CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 15: Text Similarity, Retrieval-Augmented Generation

Kuan-Hao Huang

Spring 2026

(Some slides adapted from ACL 2023 Tutorial: Retrieval-based Language Models and Applications)

Quiz 2

- Average: 90.67
- Std: 8.67
- Q1: 84
- Median: 93
- Q3: 97

Check Gradescope for details. For questions, send emails to csce638-ta-26s@lists.tamu.edu with "[CSCE 638] Subject …" or check with TA in office hours

Quiz 3

- Mar 18 (Wednesday)
- Coverage: mainly Lecture 11 to 15
  - Naturally include some concepts in Lecture 1 to 10
- In-class, 20 minutes, closed book, no cheat sheet
- Written quiz, 5 questions
  - Please bring a pen
- Tips
  - Get familiar with formula
  - Understand the intuition behind the formula and the design
  - Know the pros and cons of different approaches

Assignment 3

Assignment 3

RELEASE DATE: 03/15/2026

DUE DATE: 04/02/2026 11:59pm on Gradescope

LaTeX Template: https://www.overleaf.com/read/prmshxybxqqh#4fa8a9

Name: First-Name Last-Name UIN: 000000000

This assignment consists of two parts: a writing section and a programming section. For the writing section, please use the provided LaTeX template to prepare your solutions and remember to fill in your name and UIN. For the programming section, please follow the instructions carefully.

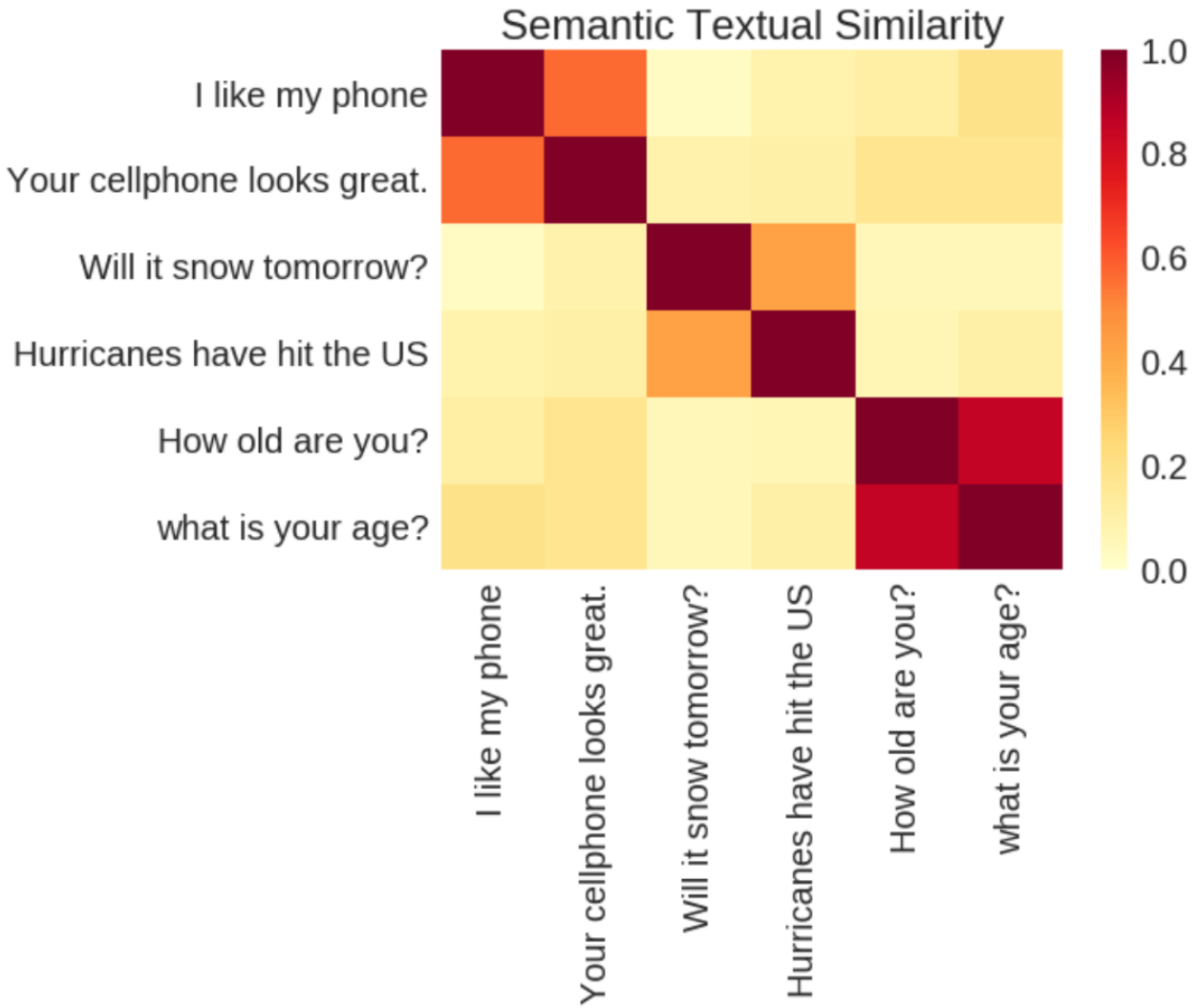
Discussions with others on course materials and assignment solutions are encouraged, and the use of AI tools as assistance is permitted. However, you must ensure that ***the final solutions are written in your own words****. It is your responsibility to avoid excessive similarity to others' work. Additionally, please clearly* ***indicate any parts where AI tools were used*** *as assistance.*

If you have any question, please send an email to csce638-ta-26s@list.tamu.edu

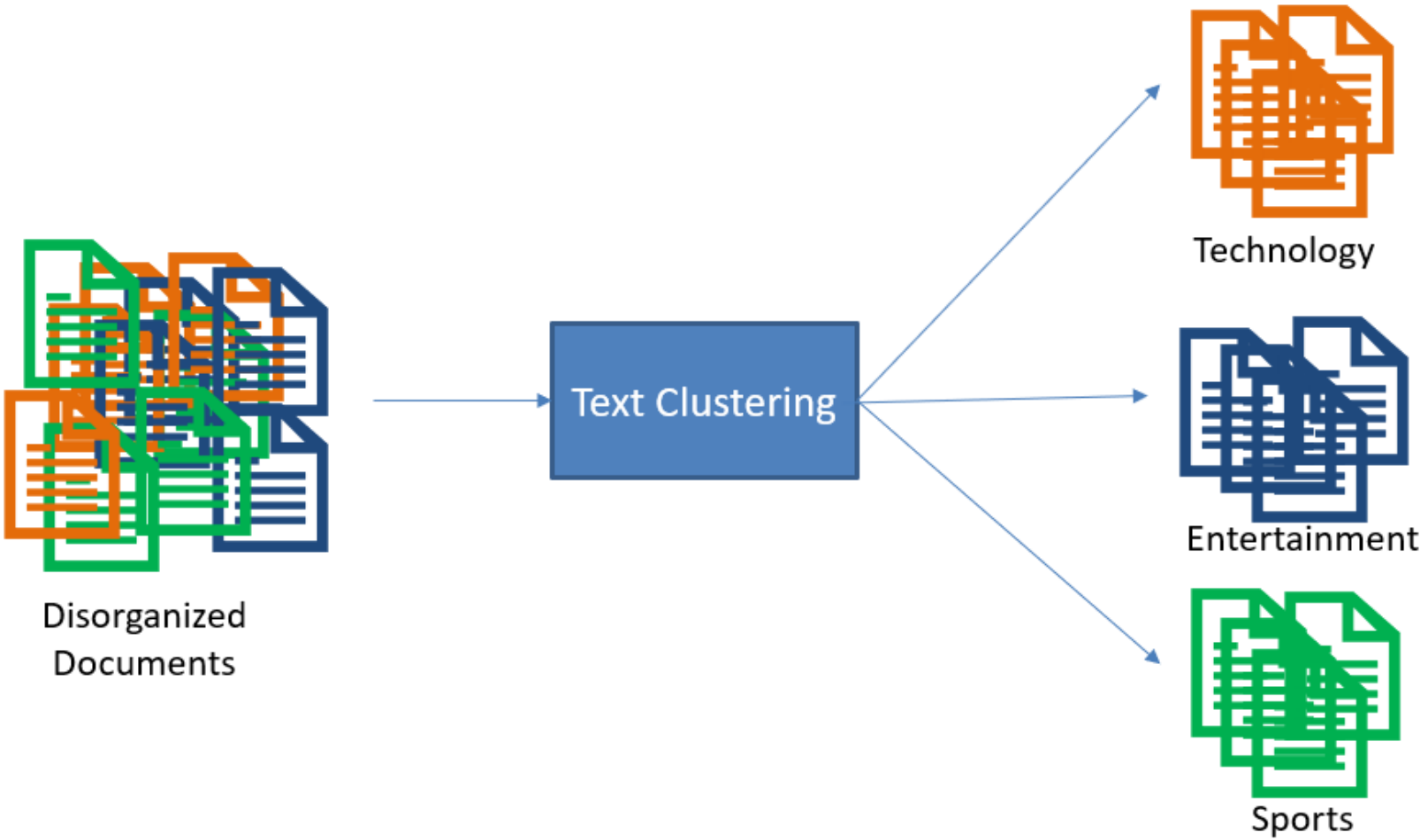
Lecture Plan

- Text Similarity
- Retrieval-Augmented Generation

Text Similarity

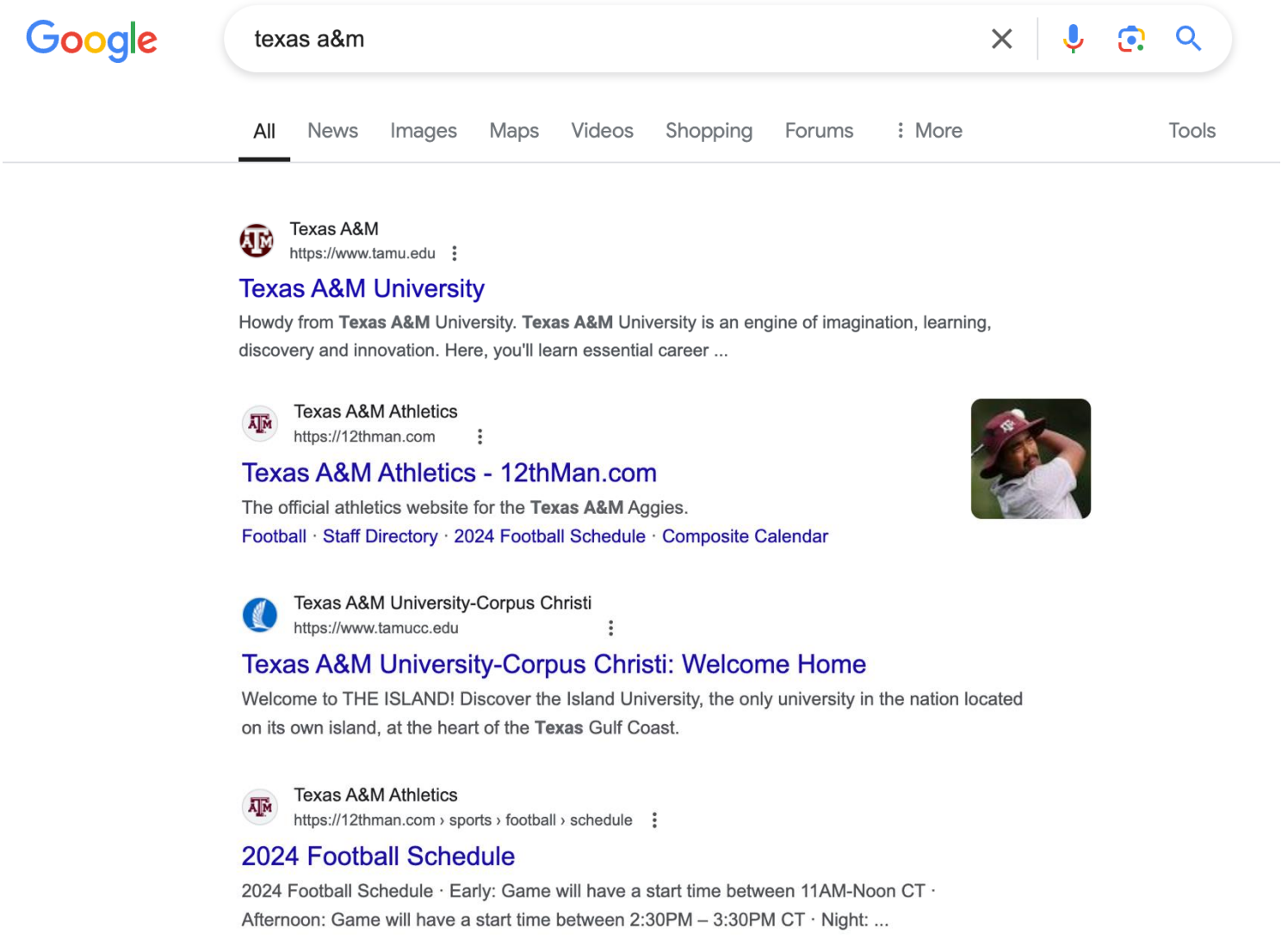


Document Clustering



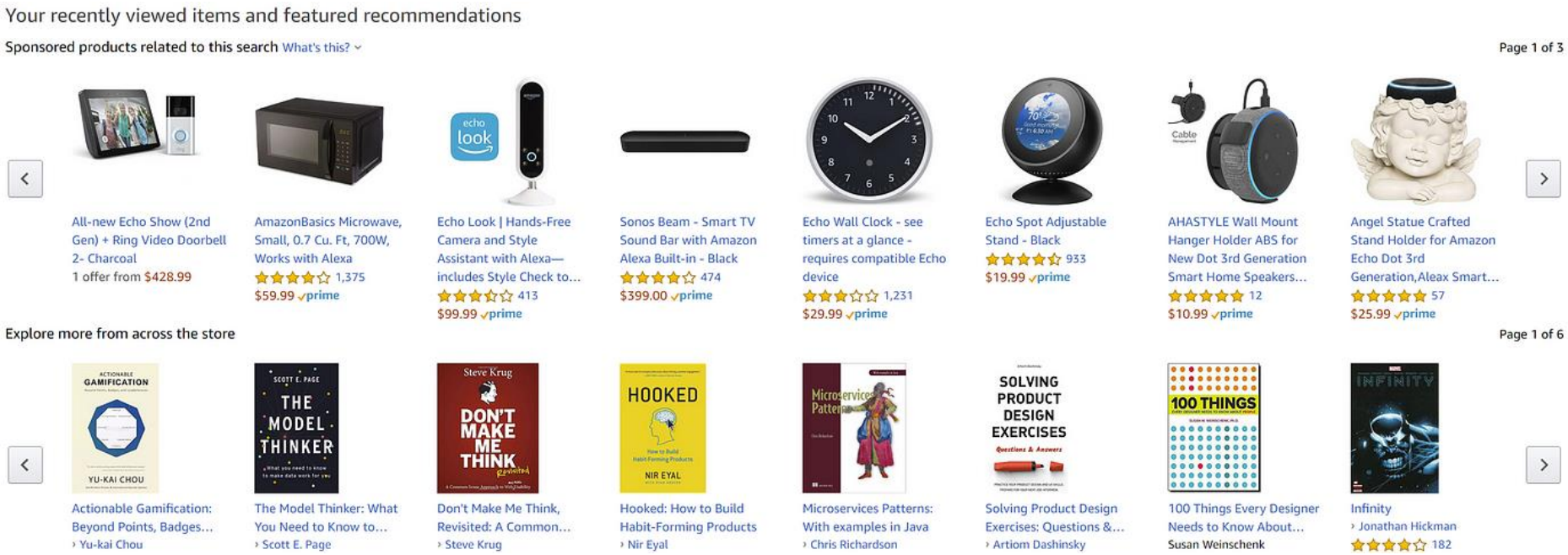
https://medium.com/@danielafrimi/text-clustering-using-nlp-techniques-c2e6b08b6e95

Information Retrieval



https://www.google.com/

Recommendation Systems



https://www.amazon.com/

Semantic Quality Control

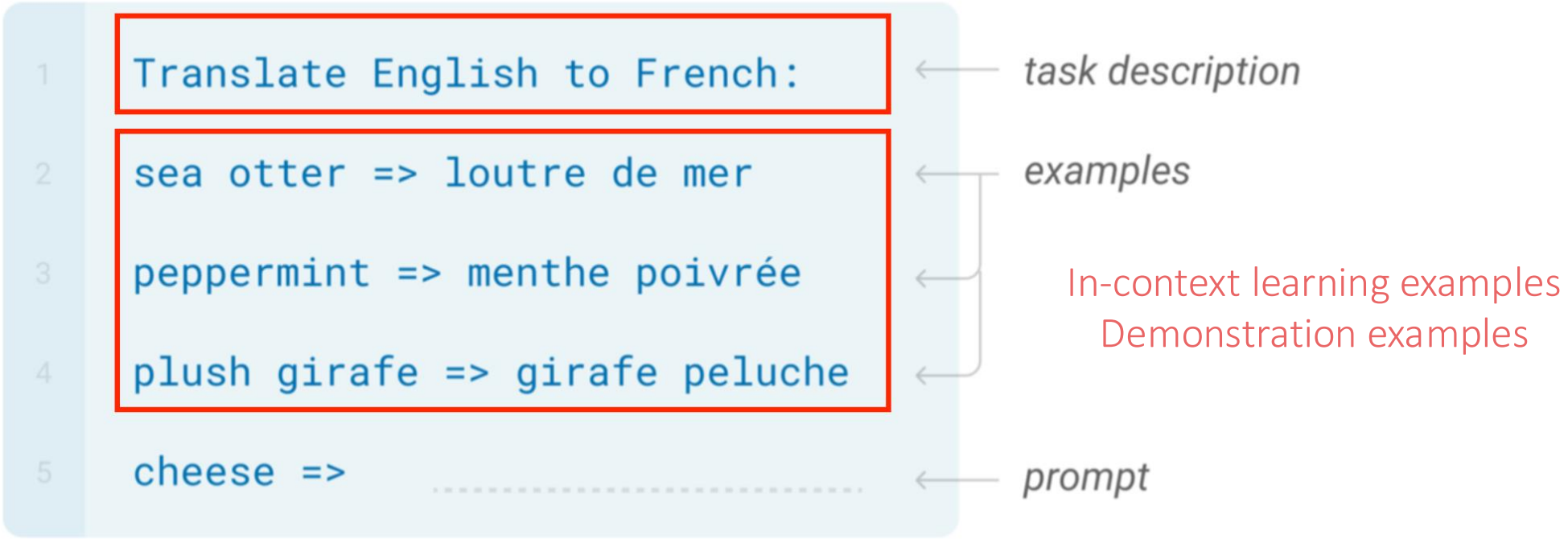
- Paraphrase generation

We will go hiking if tomorrow is a sunny day.

If it is sunny tomorrow, we will go hiking.

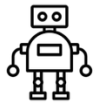
- Style transfer
- Plagiarism detection
- …

In-Context Example Selection



Language Models are Few-Shot Learners, 2020

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

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

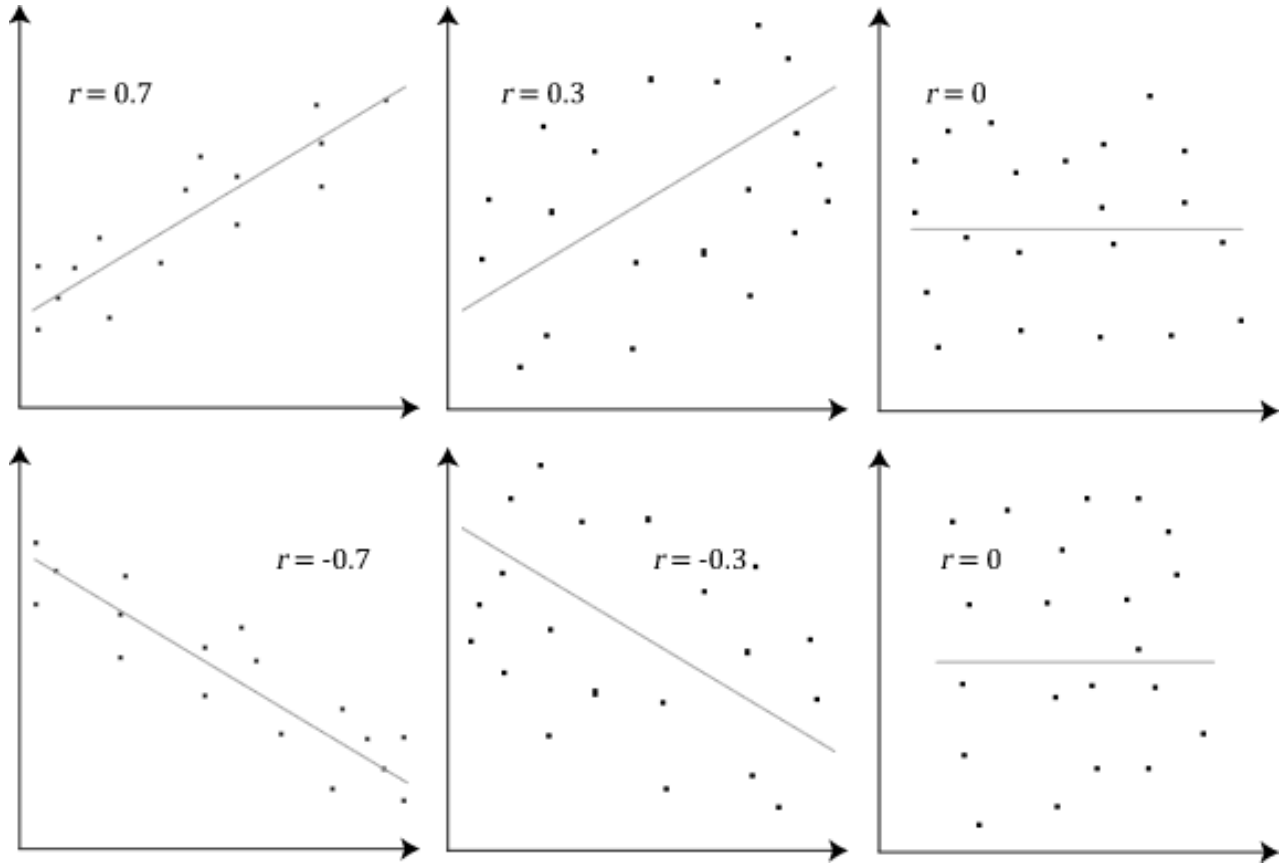
r = correlation coefficient

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

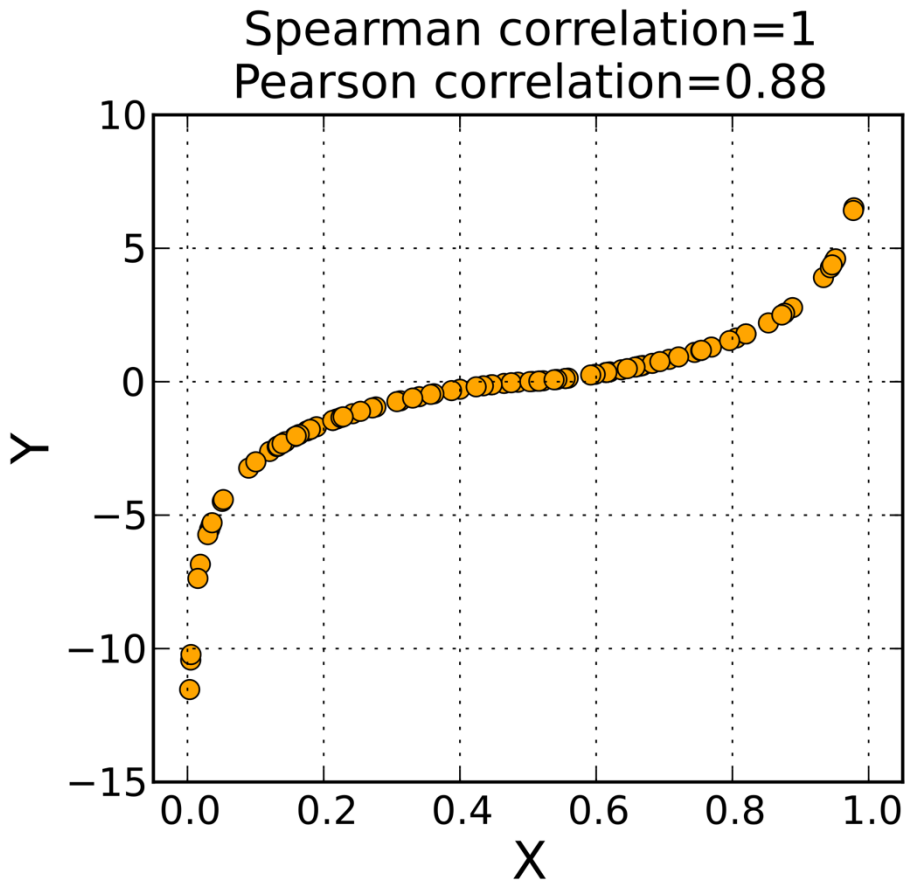
$\bar{y}$ = mean of the values of the y-variable



https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

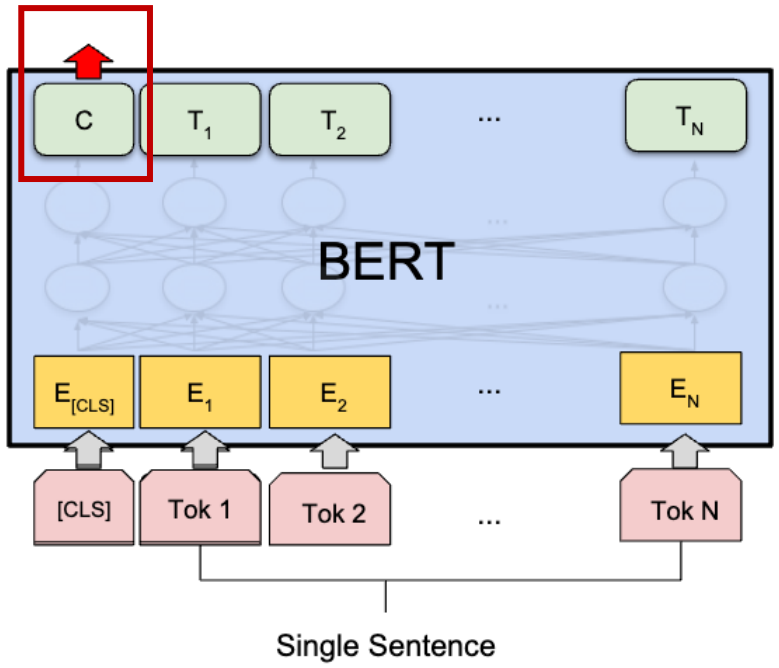
Spearman's Correlation Coefficient

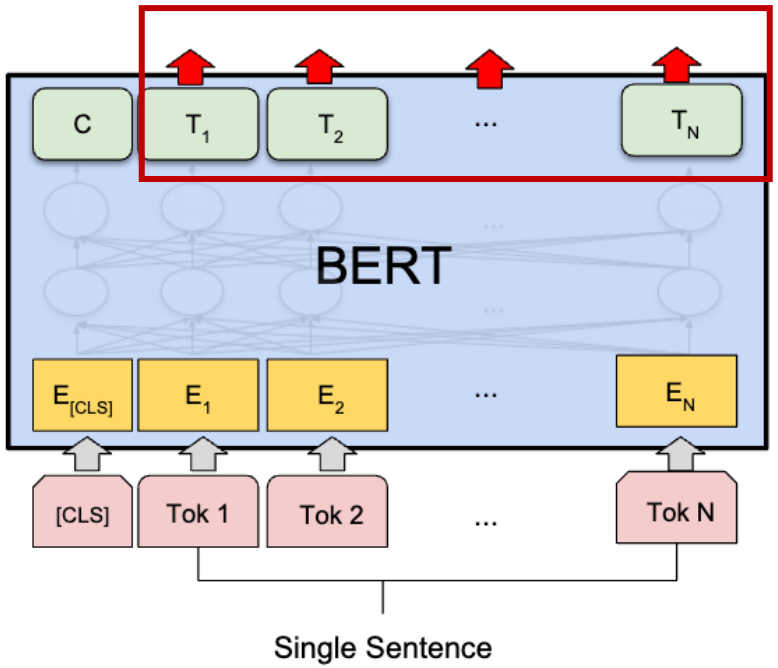
- Pearson's correlation coefficient on rank
- Score
  - Human: [1.2, 3.4, 2.5, 0.7, 4.0]
  - Machine: [0.5, 3.3, 1.0, 1.2, 3.4]
- Rank
  - Human: [4, 2, 3, 5, 1]
  - Machine: [5, 2, 4, 3, 1]
- Assesses monotonic relationships
  - whether linear or not



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

A Simple Approach: Text Encoder + Cosine Similarity





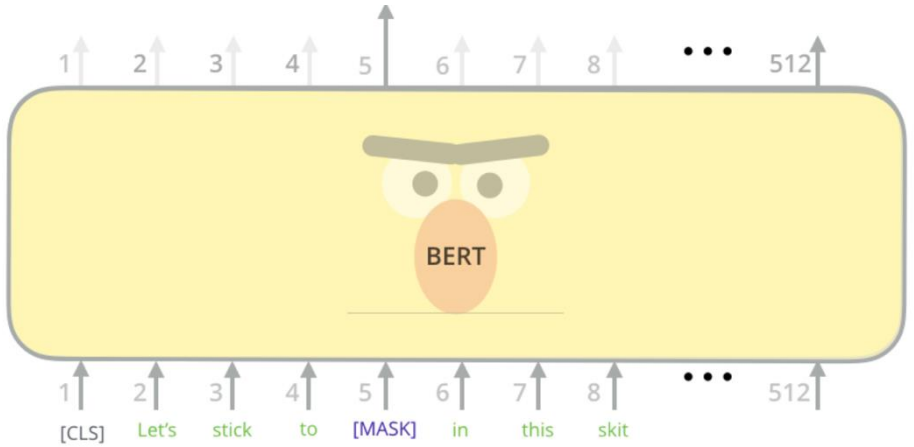
$$E_1 = \text{Encoder}(S_1)$$

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

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

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.

We will go hiking if tomorrow is a sunny day.

Good classification performance ≠ Good similarity

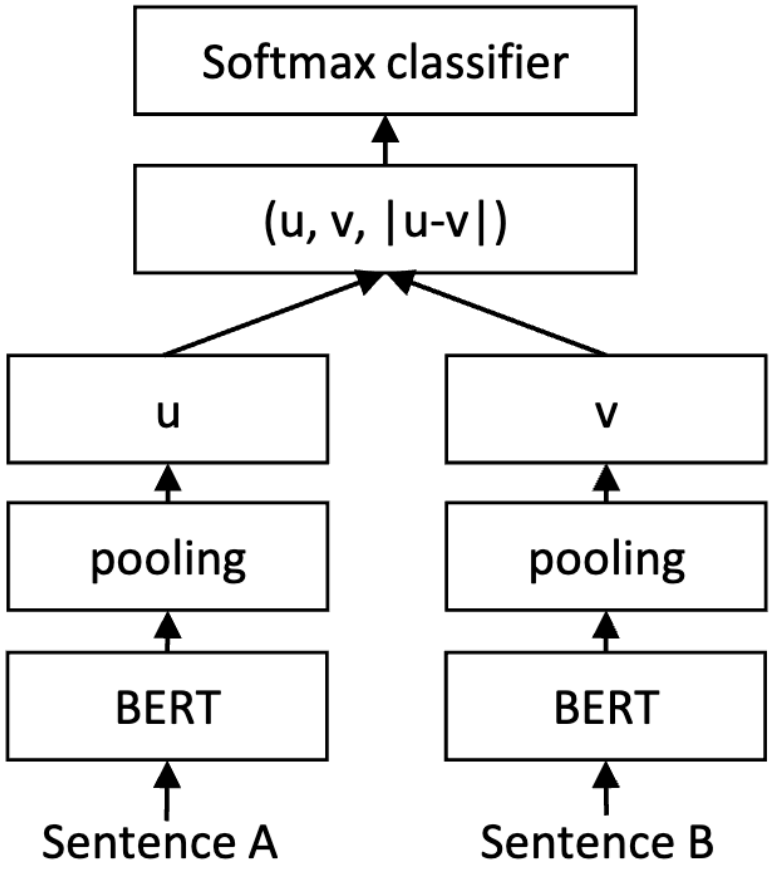
Sentence-BERT

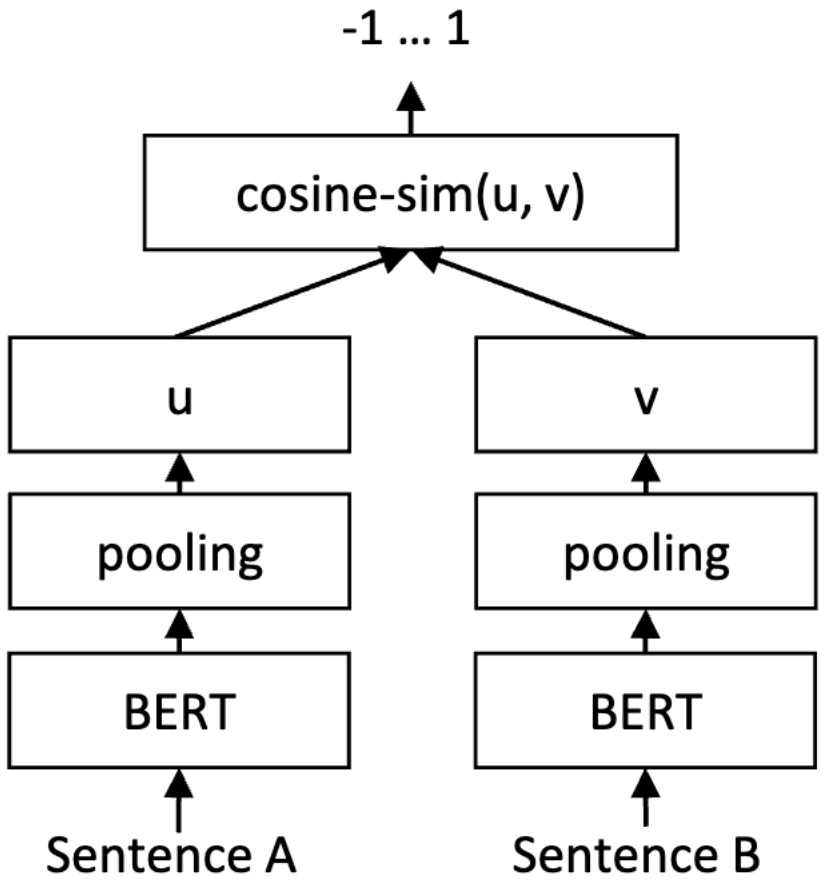
- Consider SNLI dataset
  - Stanford Natural Language Inference

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

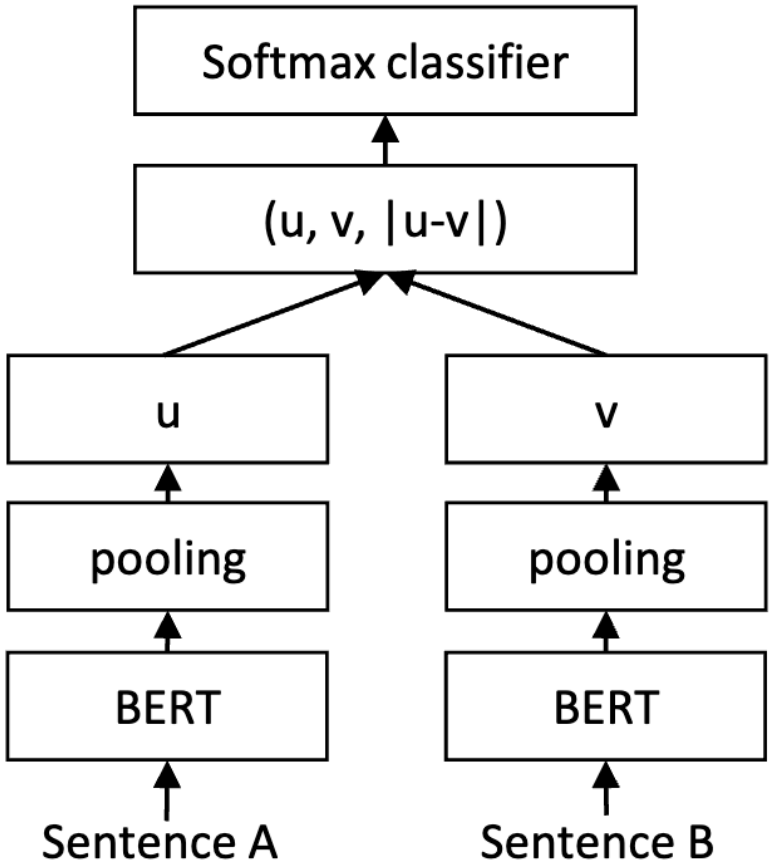




Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

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: Sentence Embeddings using Siamese BERT-Networks

Sentence-BERT

If it is sunny tomorrow, we will go hiking.

We will go hiking if tomorrow is a sunny day.

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	**76.69**	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	**78.46**	**74.90**	80.99	76.25	**79.23**	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	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

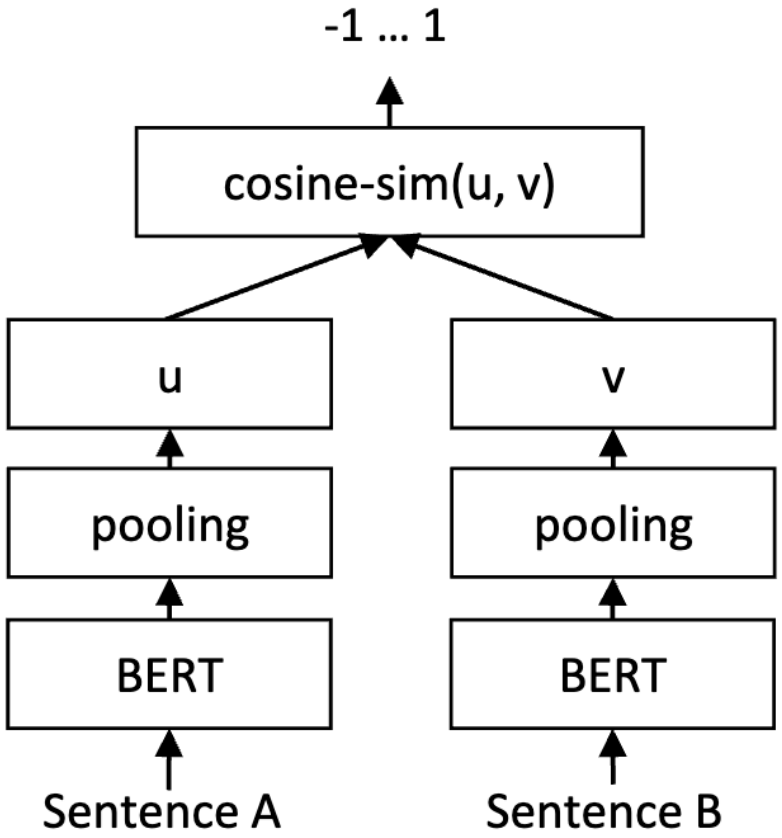
SentenceTransformers

- https://sbert.net/

```python
from sentence_transformers import SentenceTransformer

# 1. Load a pretrained Sentence Transformer model
model = SentenceTransformer("all-MiniLM-L6-v2")

# The sentences to encode
sentences = [
    "The weather is lovely today.",
    "It's so sunny outside!",
    "He drove to the stadium.",
]

# 2. Calculate embeddings by calling model.encode()
embeddings = model.encode(sentences)
print(embeddings.shape)
# [3, 384]

# 3. Calculate the embedding similarities
similarities = model.similarity(embeddings, embeddings)
print(similarities)
# tensor([[1.0000, 0.6660, 0.1046],
#         [0.6660, 1.0000, 0.1411],
#         [0.1046, 0.1411, 1.0000]])
```

SimCSE

- Simple Contrastive Learning of Sentence Embeddings



Contrastive Loss

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

SimCSE: Supervised Contrastive Learning

Sentence 1A	*Sentence 1B*	Positive
Sentence 2A	*Sentence 2B*	Negative
Sentence 3A	*Sentence 3B*	Negative
Sentence 4A	*Sentence 4B*	Negative
Sentence 5A	*Sentence 5B*	Negative

Contrastive Loss

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

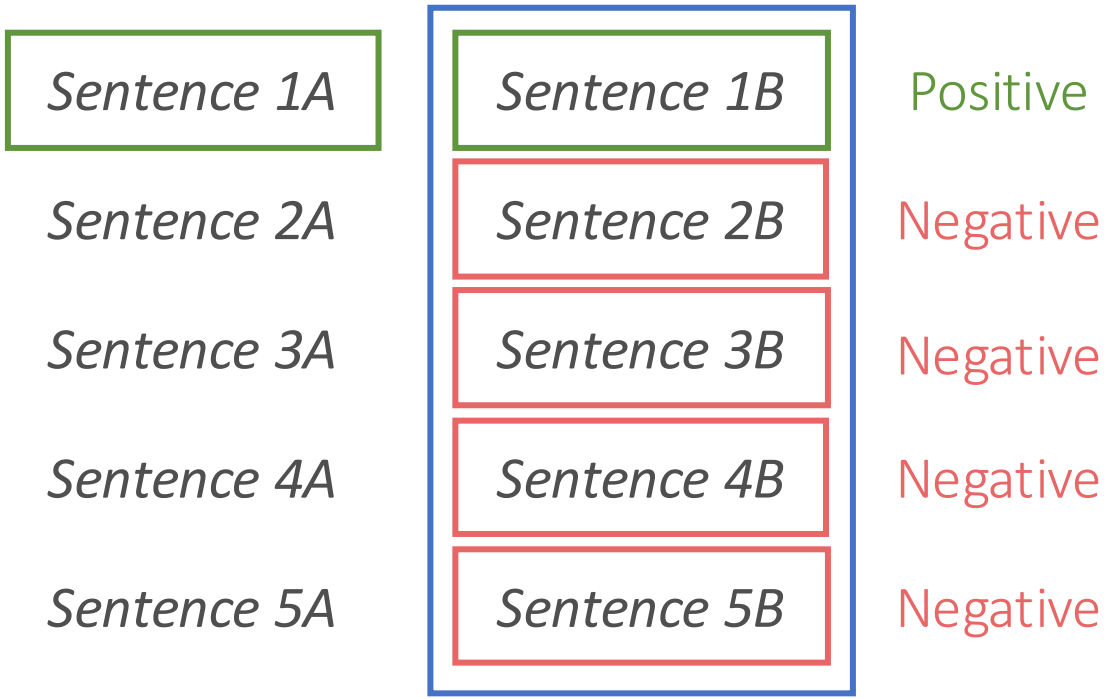
SimCSE: Supervised Contrastive Learning

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

Contrastive Loss

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

SimCSE: Supervised Contrastive Learning



Contrastive Loss

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

SimCSE: Supervised Contrastive Learning

Sentence 1A	*Sentence 1B*	Negative
Sentence 2A	*Sentence 2B*	Positive
Sentence 3A	*Sentence 3B*	Negative
Sentence 4A	*Sentence 4B*	Negative
Sentence 5A	*Sentence 5B*	Negative

Contrastive Loss

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

SimCSE: Supervised Contrastive Learning

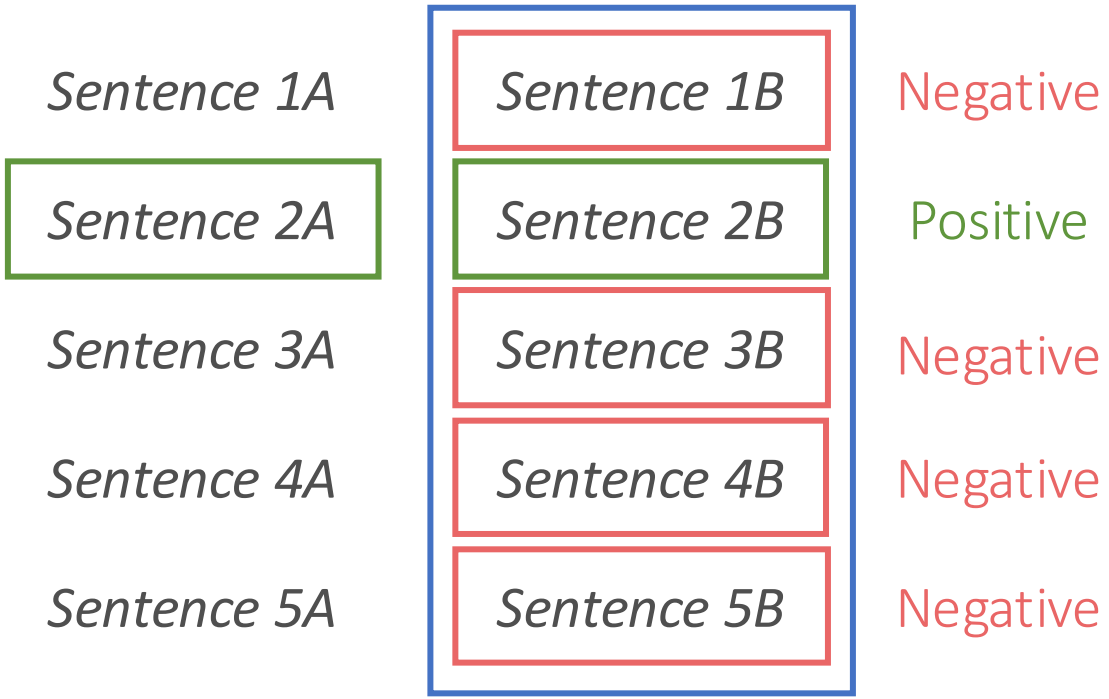
Sentence 1A	*Sentence 1B*
Sentence 2A	*Sentence 2B*
Sentence 3A	*Sentence 3B*
Sentence 4A	*Sentence 4B*
Sentence 5A	*Sentence 5B*

Positive

Contrastive Loss

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

SimCSE: Supervised Contrastive Learning

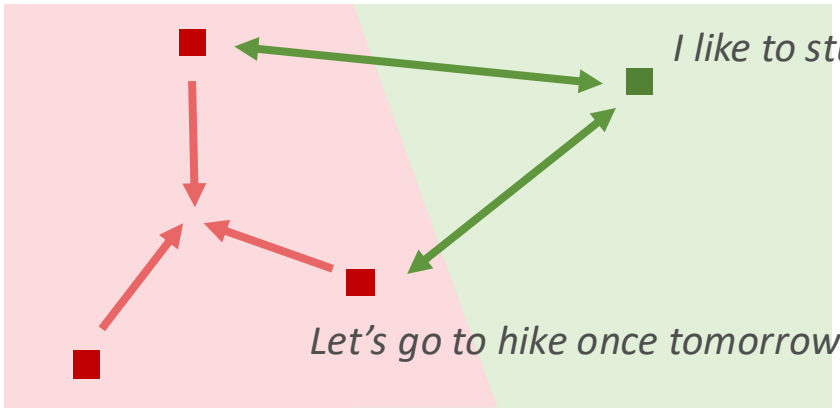


Contrastive Loss

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

SimCSE: Supervised Contrastive Learning

If it is sunny tomorrow, we will go hiking.



I like to study large language models.

Let's go to hike once tomorrow is sunny.

We will go hiking if tomorrow is a sunny day.

SimCSE: Performance

Supervised models								
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	76.69	71.22
$\text{SBERT}_{\text{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
$\text{CT-SBERT}_{\text{base}}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
∗ $\text{SimCSE-BERT}_{\text{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\text{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\text{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ $\text{SimCSE-RoBERTa}_{\text{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ $\text{SimCSE-RoBERTa}_{\text{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

Supervised Data

- 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

What if we don't have supervised data?

SimCSE: Unsupervised Contrastive Learning

Sentence 1	*Sentence 1'*	Positive
Sentence 2		Negative
Sentence 3		Negative
Sentence 4		Negative
Sentence 5		Negative

Contrastive Loss

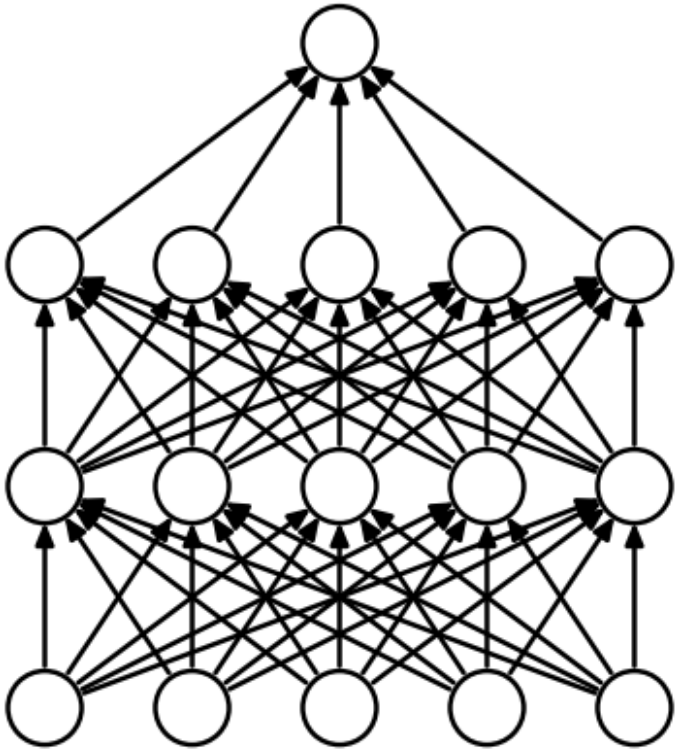
$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

Generate positive example with masking

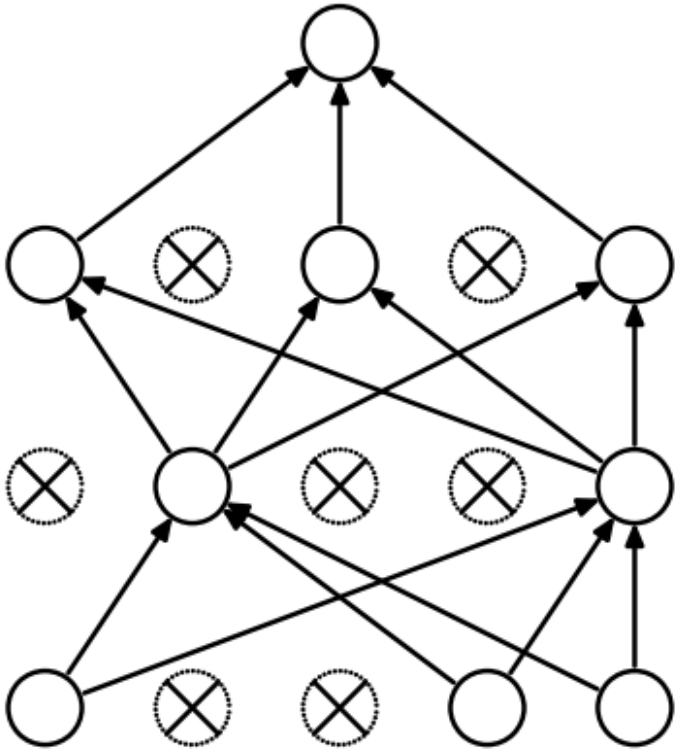
If it is sunny tomorrow, we will go hiking.

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

Dropout



(a) Standard Neural Net



(b) After applying dropout.

Generate positive example with neuron masking

Dropout: A Simple Way to Prevent Neural Networks from Overfitting

SimCSE: Unsupervised Contrastive Learning

Sentence 1 | *Sentence 1'* Positive

Sentence 2

Sentence 3

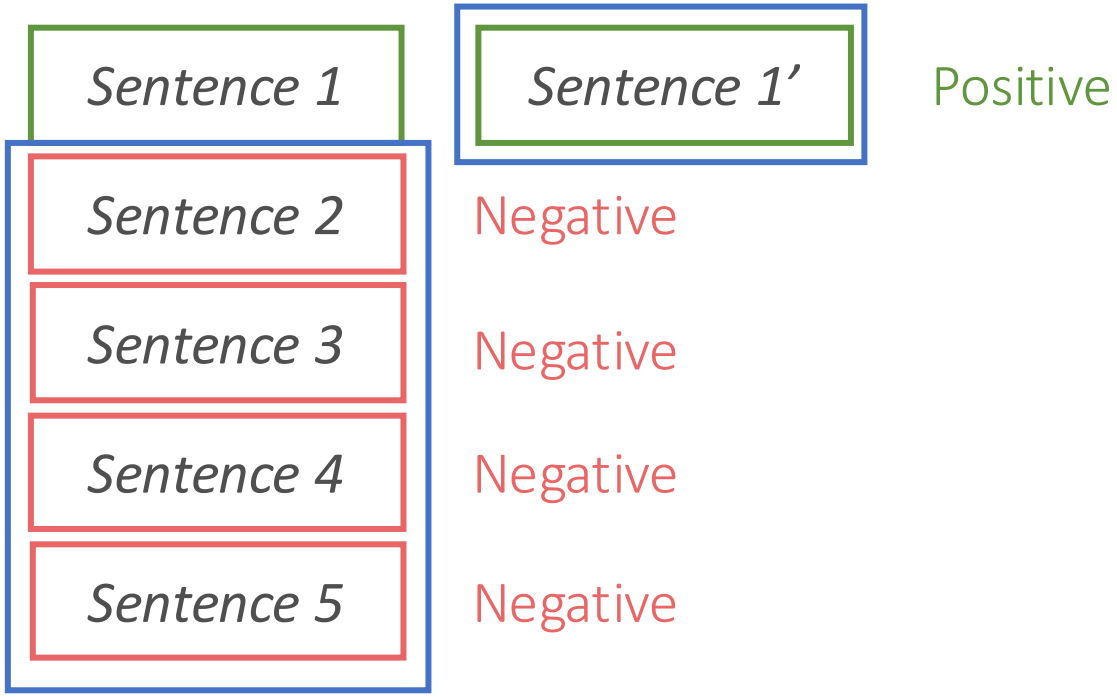
Sentence 4

Sentence 5

Contrastive Loss

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

SimCSE: Unsupervised Contrastive Learning



Contrastive Loss

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

SimCSE: Unsupervised Contrastive Learning

Sentence 1

Sentence 2 | *Sentence 2'* Positive

Sentence 3

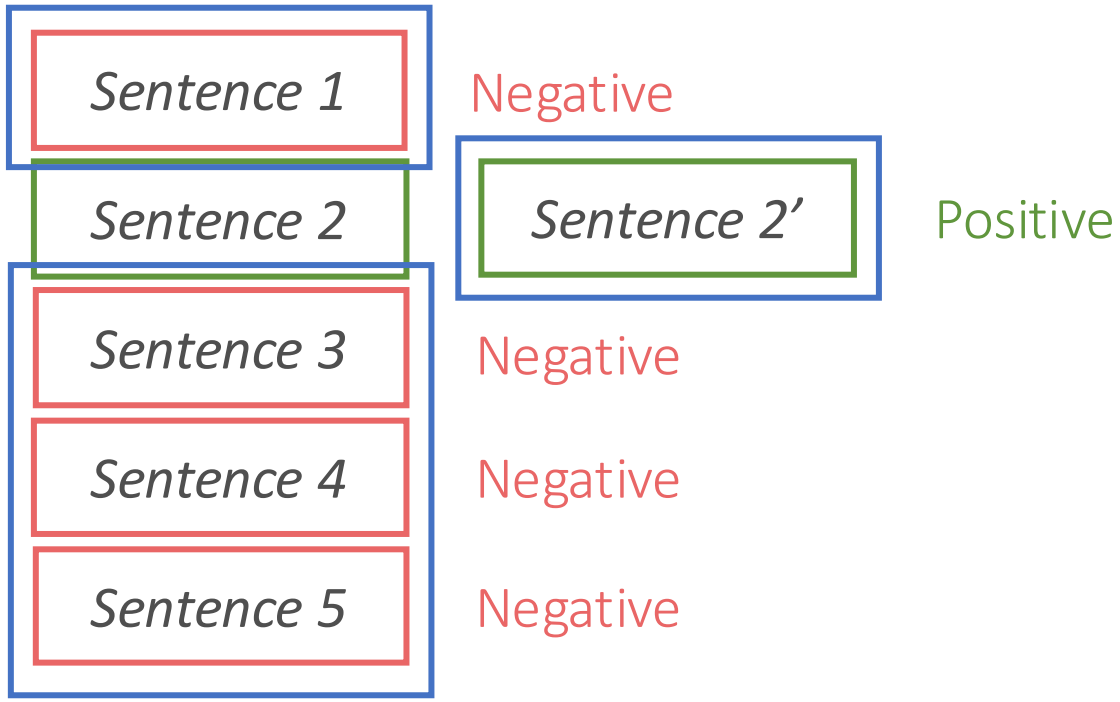
Sentence 4

Sentence 5

Contrastive Loss

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

SimCSE: Unsupervised Contrastive Learning



Contrastive Loss

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

SimCSE: Performance

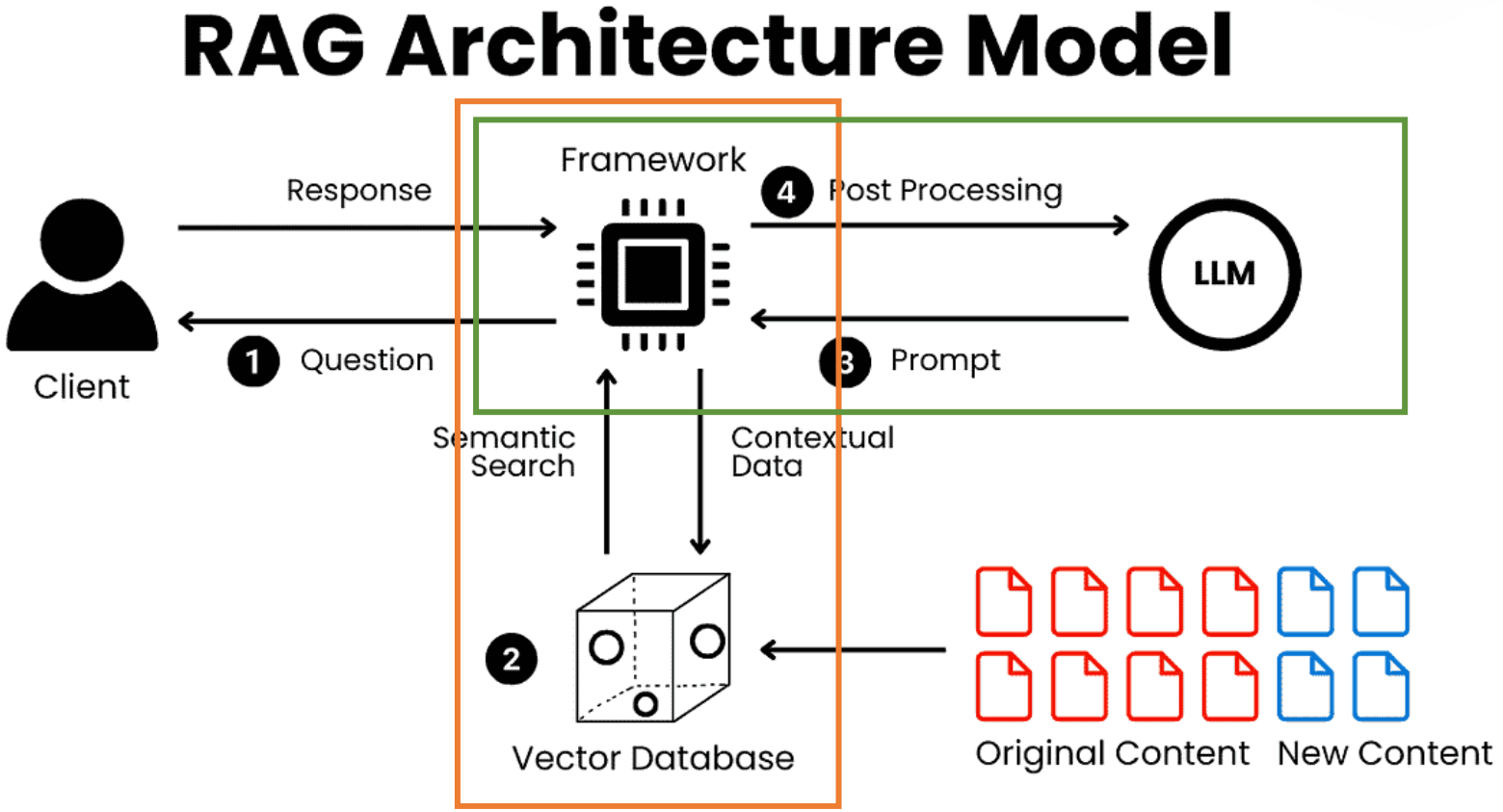
Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16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

Lecture Plan

- Text Similarity
- Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG)

- Information Retrieval + Large Language Models



https://blog.gopenai.com/langchain-and-its-role-in-building-robust-rag-systems-2808b7e5ddb9

Large Language Models

The capital city of Ontario is **Toronto**

LM

Training time

The capital city of Ontario is _____

LM

Test time

Retrieval-Augmented Generation (RAG)

The capital city of Ontario is **Toronto**

LM

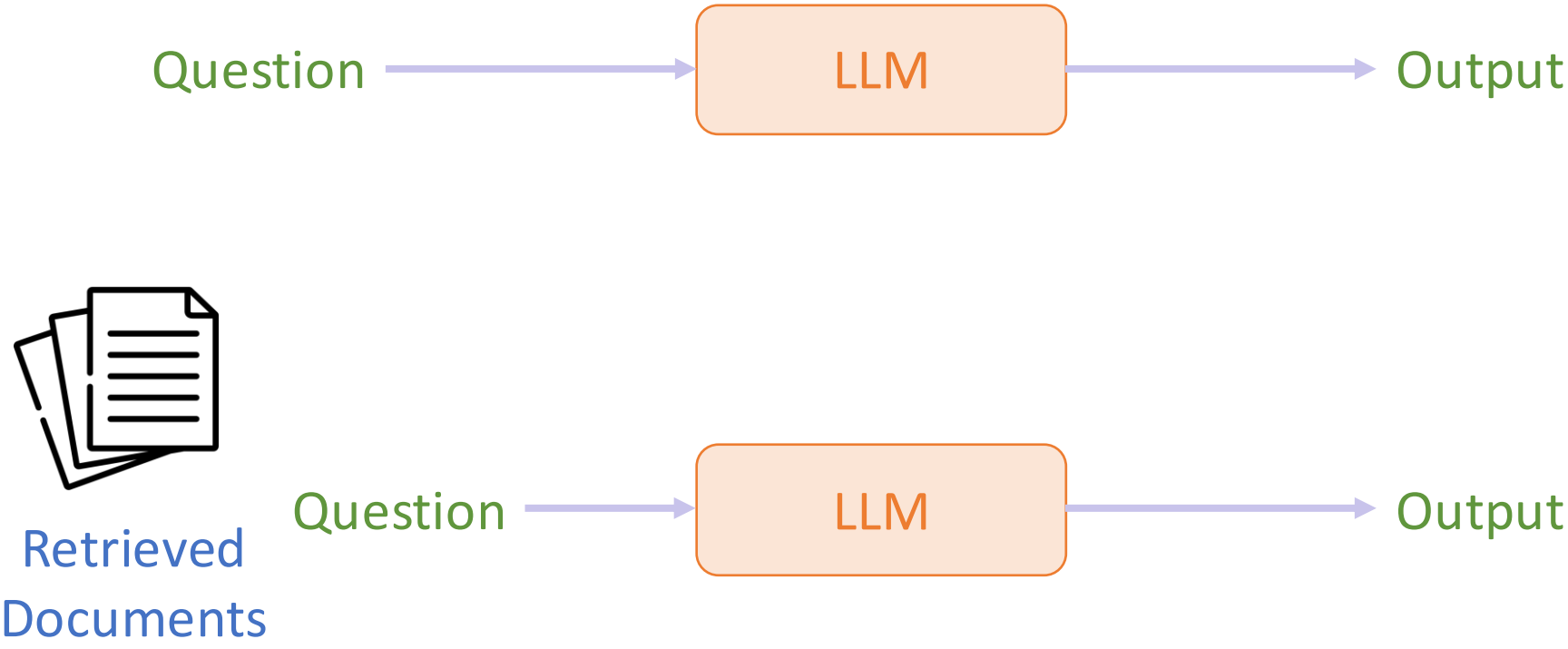
Training time

The capital city of Ontario is _____

LM

Test time

Retrieval-Augmented Generation (RAG)



Why RAG?

- Knowledge Cutoff

What is 2026 WBC US vs Canada result?

Assistant B

There is **no result** yet for a 2026 WBC game between the USA and Canada because **the tournament has not happened yet.**

The 2026 World Baseball Classic is scheduled to take place in **March 2026**.

Here is what we know so far:

- **Status:** The brackets and schedule have not been announced.
- **Qualification:** Both Team USA and Team Canada automatically qualified for the 2026 tournament based on their performance in the 2023 WBC.
- **Likelihood:** While they are often placed in the same pool due to geography, it is not guaranteed they will play each other in the group stage until the official schedule is released (likely in late 2025).

Team USA beats Canada to set up WBC semis showdown vs. D.R.

Jeff Passan
Mar 13, 2026, 11:21 PM ET

Follow

Share Like 3.3K

HOUSTON -- In the aftermath of the Dominican Republic's mercy-rule win over Korea to advance to the World Baseball Classic semifinals, Dominican fans at LoanDepot Park in Miami made clear what they wanted to see next, chanting: "Queremos a USA" -- we want the USA.

They got what they asked for.

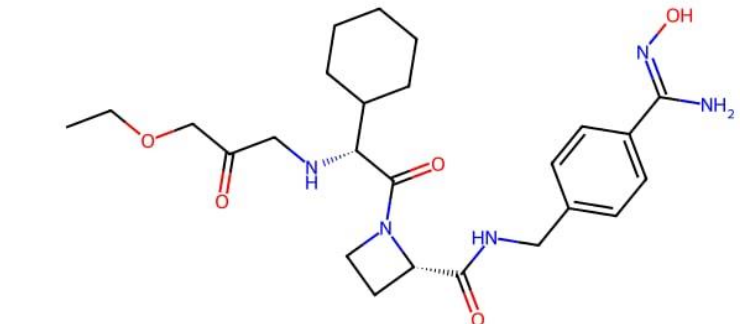
The United States booked its spot in the WBC semifinals with a 5-3 victory against Canada on Friday night, setting up a dream clash between the two most talented teams in the tournament. National League Cy Young winner Paul Skenes will start for Team USA against the formidable Dominican lineup that has wrecked opposing pitchers, while Athletics right-hander Luis Severino is expected to face an American lineup big on names but not so much on production thus far.

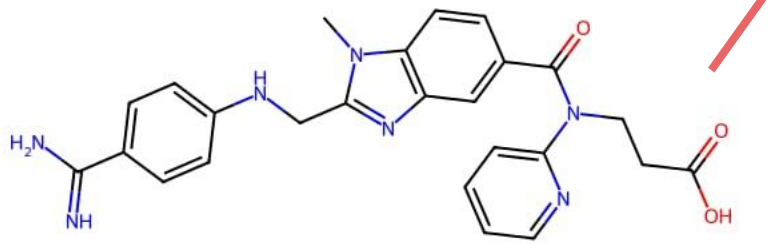
https://arena.ai/

https://www.espn.com/mlb/story/_/id/48201228/team-usa-beats-canada-set-world-baseball-classic-semifinal-showdown-vs-dominican-republic

Why RAG?

- Domain Knowledge





what is the name of this molecule:
CCOCC(=O)CN[C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1

The SMILES you gave:

```
CCOCC(=O)CN[C@@H](C(=O)N1CC[C@H]1C(=O)NCc1ccc(/C(N)=N/O)cc1)C1CCCCC1
```

corresponds to the **direct thrombin inhibitor drug**:

Dabigatran

More precisely, the structure matches **dabigatran (the active drug)** rather than the orally used prodrug **Dabigatran etexilate**.

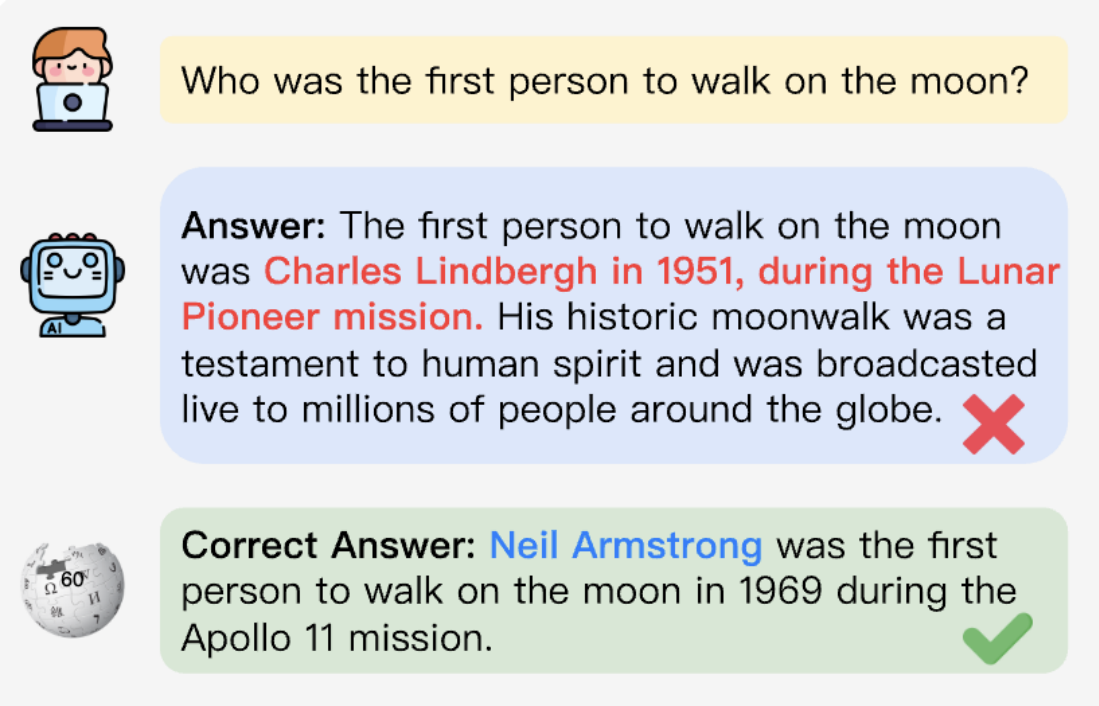
Key structural features that identify it

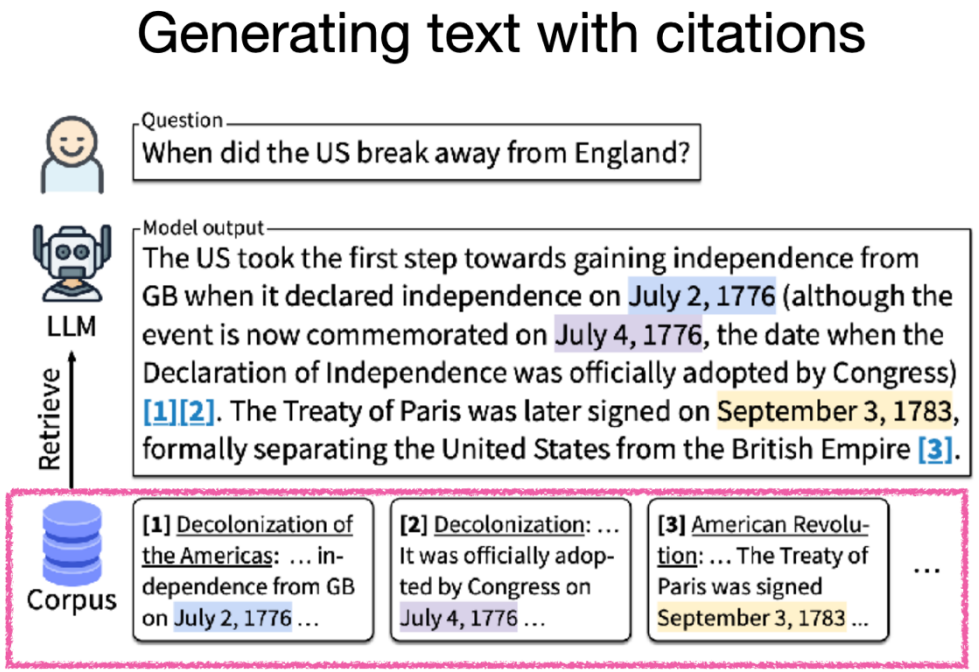
From the SMILES:

- **Amidoxime group**: `/C(N)=N/O`
  - characteristic benzamidine bioisostere used in thrombin inhibitors
- **p-substituted benzyl amide**
- **proline-like tertiary amide ring**: `N1CC[C@H]1`
- **cyclohexyl substituent**: `C1CCCCC1`

Why RAG?

- Hallucination





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

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023

Why RAG?

- Private data

Extracting Training Data from Large Language Models

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

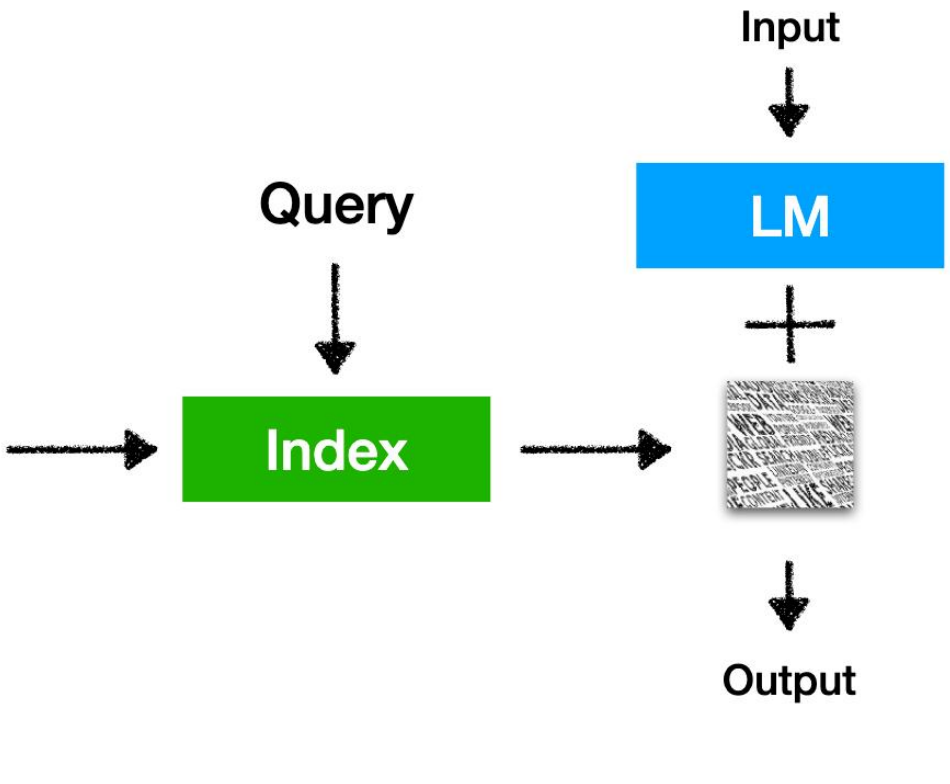
Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	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

Why RAG?

- Potentially leverage other modalities
  - Knowledge base
  - Tabular data
  - Graphs
  - …

Retrieval-Augmented Generation (RAG)

Datastore

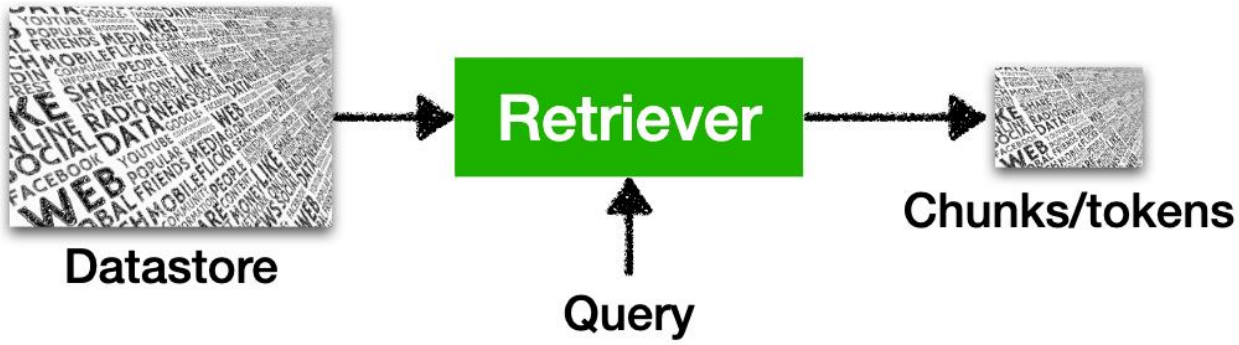


Retrieval-Augmented Generation (RAG)

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

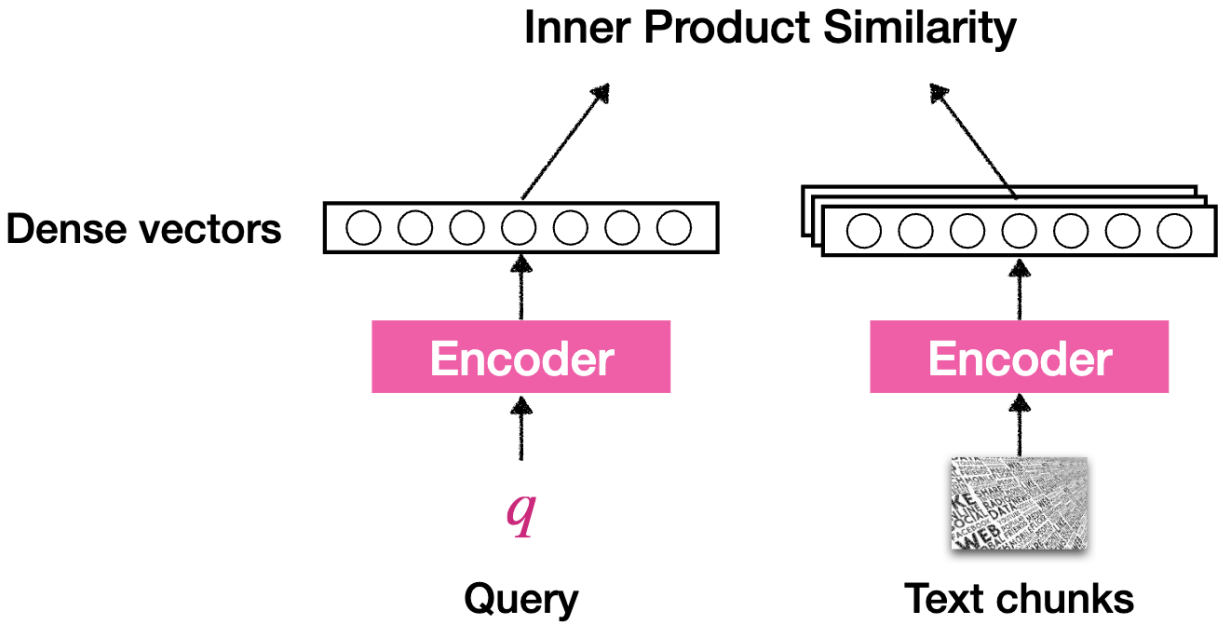
- Training language models

- Training retrieval models



How to Train A Retriever?

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



Dense Passage Retrieval for Open-Domain Question Answering, 2020

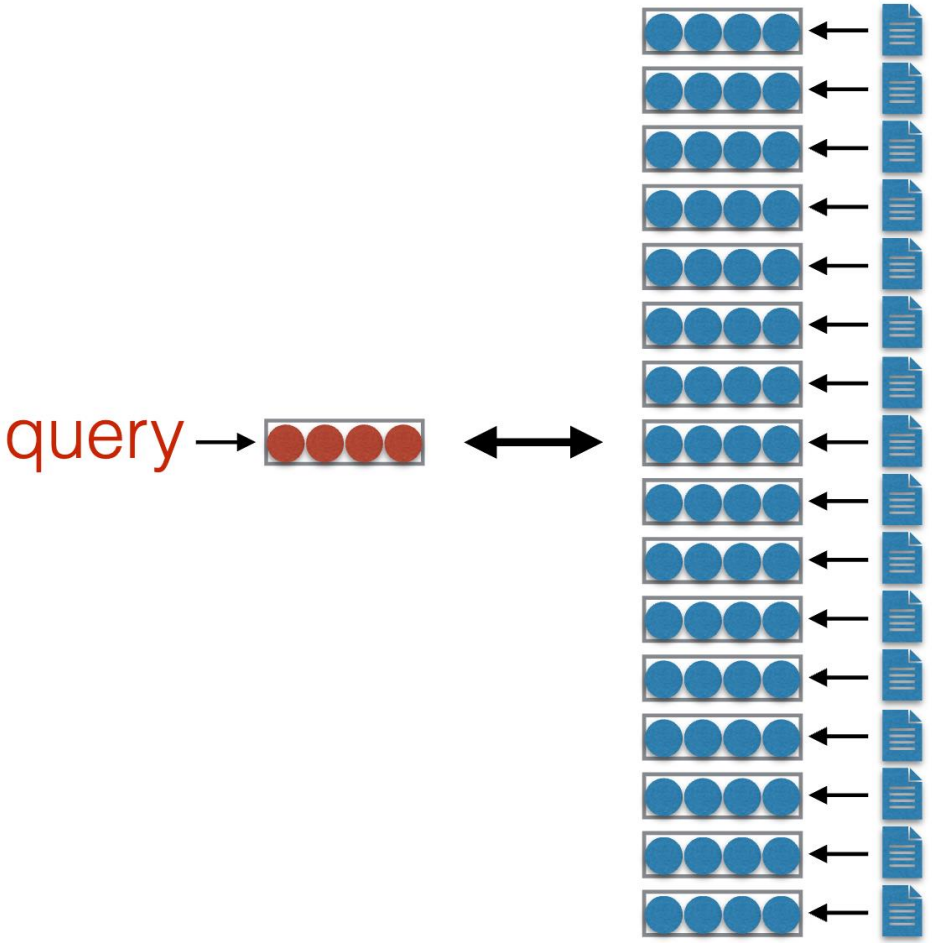
Dense Passage Retrieval (DPR)

Similarity between two sentences

We will go hiking if tomorrow is a sunny day.

If it is sunny tomorrow, we will go hiking.

Similarity between query and documents

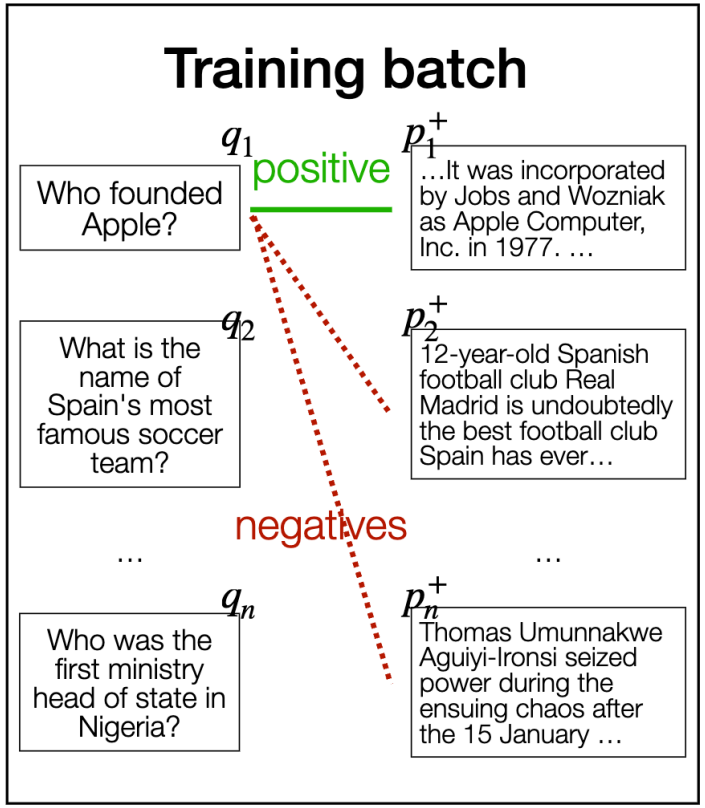


Dense Passage Retrieval for Open-Domain Question Answering, 2020

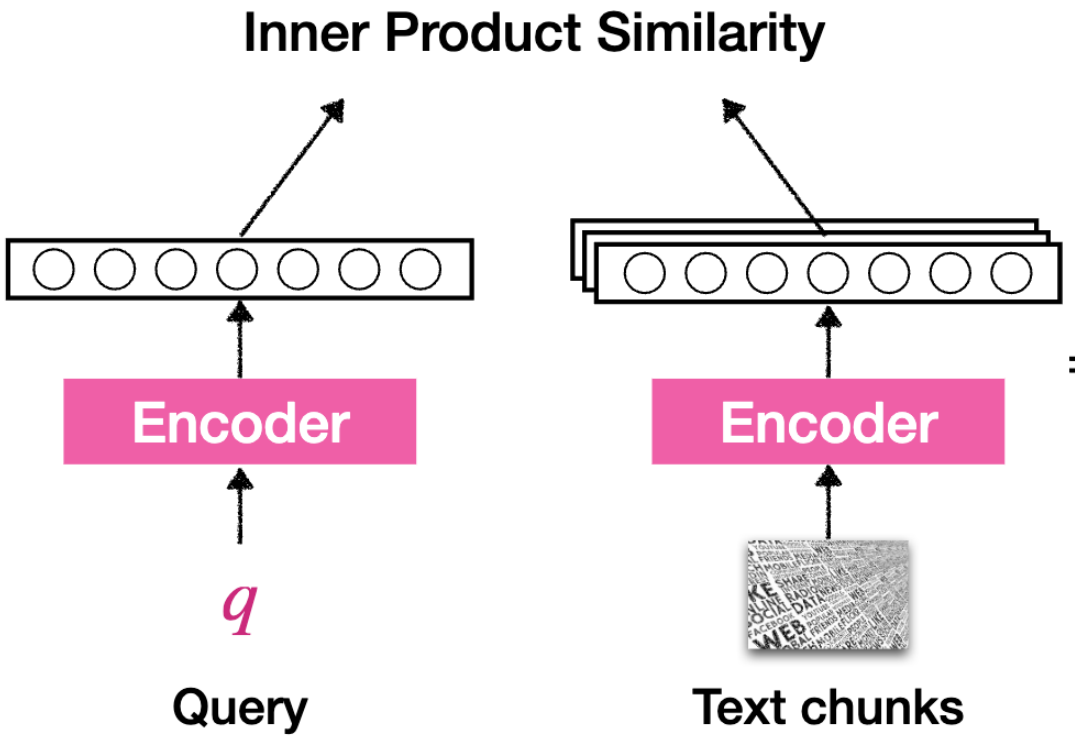
Dense Passage Retrieval (DPR)

$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

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



Dense Passage Retrieval (DPR)

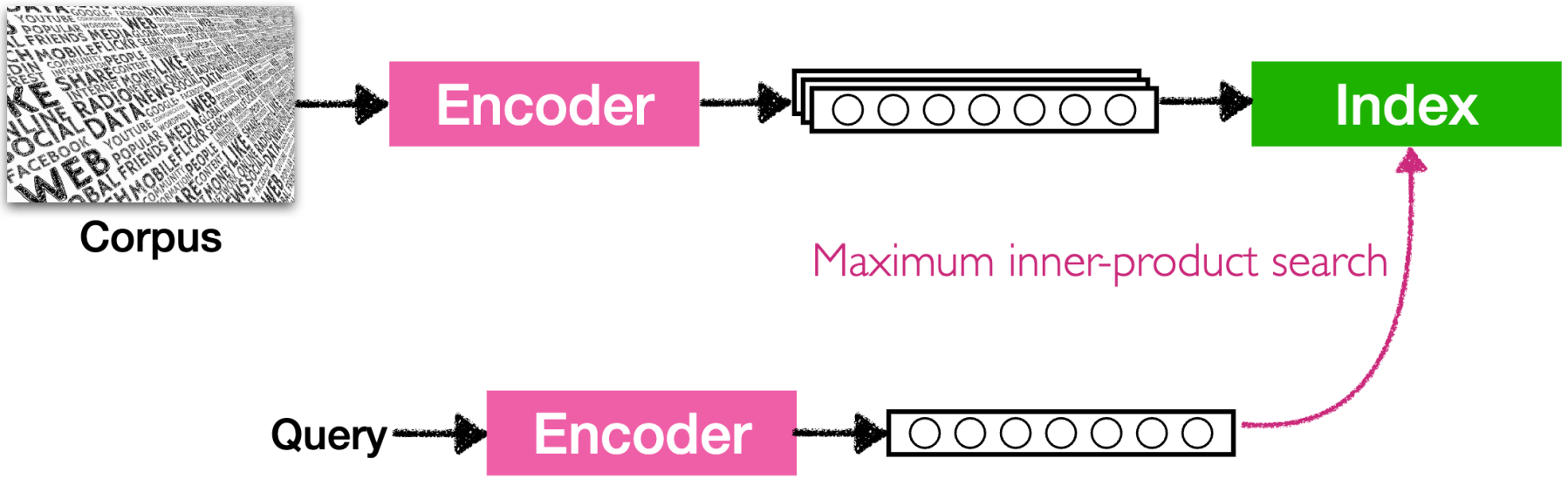


$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

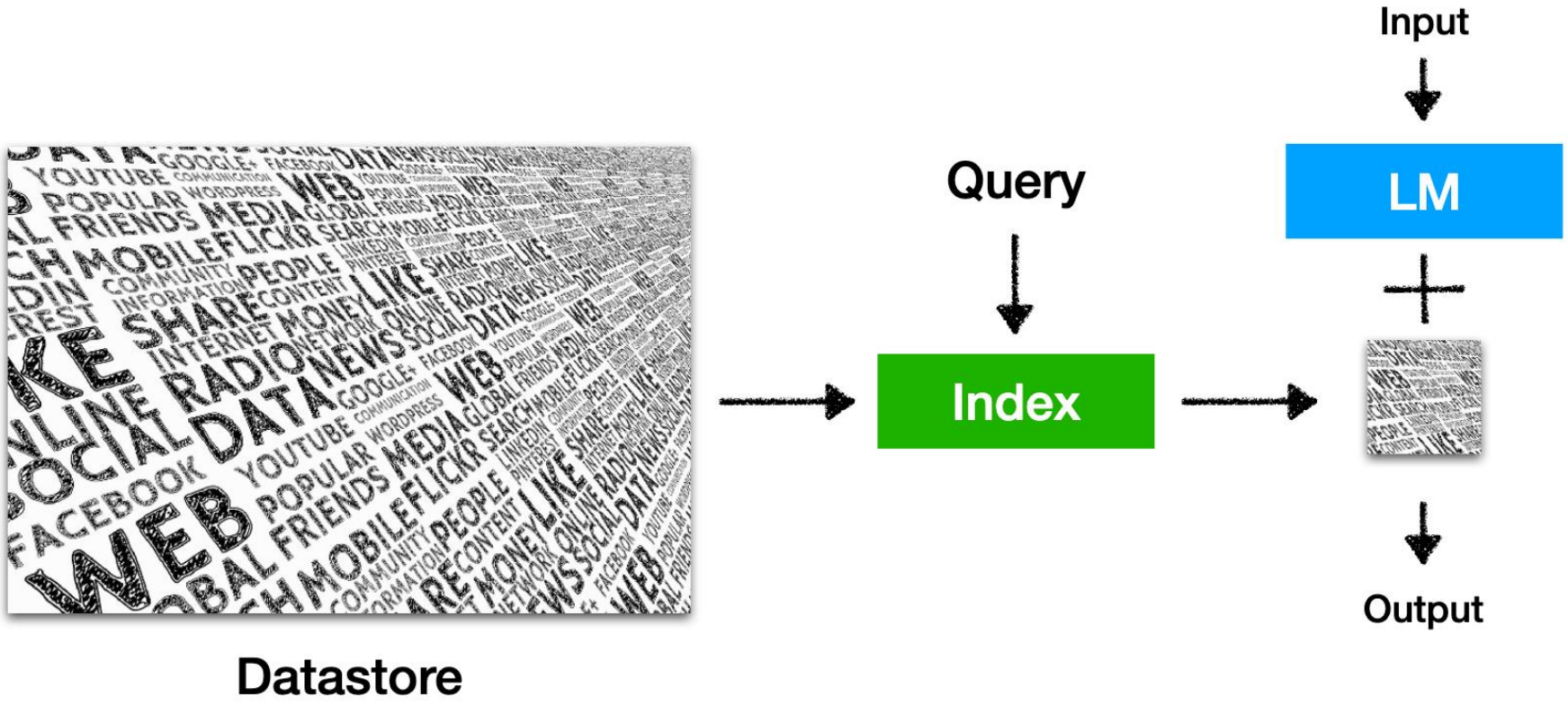
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\text{sim}(q, p_j^-))}$$

Contrastive learning

Dense Passage Retrieval (DPR)



Retrieval-Augmented Generation (RAG)

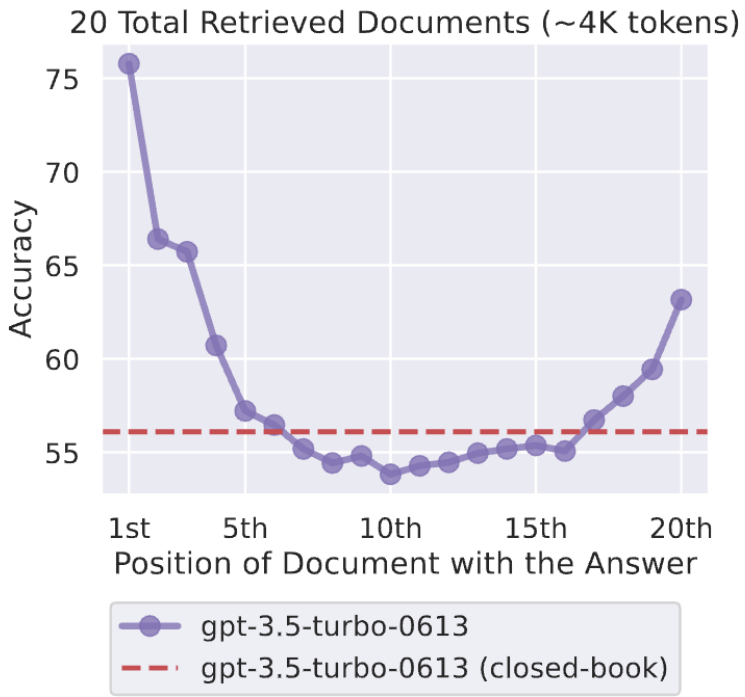


Challenges with RAG

- Longer input text
  - Length generalization
  - KV cache
- The lost-in-the-middle problem

The Lost-in-the-Middle Problem

Retrieved Documents



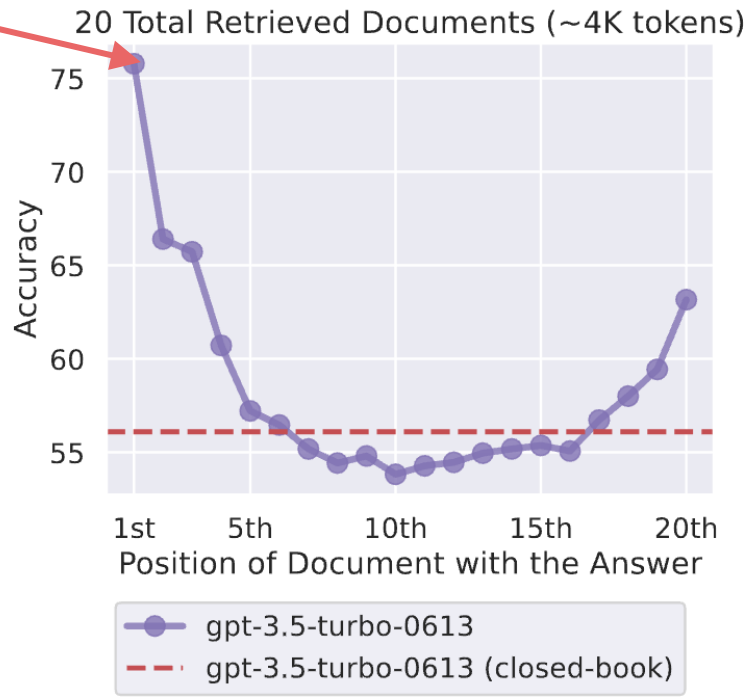
Lost in the Middle: How Language Models Use Long Contexts, 2023

The Lost-in-the-Middle Problem

Ground Truth

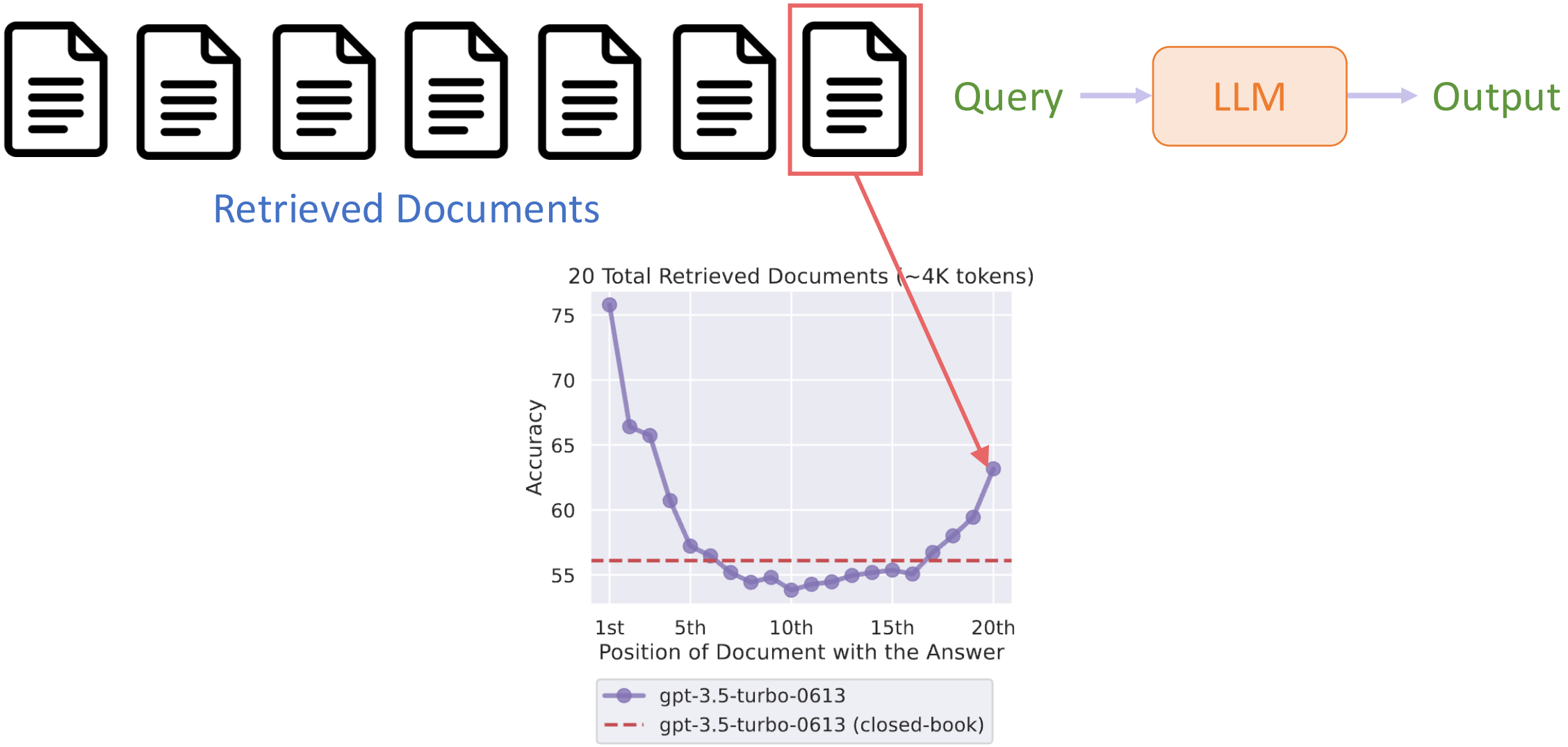
Retrieved Documents

Query → LLM → Output



Lost in the Middle: How Language Models Use Long Contexts, 2023

The Lost-in-the-Middle Problem

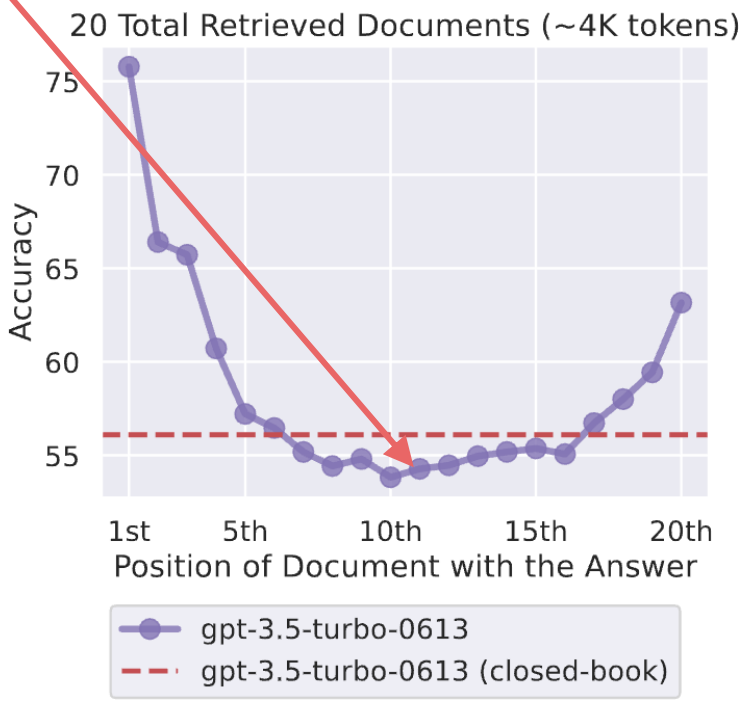


Lost in the Middle: How Language Models Use Long Contexts, 2023

The Lost-in-the-Middle Problem

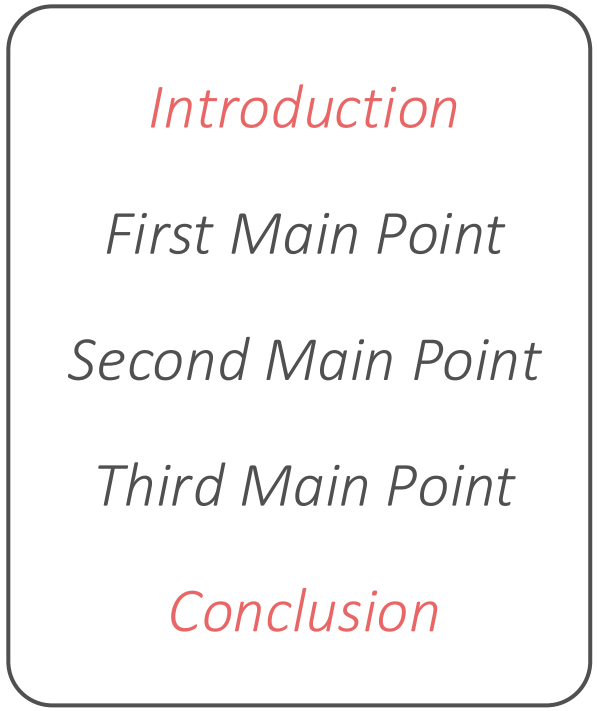
Retrieved Documents

Query → LLM → Output

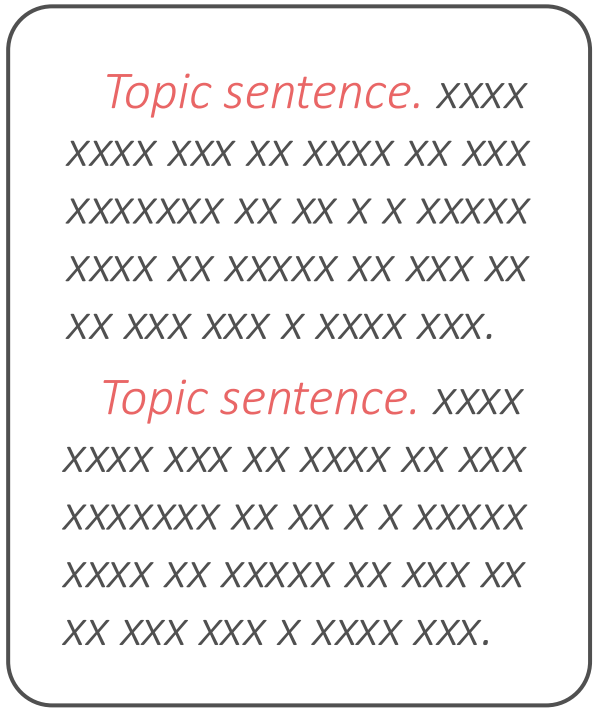


Lost in the Middle: How Language Models Use Long Contexts, 2023

Reasons for Positional Bias: Pre-Training Data



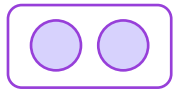
The 5 Paragraph Essay Outline



Topic Sentence

Reasons for Positional Bias: Attention Mechanism

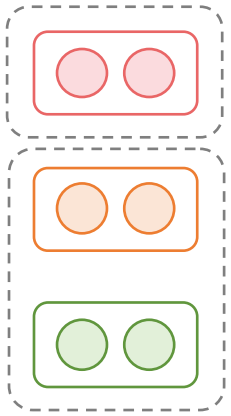
Output



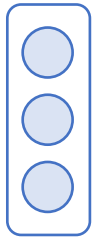
$q_i = W^Q x_i$

$k_i = W^K x_i$

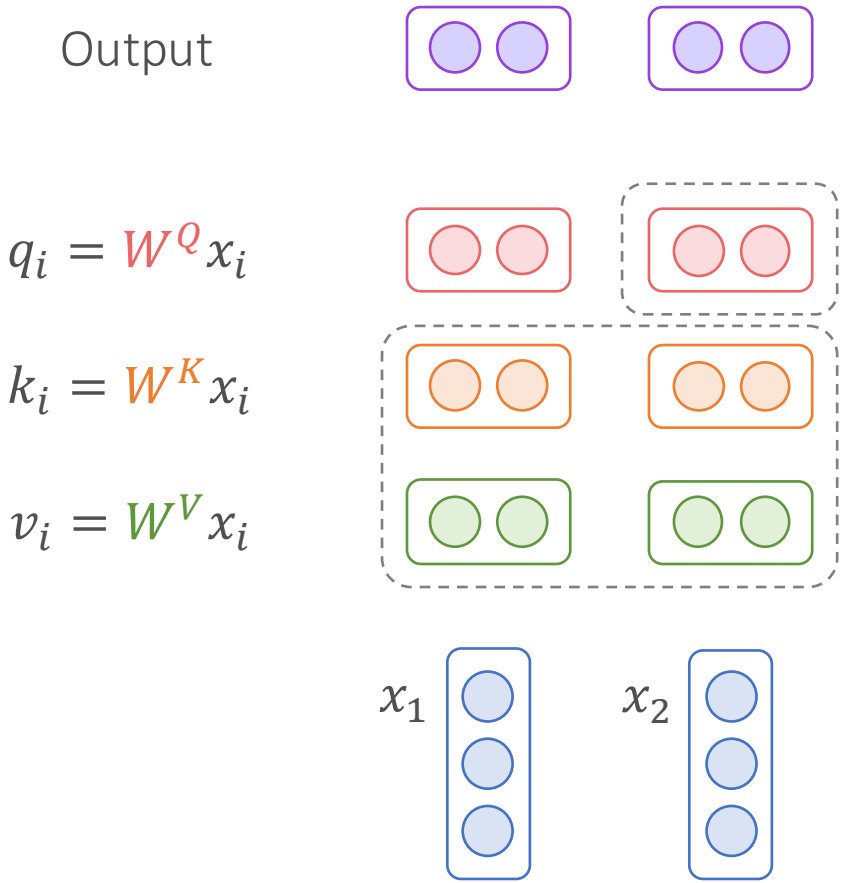
$v_i = W^V x_i$



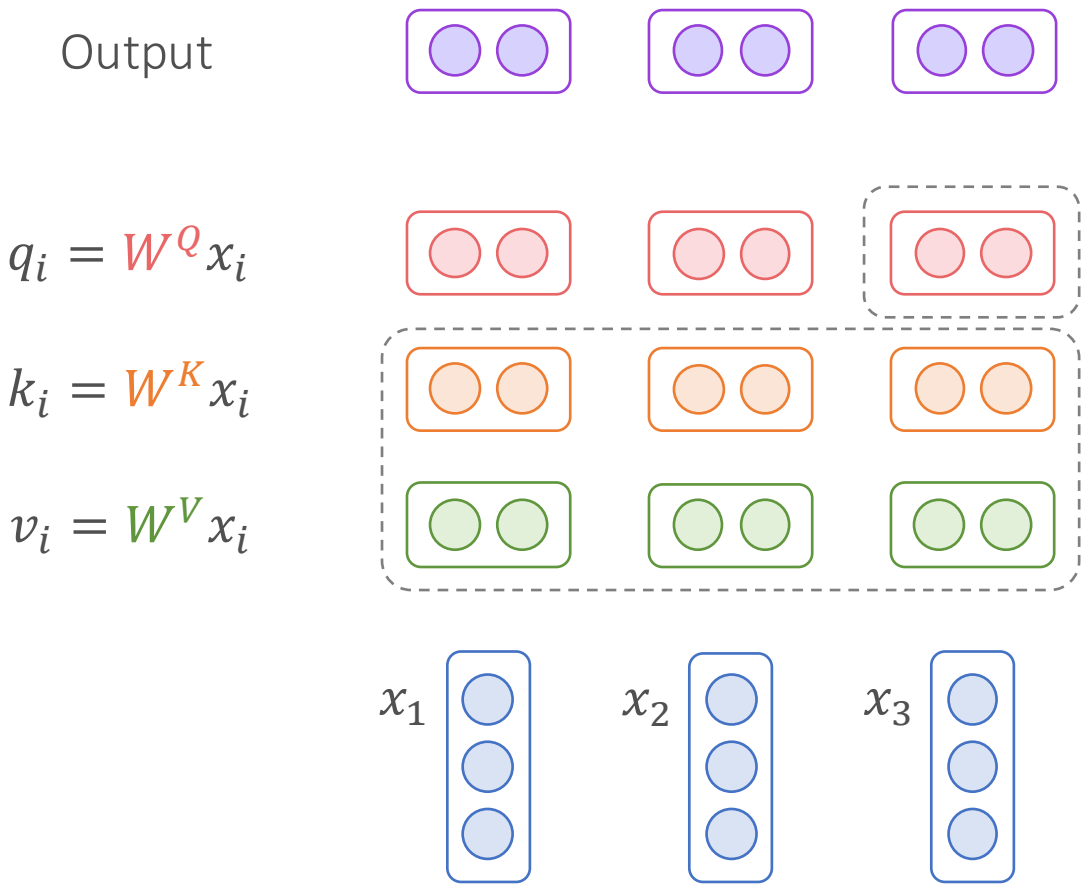
x_1



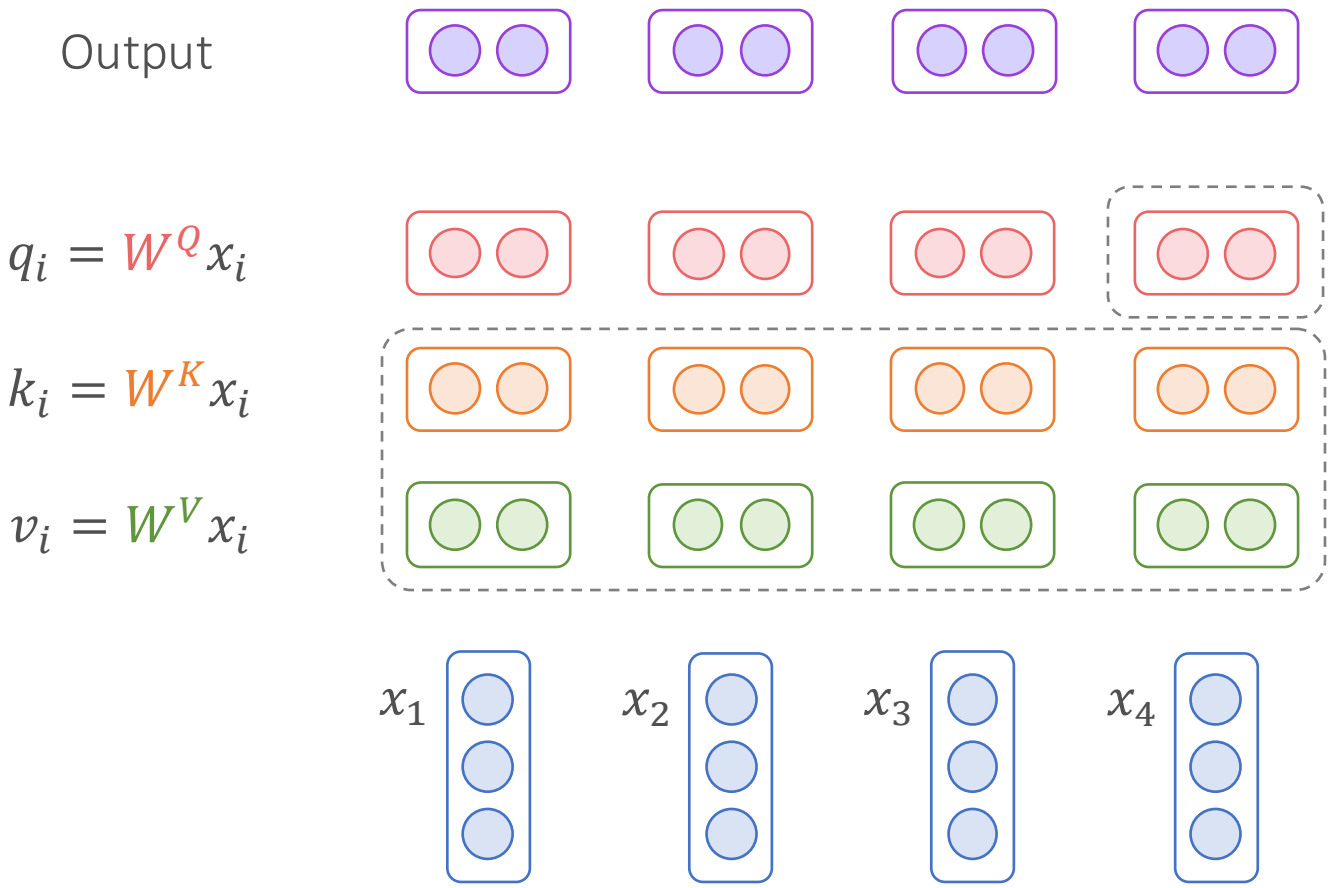
Reasons for Positional Bias: Attention Mechanism



Reasons for Positional Bias: Attention Mechanism



Reasons for Positional Bias: Attention Mechanism



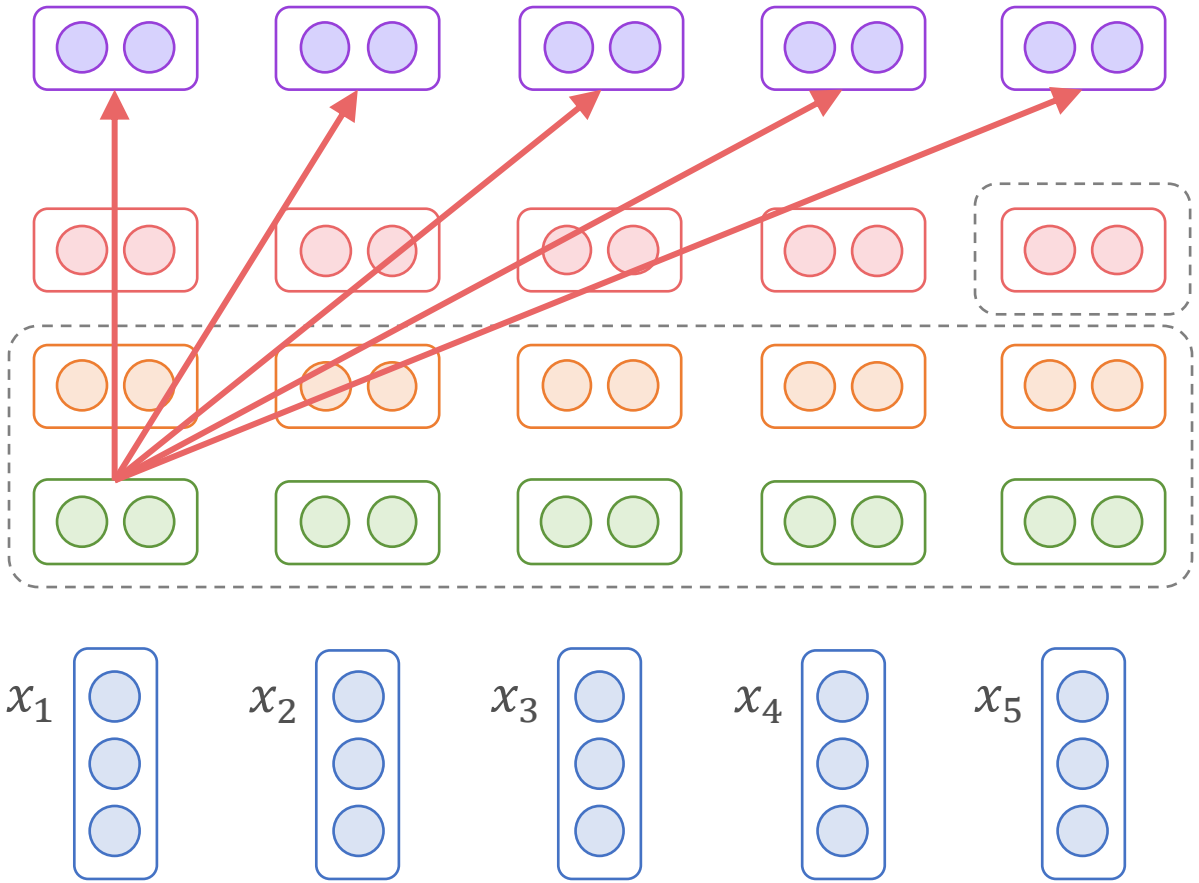
Reasons for Positional Bias: Attention Mechanism

Output

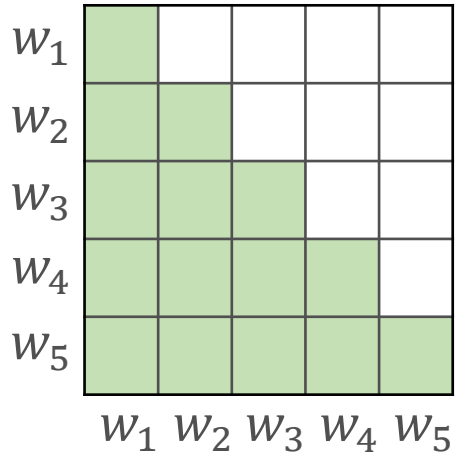
$q_i = W^Q x_i$

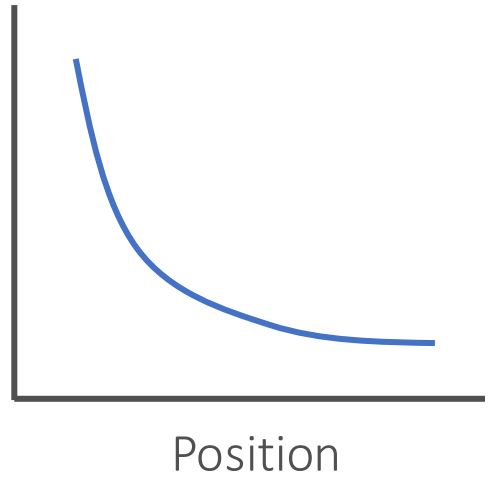
$k_i = W^K x_i$

$v_i = W^V x_i$



Causal Attention Mask



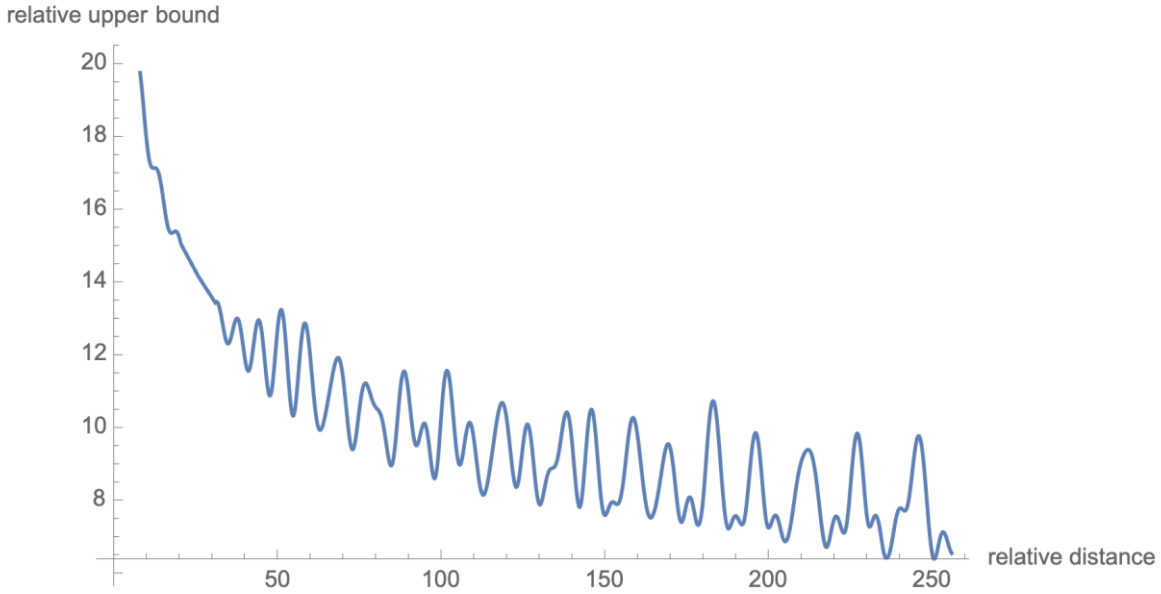


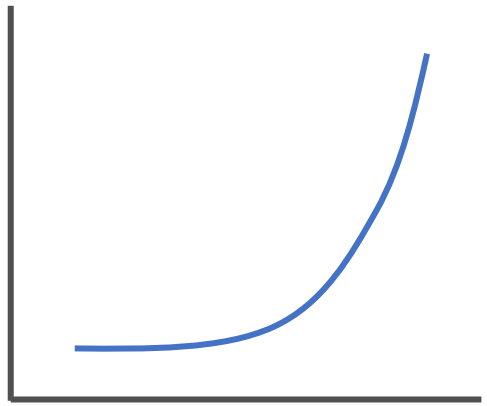
Reasons for Positional Bias: Positional Encoding

Rotary Position Embedding (RoPE)

$$\boldsymbol{q}_m = f_q(\boldsymbol{x}_m, m)$$
$$\boldsymbol{k}_n = f_k(\boldsymbol{x}_n, n)$$
$$\boldsymbol{v}_n = f_v(\boldsymbol{x}_n, n)$$

$$f_q(\boldsymbol{x}_m, m) = (\boldsymbol{W}_q \boldsymbol{x}_m) e^{im\theta}$$
$$f_k(\boldsymbol{x}_n, n) = (\boldsymbol{W}_k \boldsymbol{x}_n) e^{in\theta}$$
$$\langle f_q(\boldsymbol{x}_m, m), f_k(\boldsymbol{x}_n, n) \rangle = \mathrm{Re}[(\boldsymbol{W}_q \boldsymbol{x}_m)(\boldsymbol{W}_k \boldsymbol{x}_n)^* e^{i(m-n)\theta}]$$





Position

RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021

Combine All Together

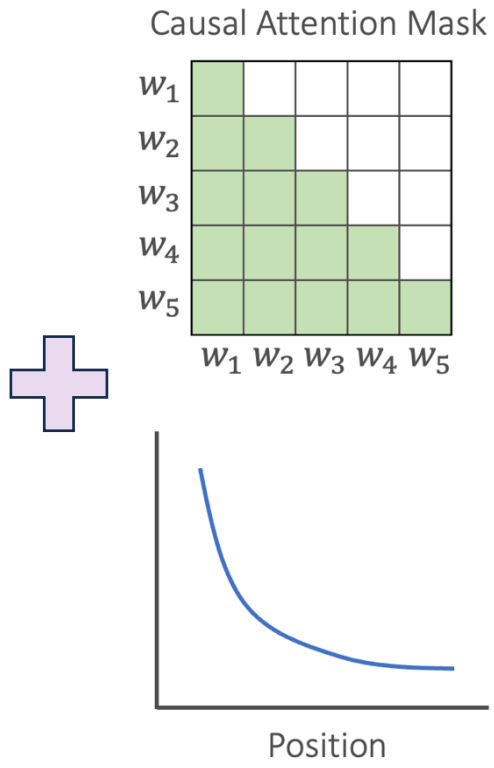
Introduction

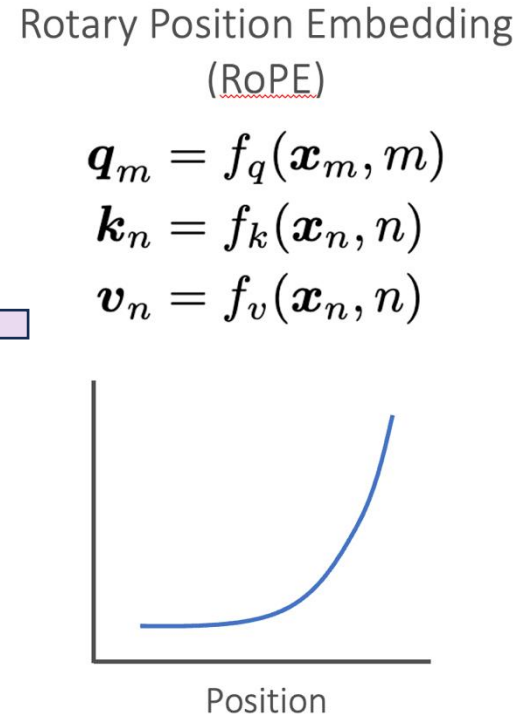
First Main Point

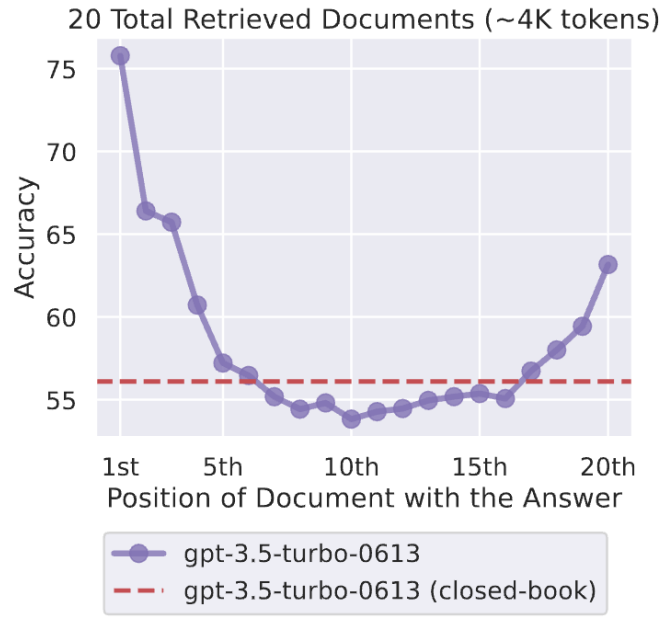
Second Main Point

Third Main Point

Conclusion







Lecture Plan

- Text Similarity
- Retrieval-Augmented Generation