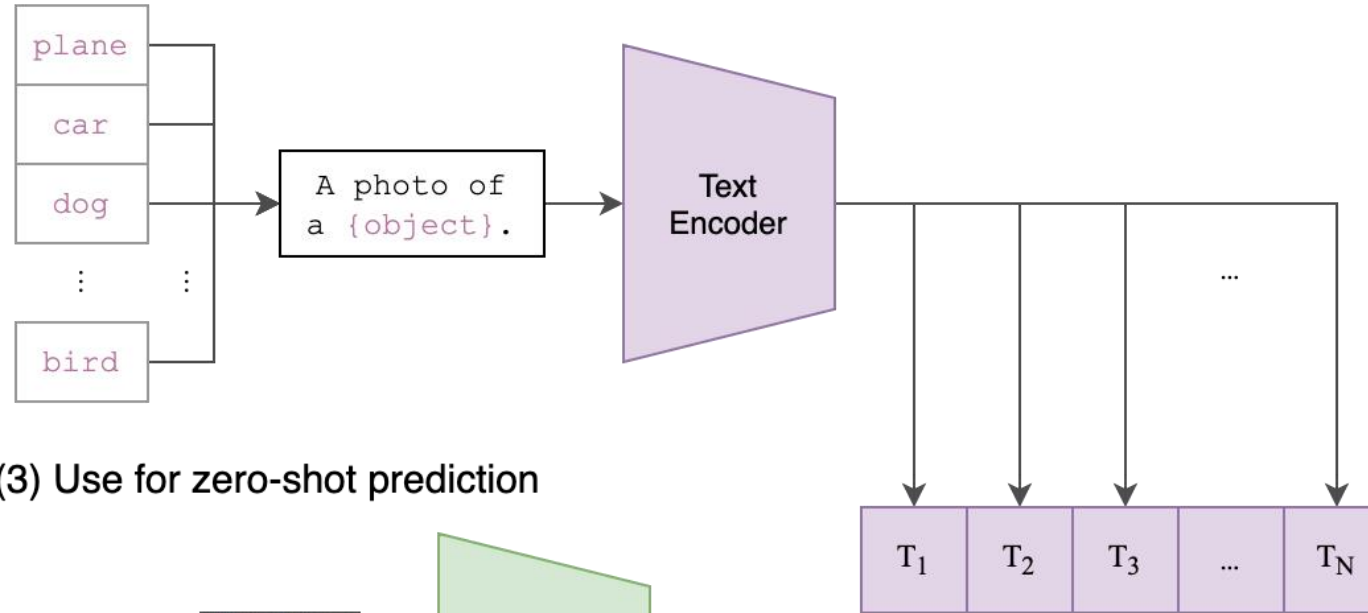
```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)
```

```
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss    = (loss_i + loss_t)/2
```

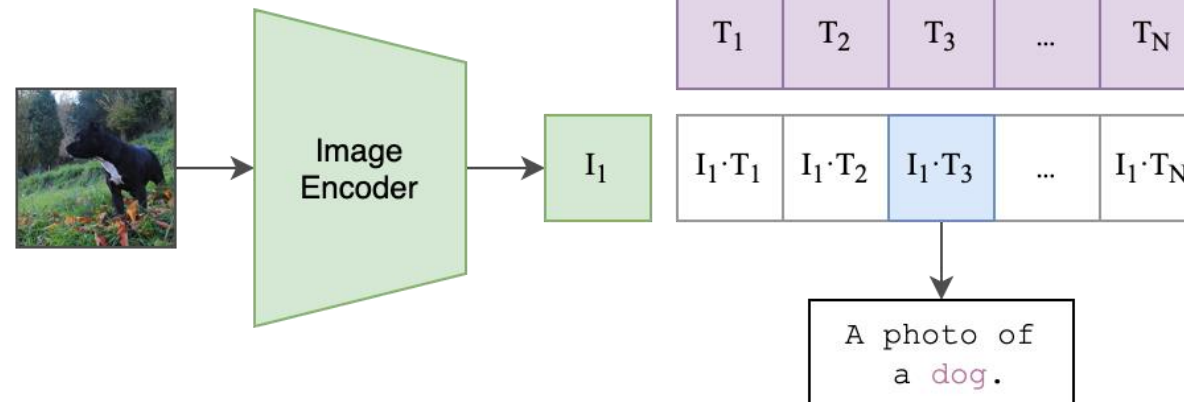
1	0	0	...	0
0	1	0	...	0
0	0	1	...	0
0	0	0	...	0
...	...	...	...	0
0	0	0	0	1

# Zero-Shot Prediction

(2) Create dataset classifier from label text

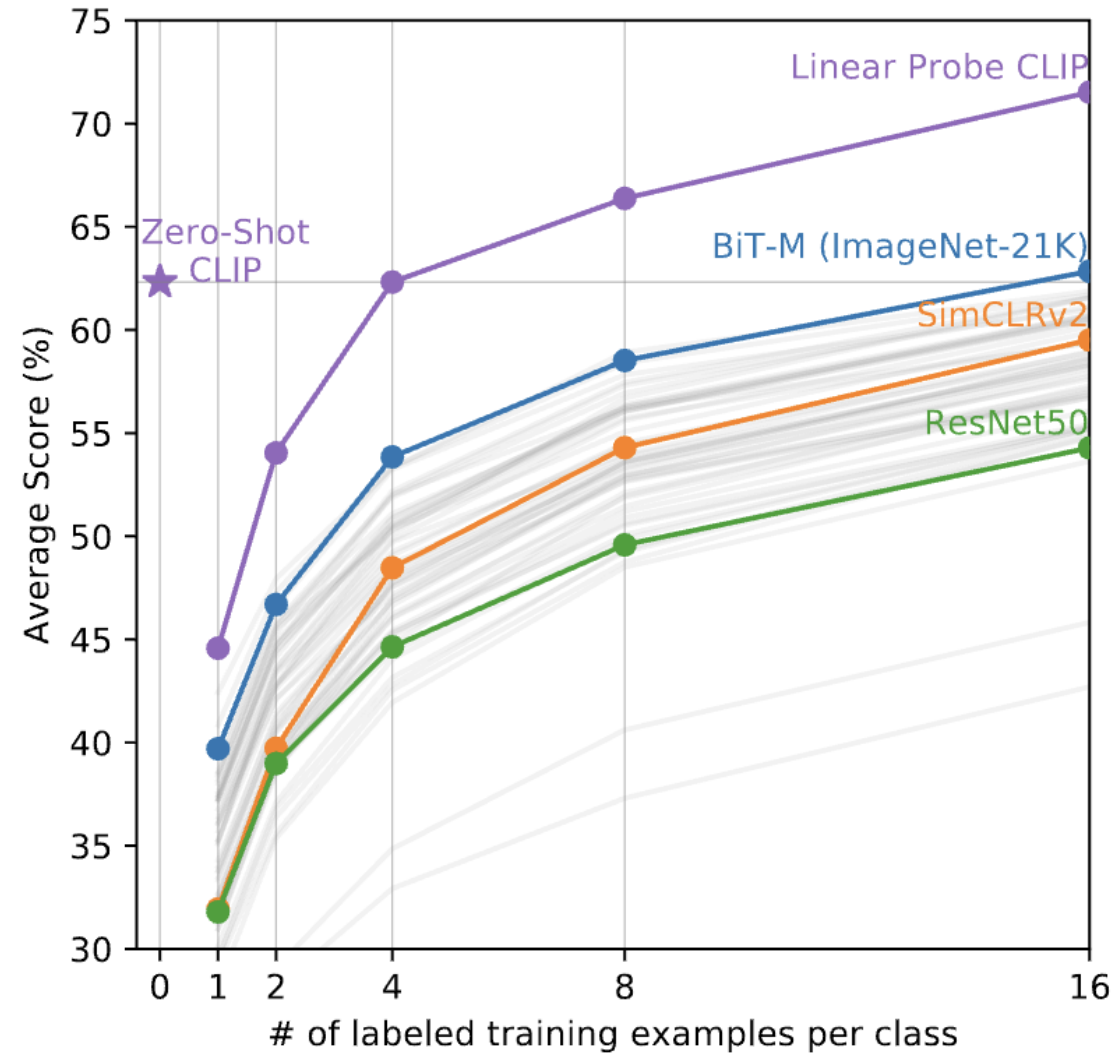


(3) Use for zero-shot prediction

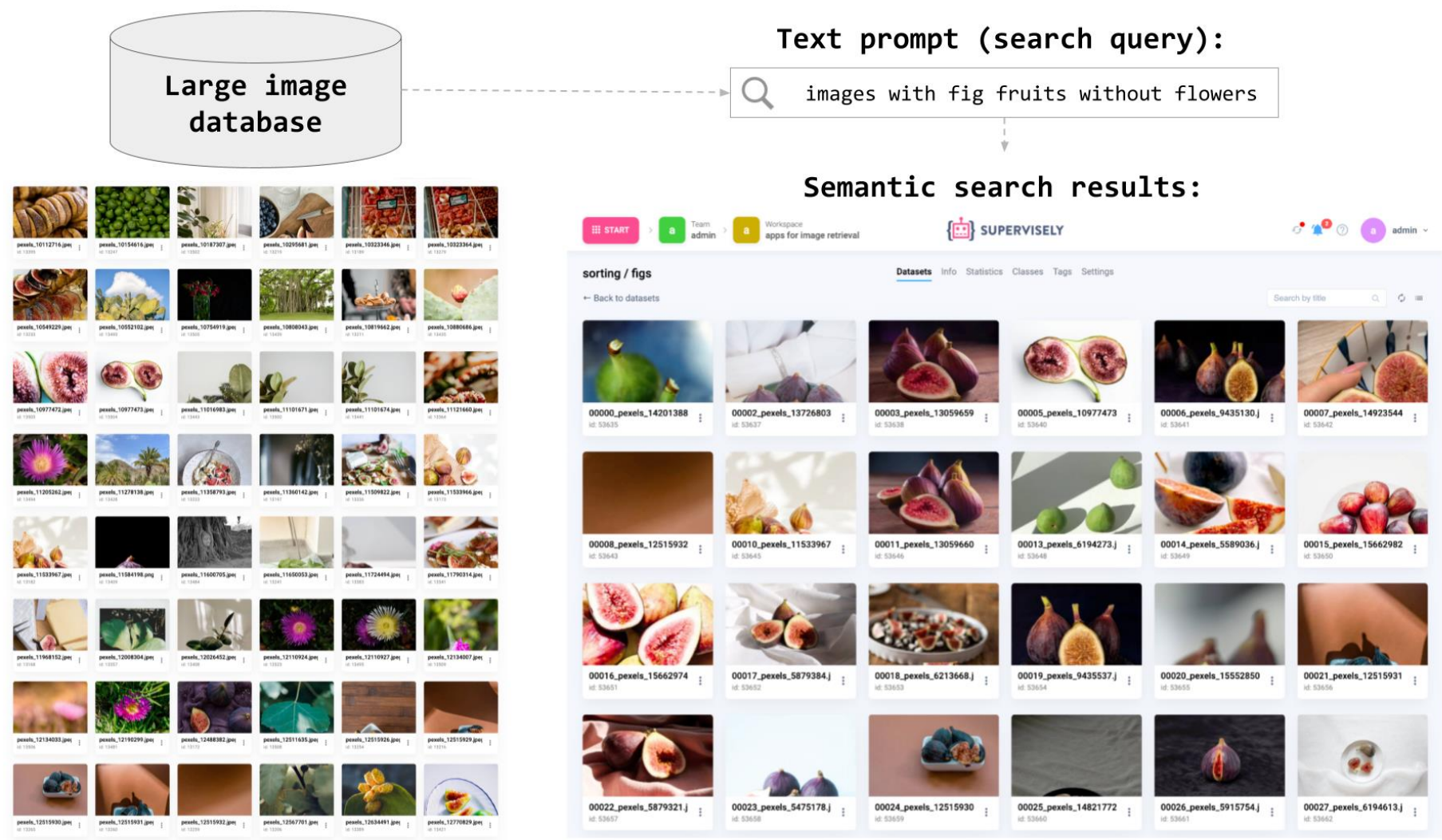




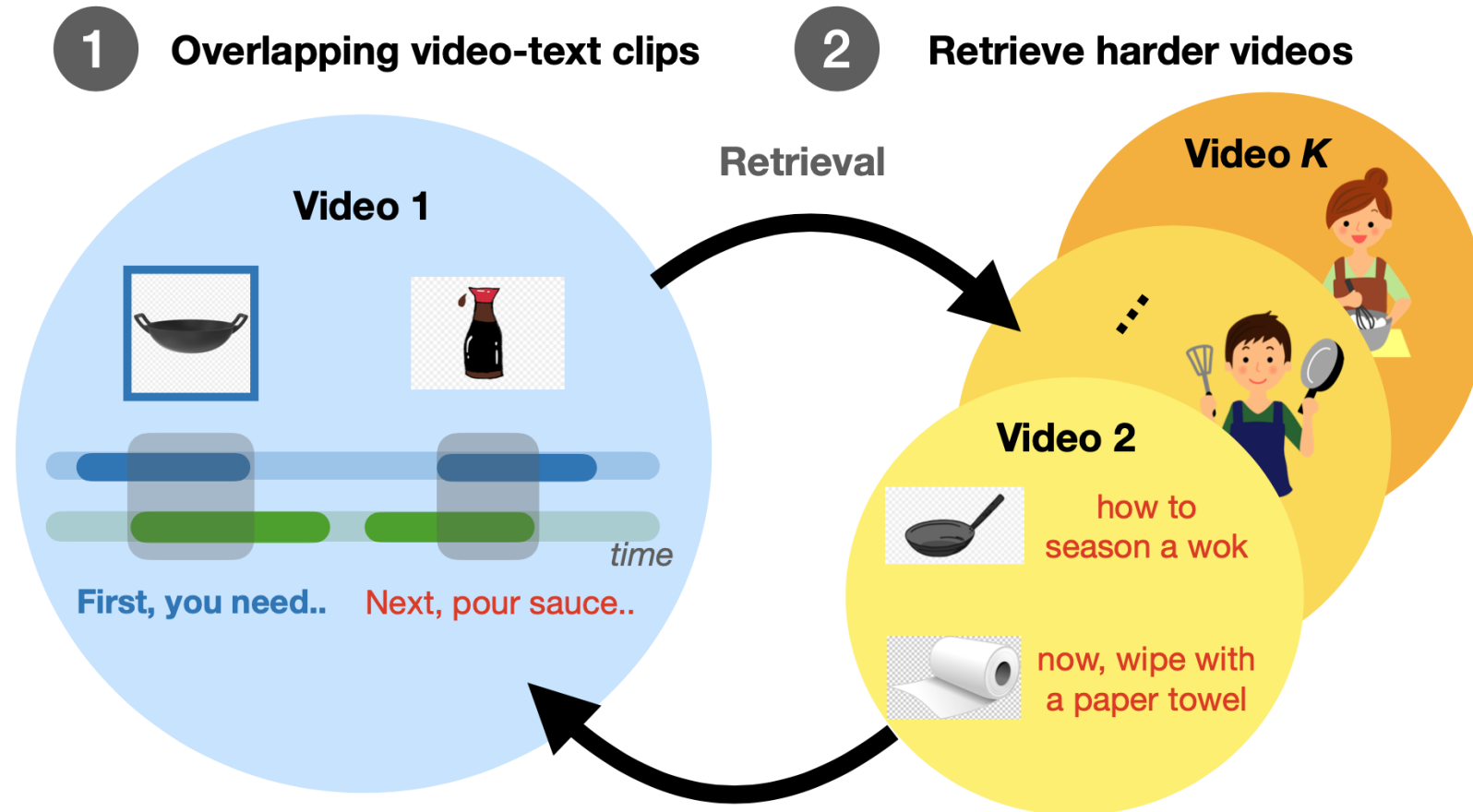
# Zero-Shot CLIP vs. Few-shot Linear Probes



# Image Retrieval with Text Query



# VideoCLIP



VideoCLIP: Contrastive learning with **hard-retrieved negatives** and **overlapping positives** for video-text pre-training.