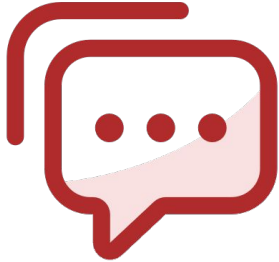


# Announcement

---

- HW2 deadline is extended to next Tuesday Oct. 1<sup>st</sup>
- HW3 release date is moved to Oct. 1<sup>st</sup> accordingly



# Audience Q&A

① Start presenting to display the audience questions on this slide.

# Tool Use and Embedding Learning

---

## **Large Language Models: Methods and Applications**

Daphne Ippolito and Chenyan Xiong

# Learning Objectives

---

Tool use:

- Understand why and when to aid LLMs with Tools
- Learn how to adapt LLMs to use tools (HW3)

Embedding Learning:

- Learn how to finetune LLMs into embedding models (HW3)
- Understand state-of-the-arts embedding learning methods

# Tool Use

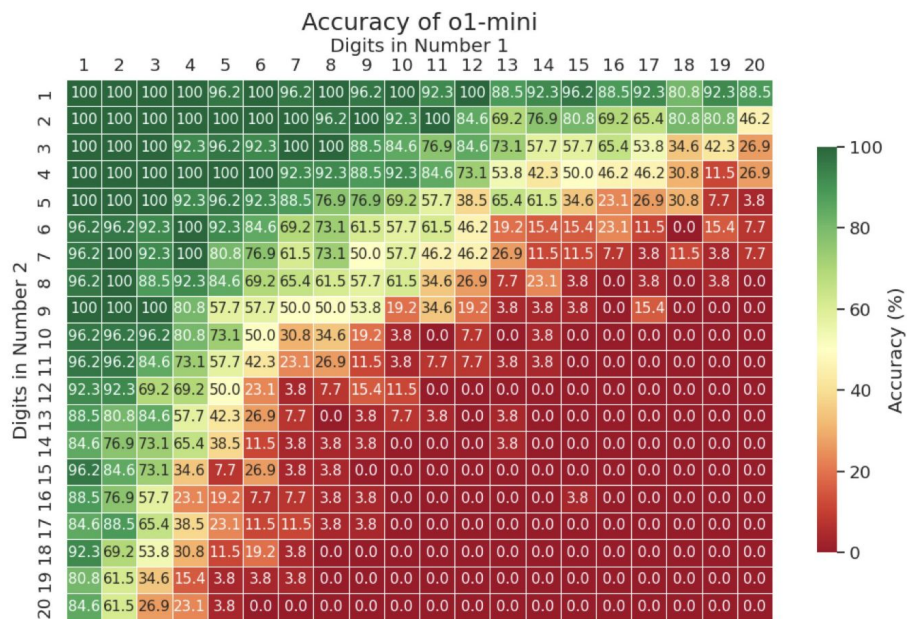
---

Why LLM needs tools and what are they

Make LLMs effective tool users

# Why Tools

LLMs are not the solution for everything. (Not AGI yet. Surprise?)



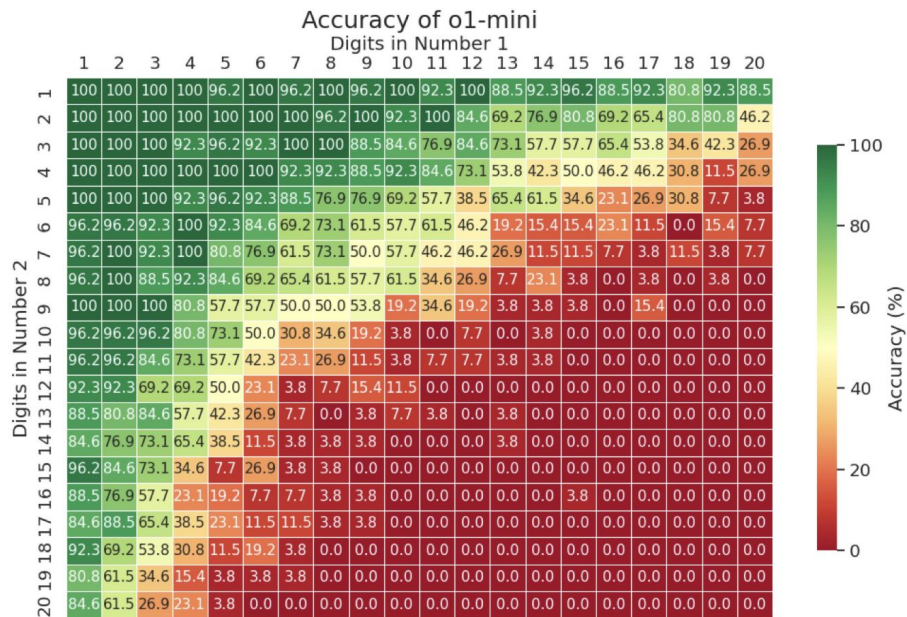
O1 cannot solve multiplications of 10+ digits...

Multiplication Accuracy of OpenAI O1 (Yuantian Deng, X)



# Why Tools

LLMs are not the solution for everything. (Not AGI yet. Surprise?)



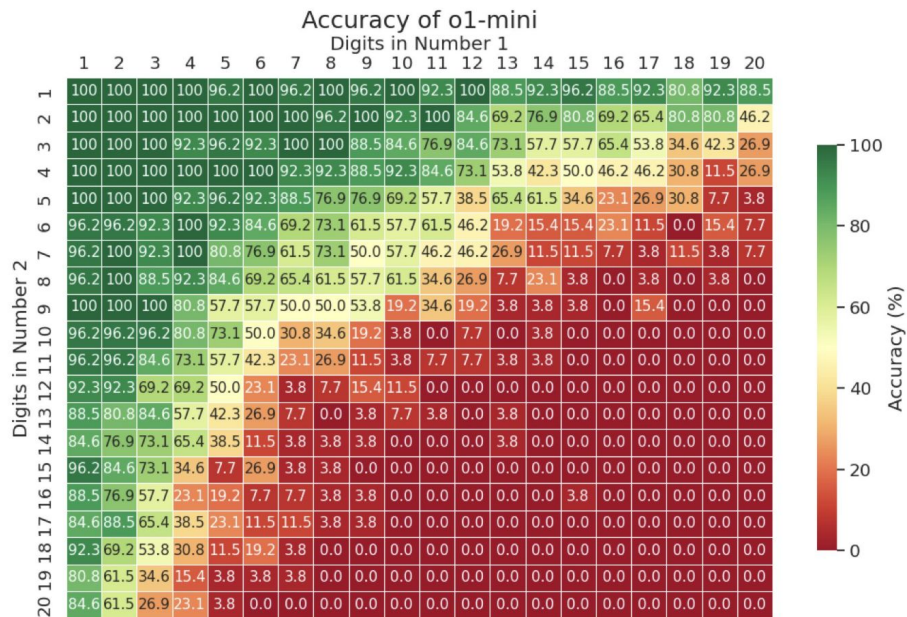
O1 cannot solve multiplications of 10+ digits...

- But does it really need to?
- We humans can use a calculator to do it...
- So do LLMs

Multiplication Accuracy of OpenAI O1 (Yuantian Deng, X)

# Why Tools

LLMs are not the solution for everything. (Not AGI yet. Surprise?)



O1 cannot solve multiplications of 10+ digits...

- But does it really need to?
- We humans can use a calculator to do it...
- So do LLMs

“Tool”

Multiplication Accuracy of OpenAI O1 (Yuantian Deng, X)





# What tasks LLMs are bad at?

① Start presenting to display the poll results on this slide.

# Why Tools: Things LLMs are Bad At

---

Numerical/symbolic operations

1. Calculation
2. Logic deduction
3. Exact operations

Knowledge not in their pretraining corpus

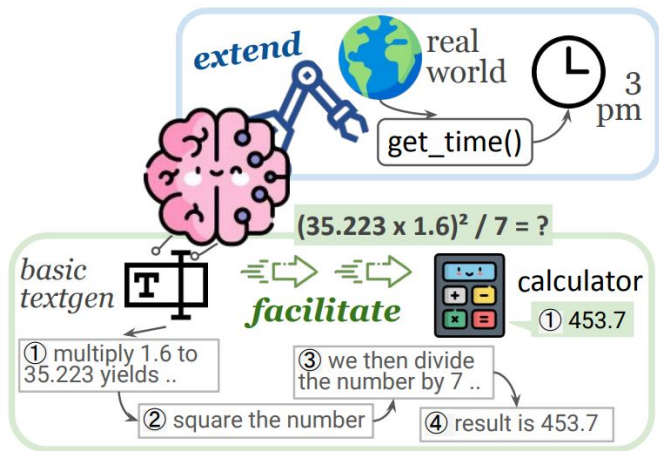
1. Tail factual knowledge
2. New information
3. Private information

Interaction with the external world

1. Non natural language interfaces
2. Physical world
3. Environmental information (time, e.g.)

# What is a Tool for LLM?

Definition: An LM-used tool is a function interface to a computer program that runs externally to the LM, where the LM generates the function calls and input arguments in order to use the tool [1]








LLM with Tools [1]

**A tool is:**

- A Computer Program
- External to the LM
- Used through generated function calls

# What are Tools

Category	Example Tools
 Knowledge access	<code>sql_executor(query: str) -&gt; answer: any</code> <code>search_engine(query: str) -&gt; document: str</code> <code>retriever(query: str) -&gt; document: str</code>
 Computation activities	<code>calculator(formula: str) -&gt; value: int   float</code> <code>python_interpreter(program: str) -&gt; result: any</code> <code>worksheet.insert_row(row: list, index: int) -&gt; None</code>
 Interaction w/ the world	<code>get_weather(city_name: str) -&gt; weather: str</code> <code>get_location(ip: str) -&gt; location: str</code> <code>calendar.fetch_events(date: str) -&gt; events: list</code> <code>email.verify(address: str) -&gt; result: bool</code>
 Non-textual modalities	<code>cat_image.delete(image_id: str) -&gt; None</code> <code>spotify.play_music(name: str) -&gt; None</code> <code>visual_qa(query: str, image: Image) -&gt; answer: str</code>
 Special-skilled LMs	<code>QA(question: str) -&gt; answer: str</code> <code>translation(text: str, language: str) -&gt; text: str</code>

## Common Tool Categories and Examples [1]

# How to Enable LLMs to Use Tools?

Out of 1400 participants, 400 (or `Calculator(400 / 1400)`  
→ `0.29`) 29%) passed the test.

The Brown Act is California's law `WikiSearch("Brown Act")` → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

## Example Tool Usage from LLMs [2]:

- **Function calls** predicted by LLMs
- **Tool execute that function call**
- **Returned results as part of LLMs context**

# How to Enable LLMs to Use Tools?

Out of 1400 participants, 400 (or `Calculator(400 / 1400)`  
→ `0.29`) 29%) passed the test.

The Brown Act is California's law `WikiSearch("Brown Act")` → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

## Example Tool Usage from LLMs [2]:

- **Function calls** predicted by LLMs
- **Tool execute that function call**
- **Returned results as part of LLMs context**

## LLMs need to learn:

- **When to use tools**
- **Which tool to use**
- **How to incorporate tool's results**



# How to Enable LLMs to Use Tools?

Out of 1400 participants, 400 (or `Calculator(400 / 1400)`  
→ 0.29) 29%) passed the test.

The Brown Act is California's law `WikiSearch("Brown Act")` → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

## Example Tool Usage from LLMs [2]:

- **Function calls** predicted by LLMs
- **Tool execute** that function call
- **Returned results** as part of LLMs context

## LLMs need to learn:

- **When to use tools**
- **Which tool to use**
- **How to incorporate tool's results**

Pretraining  
(with Code)

Standard Pretraining  
with Program Codes  
as Part of the Corpus

# How to Enable LLMs to Use Tools?

Out of 1400 participants, 400 (or `Calculator(400 / 1400)`  
→ 0.29) 29%) passed the test.

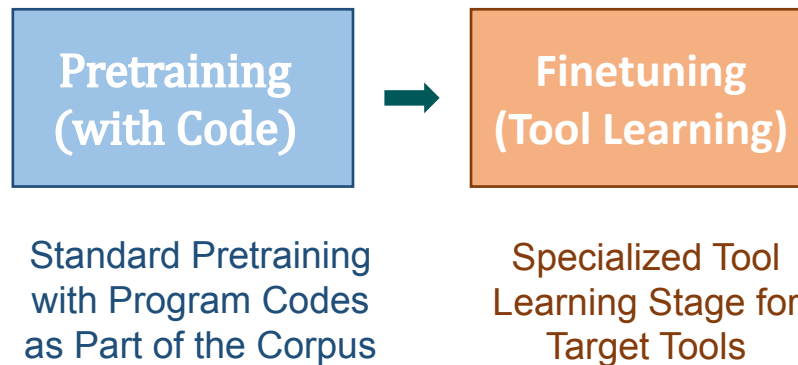
The Brown Act is California's law `WikiSearch("Brown Act")` → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

## Example Tool Usage from LLMs [2]:

- **Function calls** predicted by LLMs
- **Tool execute** that function call
- **Returned results** as part of LLMs context

## LLMs need to learn:

- **When to use tools**
- **Which tool to use**
- **How to incorporate tool's results**



# How to Enable LLMs to Use Tools?

Out of 1400 participants, 400 (or `Calculator(400 / 1400)`  
→ 0.29) 29%) passed the test.

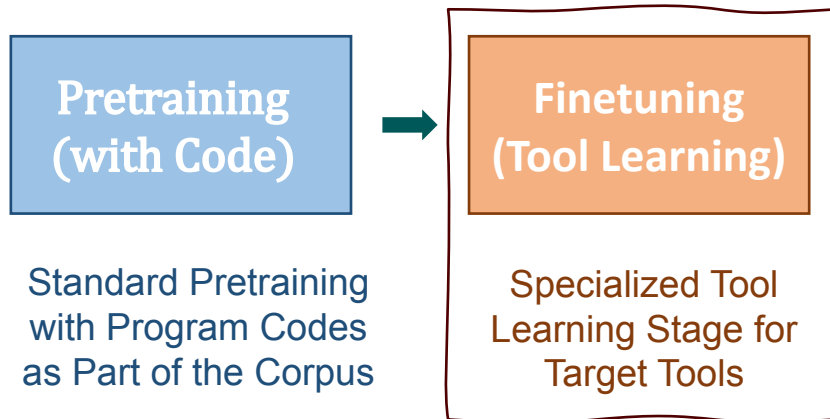
The Brown Act is California's law `WikiSearch("Brown Act")` → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

## Example Tool Usage from LLMs [2]:

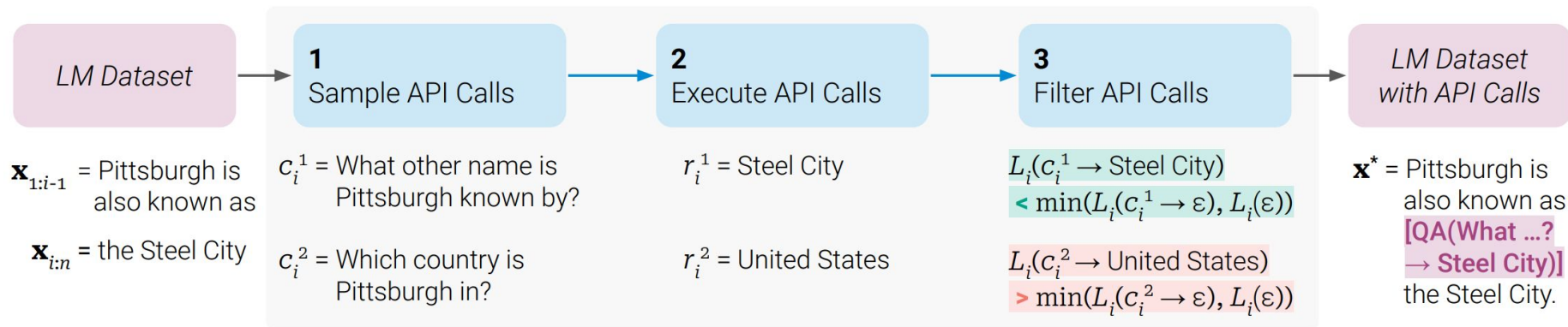
- **Function calls** predicted by LLMs
- **Tool execute** that function call
- **Returned results** as part of LLMs context

## LLMs need to learn:

- **When to use tools**
- **Which tool to use**
- **How to incorporate tool's results**



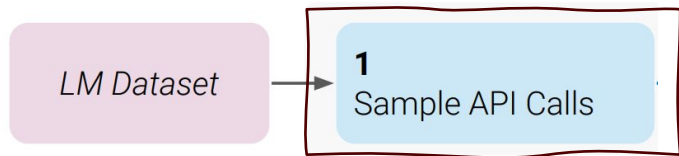
# Make LLMs Tool Users



Key idea (Toolformer [2]): Curate a corpus with tool calls and finetune LLMs with it

- Prompts an LLM to produce tool calls and execute them
- Filter the tool calls based on their effectiveness

# Make LLMs Tool Users: Sample API Calls



Write customized prompts to ask the LLM itself to generate potential API calls for each tool:

- Prompt it to generate APIs calls for positions likely to need API calls (e.g., digits for calculator)

*Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:*

**Input:** Joe Biden was born in Scranton, Pennsylvania.

**Output:** Joe Biden was born in [QA("Where was Joe Biden born?")] Scranton, [QA("In which state is Scranton?")] Pennsylvania.

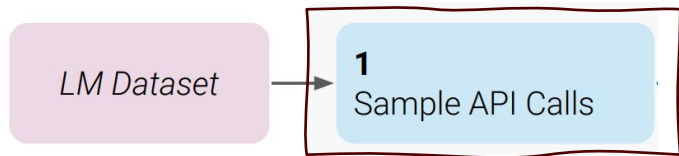
**Input:** Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

**Output:** Coca-Cola, or [QA("What other name is Coca-Cola known by?")] Coke, is a carbonated soft drink manufactured by [QA("Who manufactures Coca-Cola?")] the Coca-Cola Company.

**Input:** x

**Output:**

# Make LLMs Tool Users: Sample API Calls



Write customized prompts to ask the LLM itself to generate potential API calls for each tool:

- Prompt it to generate APIs calls for positions likely to need API calls (e.g., digits for calculator)
- Keep top K positions in a sequence with highest probability of API calls above a given threshold
- Keep m candidate API calls for each of the chosen position

*Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:*

**Input:** Joe Biden was born in Scranton, Pennsylvania.

**Output:** Joe Biden was born in [QA("Where was Joe Biden born?")] Scranton, [QA("In which state is Scranton?")] Pennsylvania.

**Input:** Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

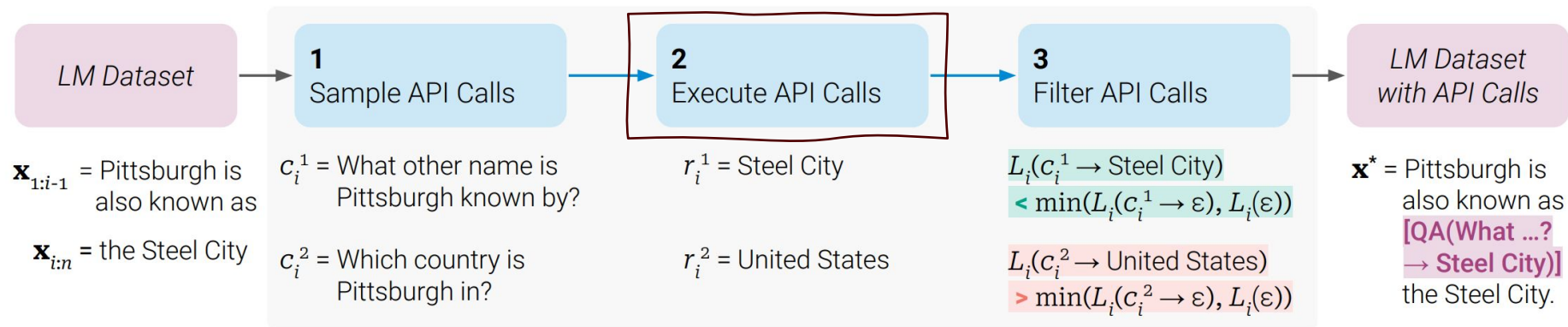
**Output:** Coca-Cola, or [QA("What other name is Coca-Cola known by?")] Coke, is a carbonated soft drink manufactured by [QA("Who manufactures Coca-Cola?")] the Coca-Cola Company.

**Input:** x

**Output:**

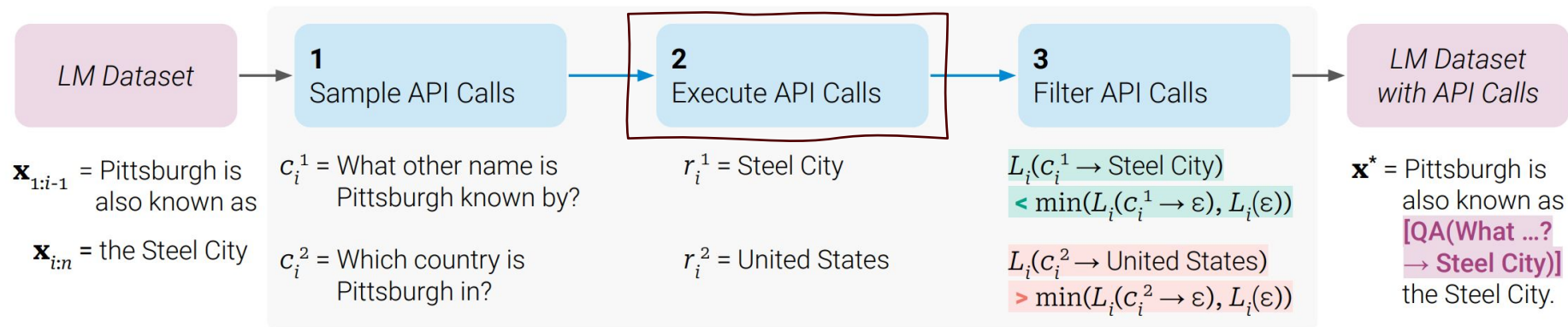


# Make LLMs Tool Users: Execute API Calls



- Perform the tools for the sampled API calls
- Add tool output to the text sequence

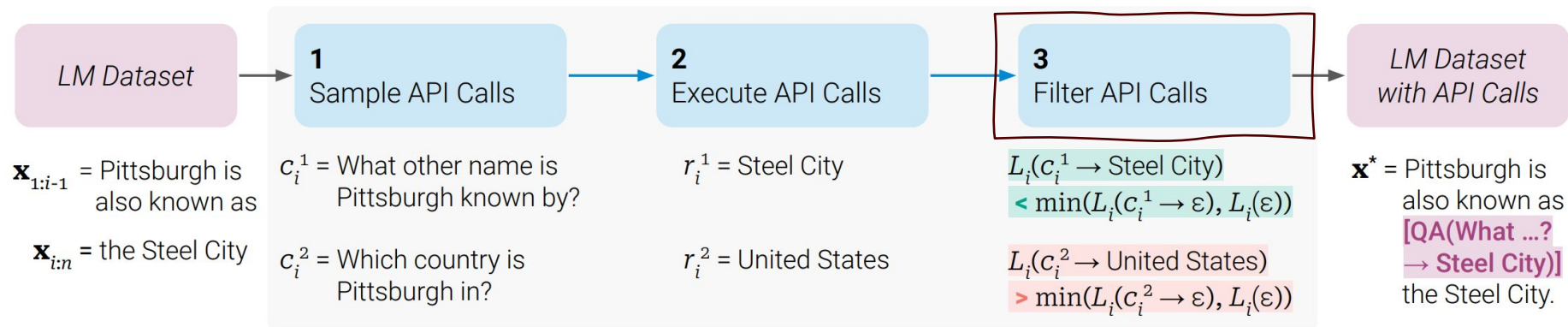
# Make LLMs Tool Users: Execute API Calls



- Perform the tools for the sampled API calls
- Add tool output to the text sequence

The Brown Act is California's law **[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]** that requires legislative bodies, like city councils, to hold their meetings open to the public.

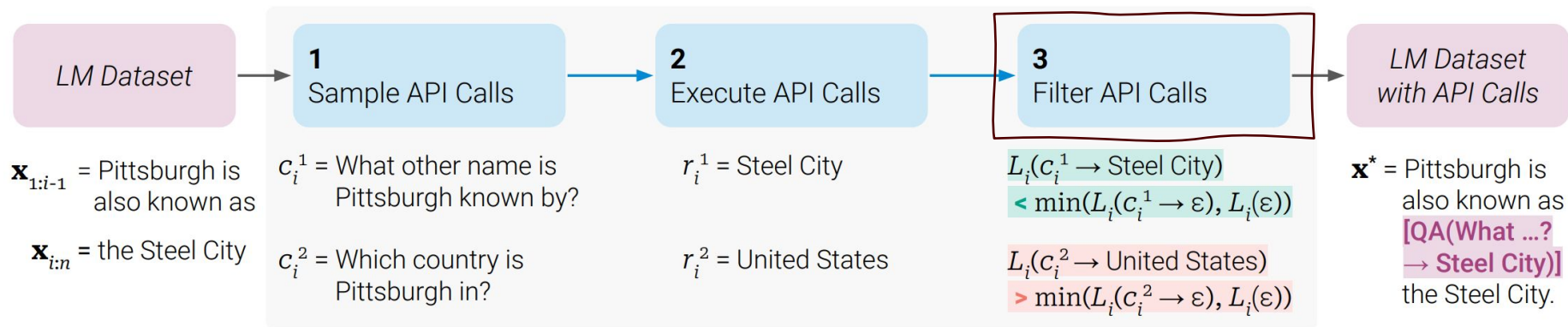
# Make LLMs Tool Users: Filter Non-Useful Tool Calls



What is a useful API call?

- LM performance in the rest tokens improved with the API call versus without
- Filter out not useful ones to remove noise/failed tool uses

# Make LLMs Tool Users: Filter Non-Useful Tool Calls



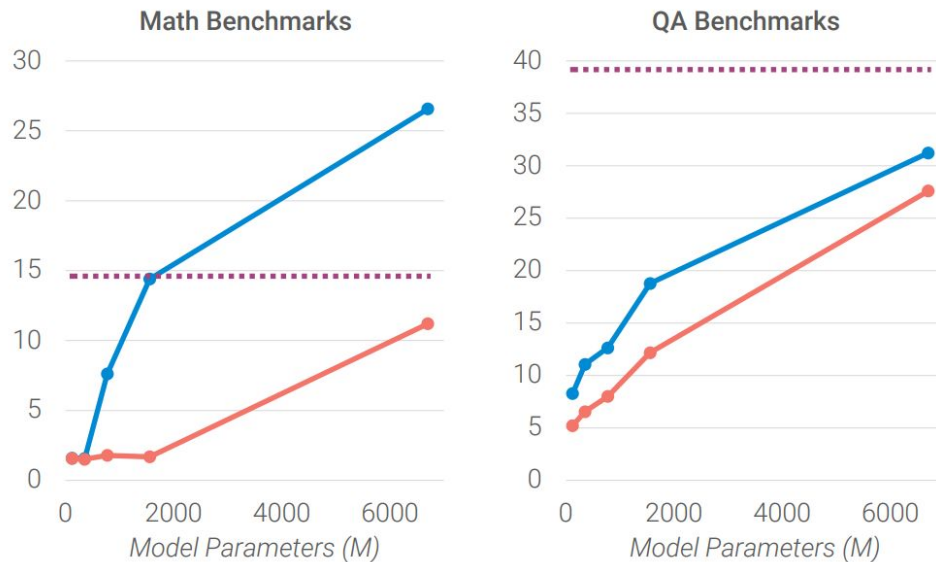
What is a useful API call?

- LM performance in the rest tokens improved with the API call versus without
- Filter out not useful ones to remove noise/failed tool uses

$$L_i(c_i^2 \rightarrow \text{United States}) > \min(L_i(c_i^2 \rightarrow \epsilon), L_i(\epsilon))$$

Performance with API call + output      Performance with API call      Performance without API call

# Tool Usage Performance



## Significantly Improving GPT's Performances [2]

- Using at Max 25k examples per API

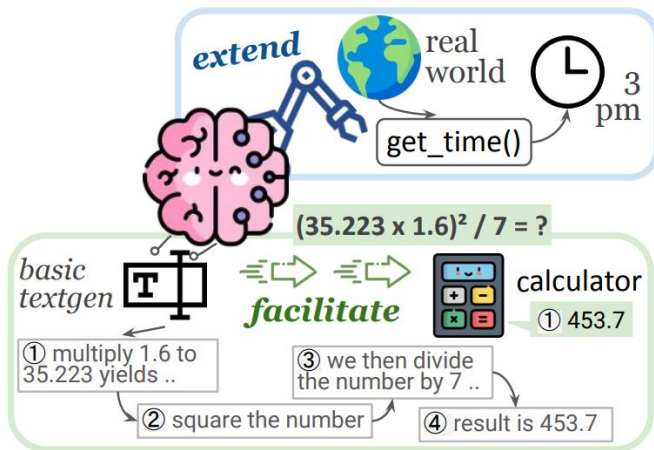
# Example Tool Calls

Example	$L_i^- - L_i^+$	Useful
Os Melhores Escolas em Jersey 2020 <API> <b>MT(Os Melhores Escolas em Jersey) → The Best Schools in Jersey</b> </API> On this page you can search for Universities, Colleges and Business schools in Jersey	0.70	✓
Enjoy these pictures from the <API> <b>Calendar() → Today is Friday, April 19, 2013.</b> </API> Easter Egg Hunt.	0.33	✓
85 patients (23%) were hospitalised alive and admitted to a hospital ward. Of them, <API> <b>Calculator(85 / 23) → 3.70</b> </API> 65% had a cardiac aetiology [...]	-0.02	✗
But hey, after the <API> <b>Calendar() → Today is Saturday, June 25, 2011.</b> </API> Disneyland fiasco with the fire drill, I think it's safe to say Chewey won't let anyone die in a fire.	-0.41	✗
The last time I was with <API> <b>QA(Who was last time I was with?) → The Last Time</b> </API> him I asked what he likes about me and he said he would tell me one day.	-1.23	✗

## Examples of kept and filtered API Calls [2]



# Tool Use: Summary

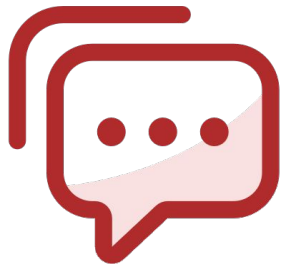


Why LLM needs tools and what are they

- To aid LLMs on tasks beyond their ability
- Mainly tools for knowledge, symbolic, and external environment operations

How to make LLMs effective tool users

- Existing LLMs with coding ability can be prompted to generate noisy tool calls
- Leverage the noisy tool call ability to curate tool use data
- Finetune LLMs on the tool use data to enhance its abilities



# Audience Q&A

① Start presenting to display the audience questions on this slide.

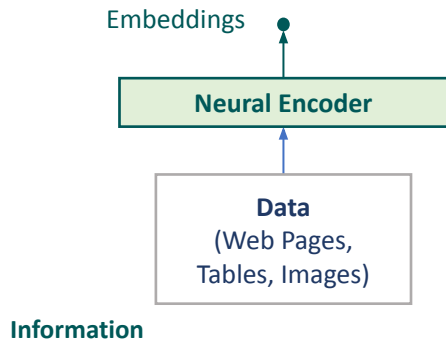


# Embedding Learning

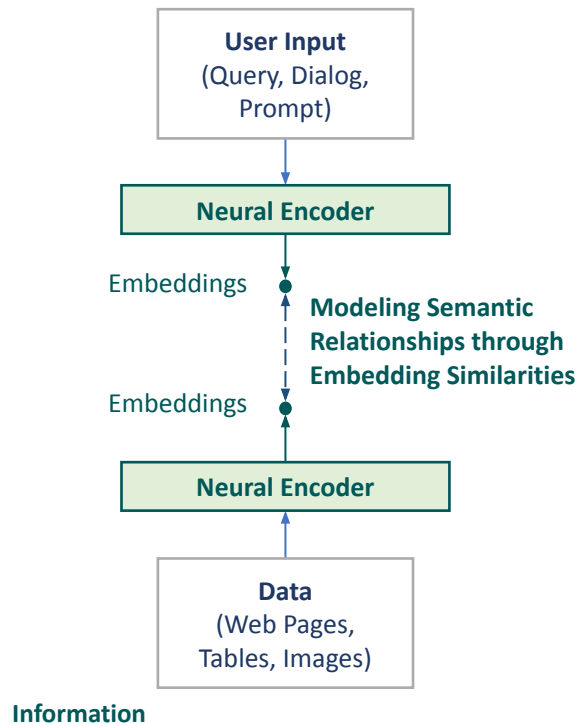
# Embedding Learning: Overview

Embedding Learning: Encode data into an embedding vector

- By finetuned LLMs



# Embedding Learning: Overview



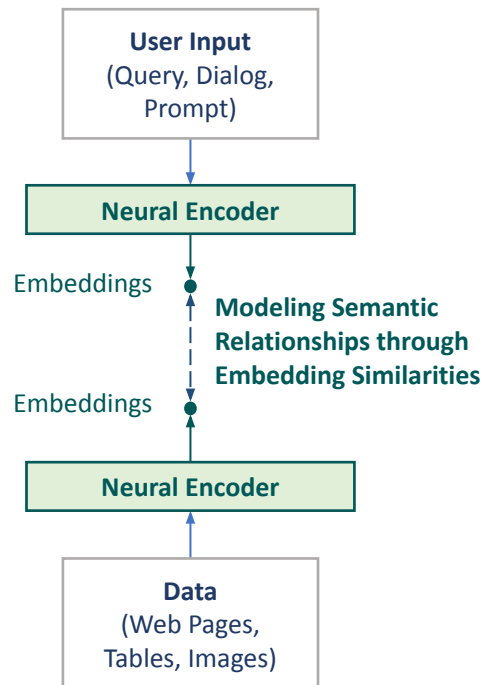
Embedding Learning: Encode data into an embedding vector

- By finetuned LLMs

Capture semantic relationships through distances in the embedding space

- Relevance in search and RAG
- User interests in recommendation systems
- Data similarity in clustering

# Embedding Learning: Why?



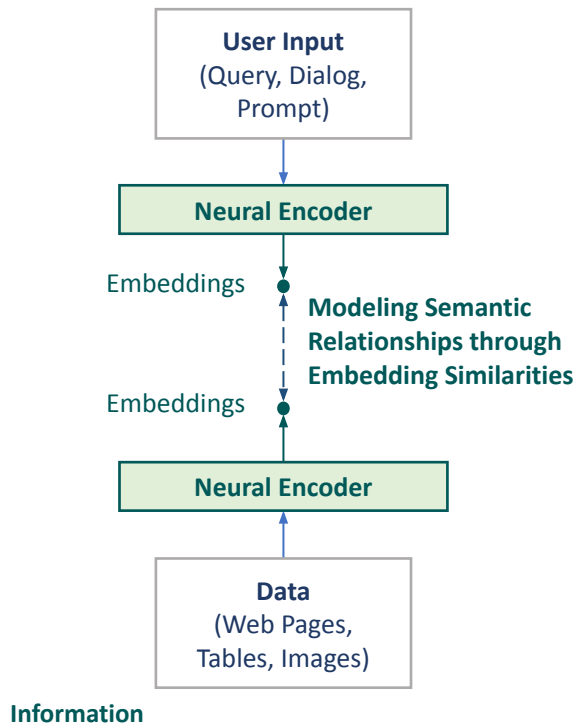
Information

Efficiency

- Offline computed embedding
- Quick similarity calculations
- Various effective nearest neighbor search methods

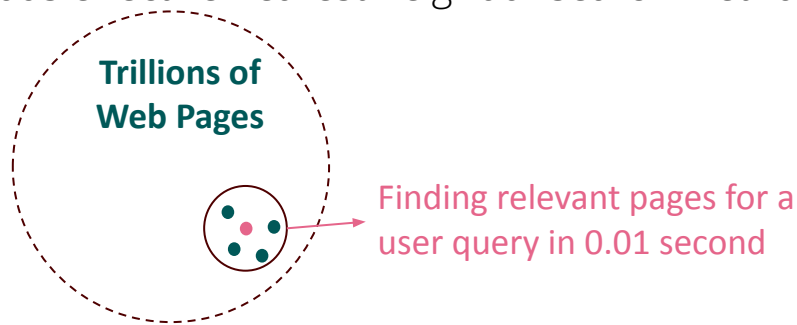


# Embedding Learning: Why?

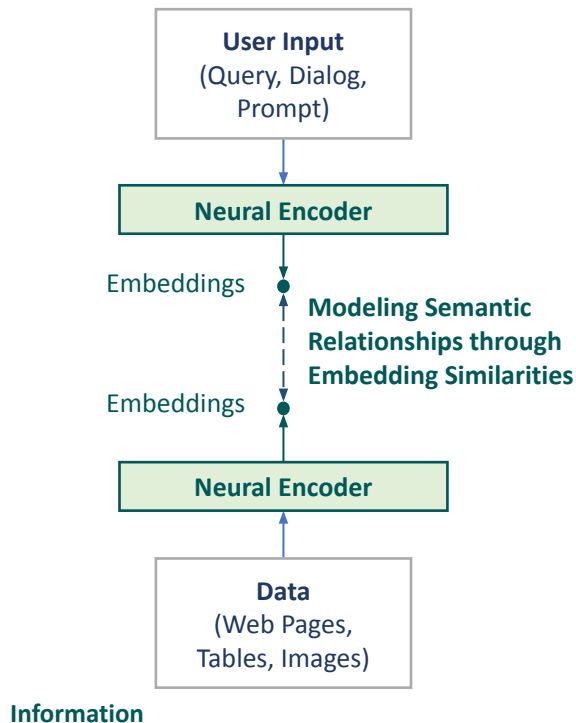


## Efficiency

- Offline computed embedding
- Quick similarity calculations
- Various effective nearest neighbor search methods



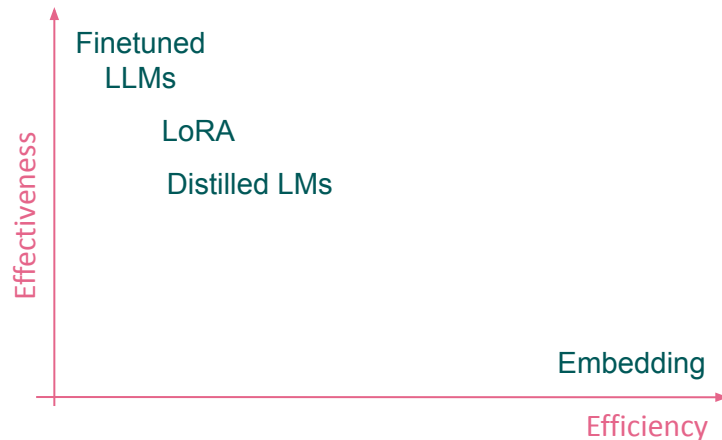
# Embedding Learning: Why?



## Efficiency

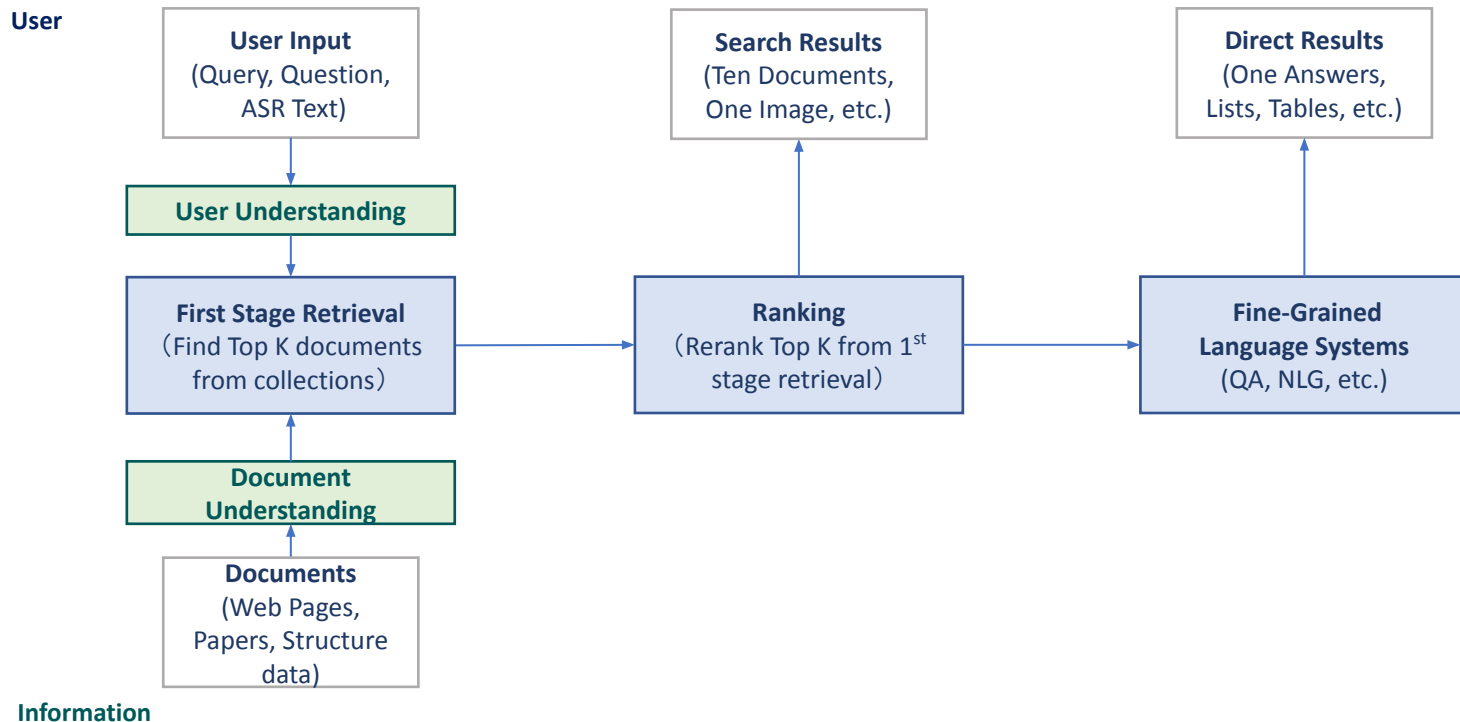
- Offline computed embedding
- Quick similarity calculations
- Various effective nearest neighbor search methods

Trading effectiveness  
for efficiency:



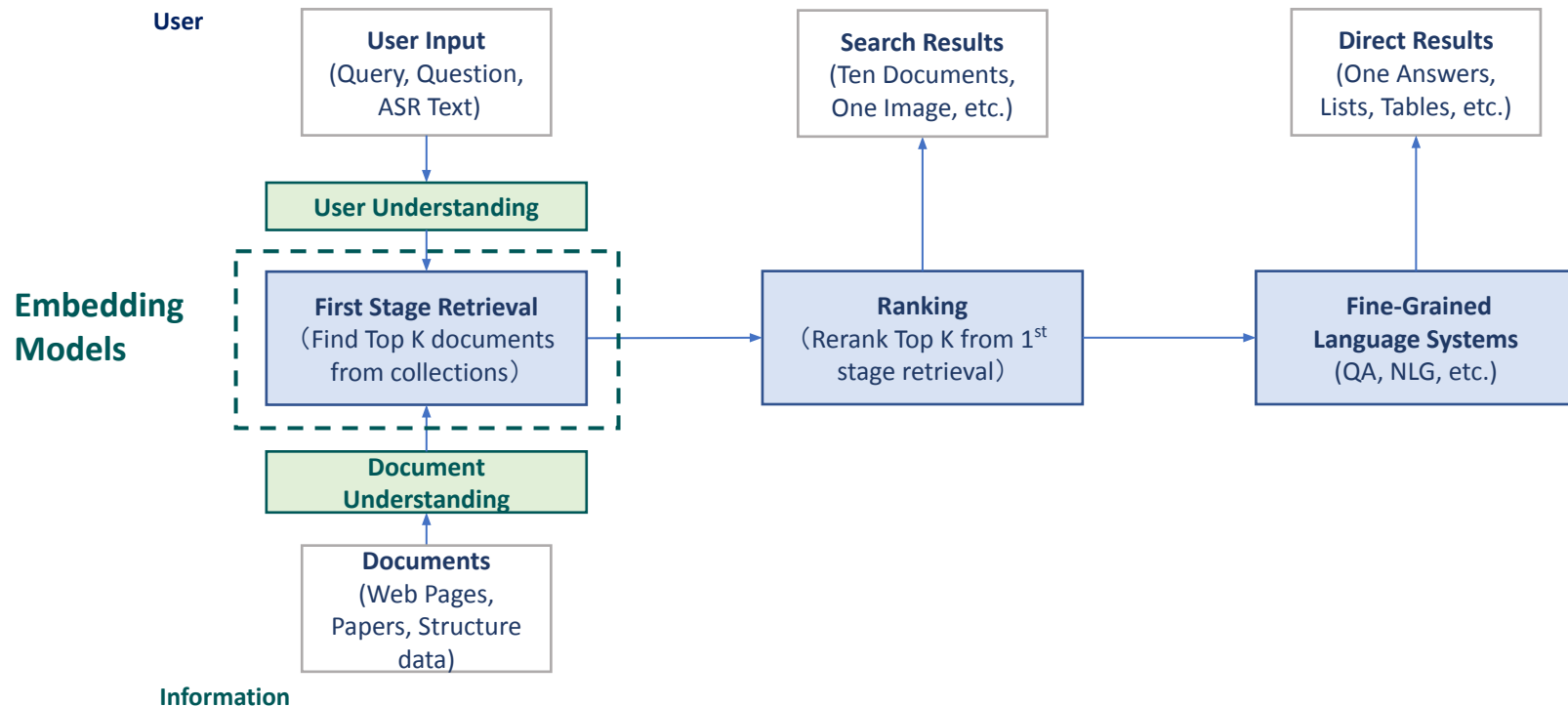
# Embedding Learning: Standard Usage

The first stage of many AI systems: search engines, QA, RAG, etc.



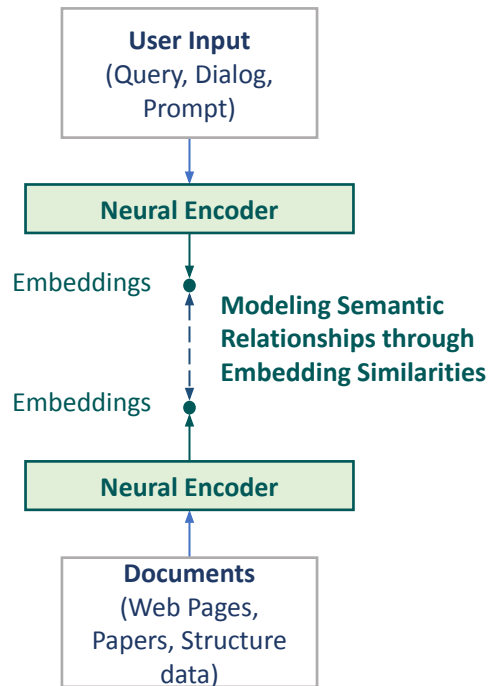
# Embedding Learning: Standard Usage

The first stage of many AI systems: search engines, QA, RAG, etc.



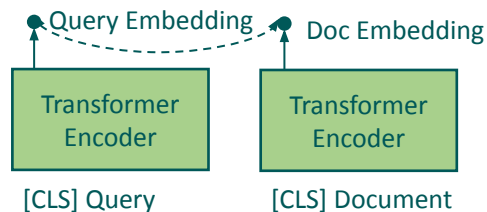
# Embedding Learning: Formulation

User



Information

## Simple Similarity Match



## A representation-centric approach:

- All system capacity from encoders, only simple vector operations afterwards

# Embedding Learning: Formulation

A standard setup with BERT Encoders [3]

**Retrieval Function:** (Dual Encoder)

$$f(q, d) = \text{BERT}(q) \cdot \text{BERT}(d) = \text{MLP}(\overrightarrow{[\text{CLS}]_q}) \cdot \text{MLP}(\overrightarrow{[\text{CLS}]_d})$$

**Inference:** (Approximate KNN Search)

$$D_q = \text{ANN}_{f(q, \circ)}$$

Finding K nearest neighbor in the corpus with approximate nearest neighbor search.



# Embedding Learning: Formulation

A standard setup with BERT Encoders [3]

**Retrieval Function:** (Dual Encoder)

$$f(q, d) = \text{BERT}(q) \cdot \text{BERT}(d) = \text{MLP}(\overrightarrow{[\text{CLS}]_q}) \cdot \text{MLP}(\overrightarrow{[\text{CLS}]_d})$$

**Inference:** (Approximate KNN Search)

$$D_q = \text{ANN}_{f(q, \circ)}$$

Finding K nearest neighbor in the corpus with approximate nearest neighbor search.

Approximate nearest neighbor search (ANNS): Gain (sub-linear) efficiency by slightly scarifying KNN accuracy [4]

- Partition-based methods: Split the space into regions and only search sub regions
  - E.g., hierarchical K-means trees
- Hash-based methods: Map data points by hashing functions and only search certain hash codes
  - E.g., Locality sensitive hash
- Graph-based methods: Connect data points by similarity edges and greedily traverse the graph
  - E.g., K-nearest neighborhood graph

Can achieve similar cost/efficiency as inverted index

# Embedding Learning: Training

Standard random negatives too weak for retrieval

**Learning:** (Contrastive Learning / Learning to Rank)

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim P_D} l(f(q, d^+), f(q, d^-))$$

Relevant q-d pairs (given)

**Negative Sampling**

Standard Ranking Loss

# Embedding Learning: Training

Standard random negatives too weak for retrieval

**Learning:** (Contrastive Learning / Learning to Rank)

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim P_D} l(f(q, d^+), f(q, d^-))$$

Relevant q-d  
pairs (given)

Negative  
Sampling

Standard  
Ranking Loss



**Dense Retrieval Training Loss with Randomly  
Sampled Negatives on MSMARCO**

# Embedding Learning: Training

Standard random negatives too weak for retrieval

**Learning:** (Contrastive Learning / Learning to Rank)

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim P_D^-} l(f(q, d^+), f(q, d^-))$$

Relevant q-d pairs (given)      Negative Sampling      Standard Ranking Loss

A severe problem because of unique properties of retrieval

- Corpus size is **huge**: millions, billions, or trillions
- 99.99% are trivially irrelevant
- Retrieval is to distinguish a **small number** of hard negatives



**Dense Retrieval Training Loss with Randomly Sampled Negatives on MSMARCO**

# Embedding Learning: Sparse Retrieval Negatives

---

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim P_{D^-}} l(f(q, d^+), f(q, d^-))$$

Sampling negatives from top results of existing sparse retrieval systems

- Negatives from existing inverted index (industry's sparse retrieval). (Bing Vector Search) [Waldburger 2019]
- Sampling from BM25 Top K. (DPR) [Karpukhin et al. 2020]
- Offline hard negative mining from production system (Facebook Embedding Search) [Huang et al. 2020]

# Embedding Learning: Sparse Retrieval Negatives

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim P_{D^-}} l(f(q, d^+), f(q, d^-))$$

Sampling negatives from top results of existing sparse retrieval systems

- Negatives from existing inverted index (industry's sparse retrieval). (Bing Vector Search) [Waldburger 2019]
- Sampling from BM25 Top K. (DPR) [Karpukhin et al. 2020]
- Offline hard negative mining from production system (Facebook Embedding Search) [Huang et al. 2020]

Pros:

- Bootstrap upon an existing system with meaningful negatives

Cons:

- Often negatives from sparse retrieval are still too trivial for dense retrieval
- Empirically, weaker generalization ability

# Embedding Learning: Training with Self Negatives

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim \operatorname{ANN}_{f(q, \circ)}} l(f(q, d^+), f(q, d^-))$$

Sampling negatives globally from the entire corpus using the dense retriever itself (ANCE [5]).

- Sampling from the top retrieved results of the dense retrieval model
- Periodically refresh the dense retrieval index to keep negatives updated
- Start from sparse retrieval negatives to warm up



# Embedding Learning: Training with Self Negatives

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim \text{ANN}_{f(q^+)}} l(f(q, d^+), f(q, d^-))$$

Sampling negatives globally from the entire corpus using the dense retriever itself (ANCE [5]).

- Sampling from the top retrieved results of the dense retrieval model
- Periodically refresh the dense retrieval index to keep negatives updated
- Start from sparse retrieval negatives to warm up

Pros:

- Aligned training and testing distribution
- Strong performance in-domain and out-of-domain

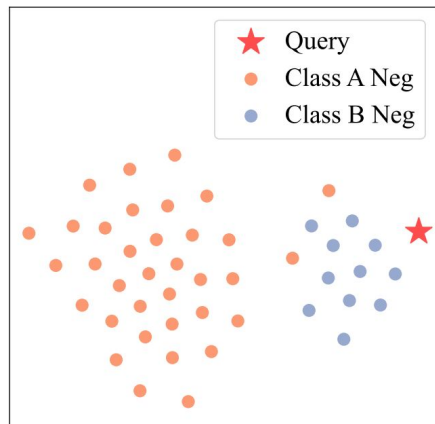
Cons:

- Overhead cost in refreshing the corpus index for negative sampling
- Instabilities from negative refreshes

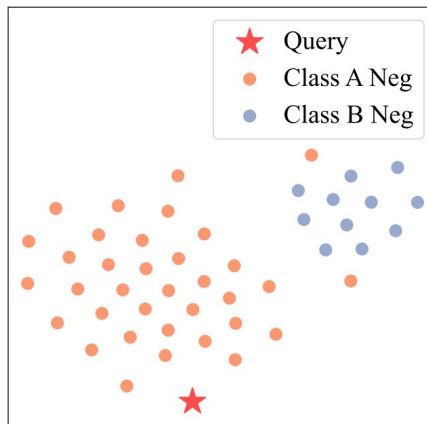
# Embedding Learning: Instabilities from Negative Sampling

Dense retriever swings between several groups of negatives [6]

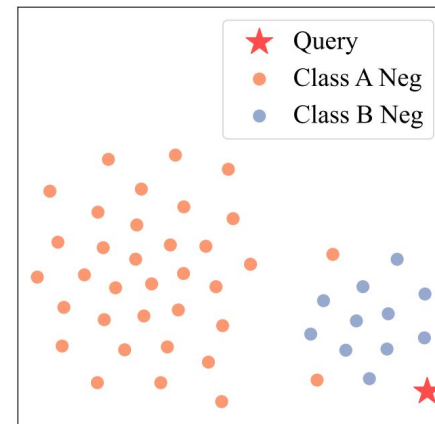
Query	Class A Negatives	Class B Negatives
most popular breed of rabbit	The Golden Retriever is one of the most <b>popular breeds</b> in the United States...	<b>Rabbit</b> habitats include meadows, woods, forests, grasslands, deserts and wetlands...



After Negative Refresh #1



After Negative Refresh #2



After Negative Refresh #3

T-SNE plots of a query and its two negative groups during ANCE training [6]

# Embedding Learning: Training with Smoothed Negatives

Smooth training by combining negatives from past samples and potential future samples (ANCE-Tele [7])

$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim \text{Tele}} l(f(q, d^+), f(q, d^-))$$

$$\text{Tele}_i = \text{ANN}_{f_i(q, \circ)} + \text{Tele}_{i-1} + \text{ANN}_{f_i(d^+, \circ)}$$

Self-Negatives from  
current (i-th) training  
episode

Negatives from  
previous episode  
(Momentum)

Approximation of future  
negatives using neighbors  
of  $d^+$  (Lookahead)

# Embedding Learning: Training with Smoothed Negatives

Smooth training by combining negatives from past samples and potential future samples (ANCE-Tele [7])

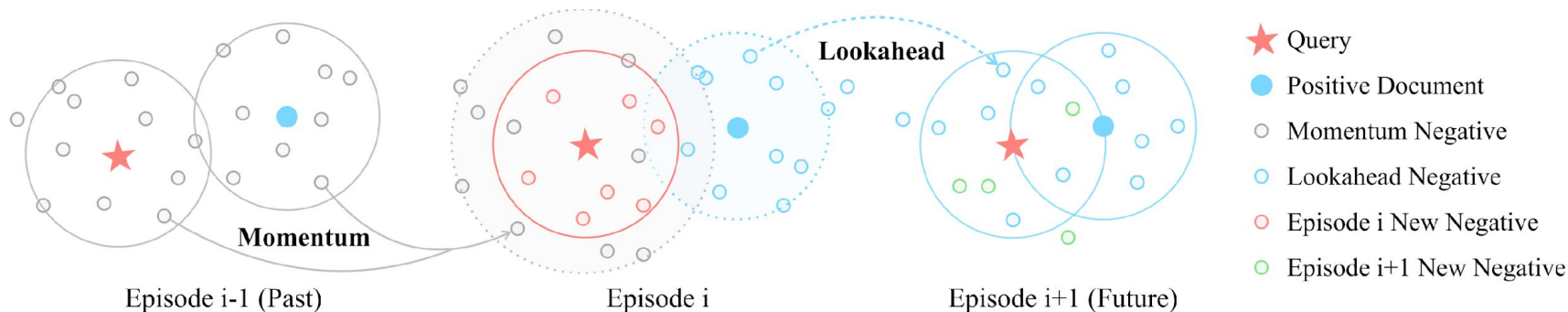
$$\theta^* = \operatorname{argmin}_{\theta} \sum_q \sum_{d^+} \sum_{d^- \sim \text{Tele}} l(f(q, d^+), f(q, d^-))$$

$$\text{Tele}_i = \text{ANN}_{f_i(q, \circ)} + \text{Tele}_{i-1} + \text{ANN}_{f_i(d^+, \circ)}$$

Self-Negatives from  
current (i-th) training  
episode

Negatives from  
previous episode  
(Momentum)

Approximation of future  
negatives using neighbors  
of  $d^+$  (Lookahead)

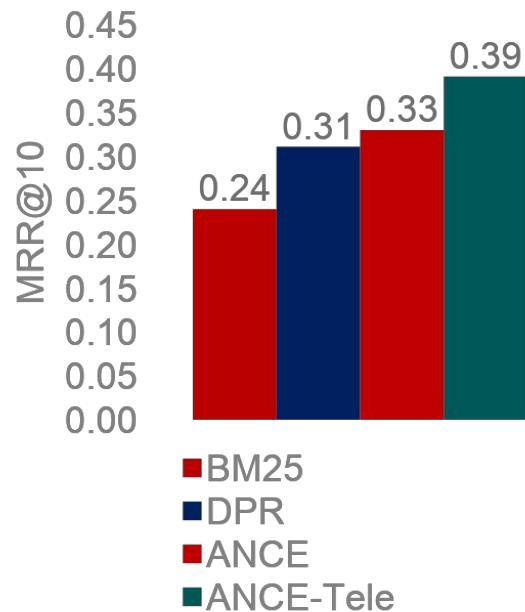


Smooth Negative Sampling with Momentum and Lookahead [6]

# Embedding Learning: Performances

Evaluation on supervised retrieval: MS MARCO Passage Task.

- Retrieve answer passages for Bing questions from a corpus of ~10M passages
- All dense retrievers start from RoBERTa base.



BM25: Standard sparse bag-of-words based retrieval

DPR: Trained with BM25 negatives + random negatives

ANCE: Trained with self-negatives (warmed up by BM25 negative)

ANCE-Tele: Trained with momentum and lookahead global negatives

**Supervised Retrieval  
Performances on MS MARCO.**

# Embedding Learning: Error Cases

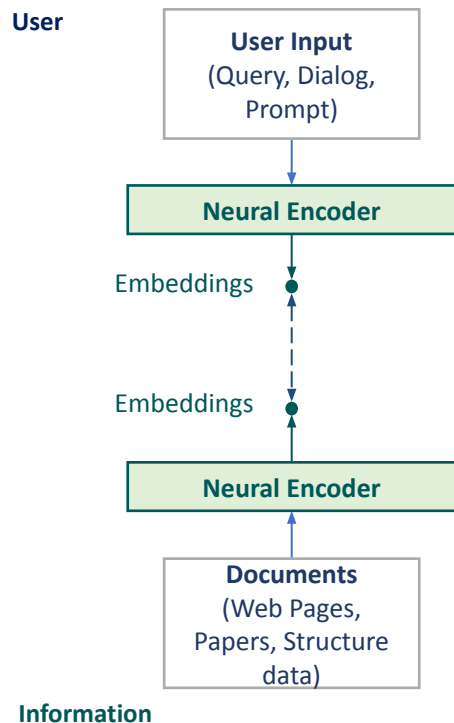
Retriever	Query	Bad Case	Relevant Document
BM25	What is the <u>most popular food</u> in Switzerland	Answers.com: <u>Most popular</u> traditional <u>food</u> dishes of Mexico	Wikipedia: Swiss cuisine
ANCE	How long to hold <u>bow</u> in yoga	Yahoo Answer: How long should you hold a yoga <u>pose</u> for	yogaoutlet.com: How to do bow pose in yoga

**Error Cases of BM25 and ANCE in TREC Deep Learning Track Document Retrieval 2019 [5]**

Sparse retrieval and dense retrieval behave quite differently.

- BM25 and ANCE only agree on 20% of their top 100 rankings
- But both find relevant document in top 3

# Embedding Learning: Recap



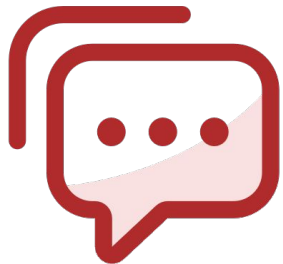
Mapping all kinds of data points into embedding vectors

- Capturing data relationships via embedding similarity
- Support efficient nearest neighbor lookup
- Trade effectiveness for efficiency

Training of embedding models require special treatment

- Contrastive loss
- Sophisticated negative selection methods





# Audience Q&A

① Start presenting to display the audience questions on this slide.



**Why embedding learning needs  
special treatment, rather than  
following standard fine tuning or  
few-shot learning?**

① Start presenting to display the poll results on this slide.

# Mismatch of Embeddings and LLMs: Anisotropy/Non-Uniformity

---

Zero-shot performance of pretrained embeddings on semantic text similarity (STS) tasks

- STS Task: producing a similarity score for a given pair of sentences
- Metric: by Pearson correlation with human rating (e.g., 5 being exact same meaning/paraphrase)

# Mismatch of Embeddings and LLMs: Anisotropy/Non-Uniformity

Zero-shot performance of pretrained embeddings on semantic text similarity (STS) tasks

- STS Task: producing a similarity score for a given pair of sentences
- Metric: by Pearson correlation with human rating (e.g., 5 being exact same meaning/paraphrase)

Model	STS12	STS13	STS14	STS15	STS16	STSb
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50

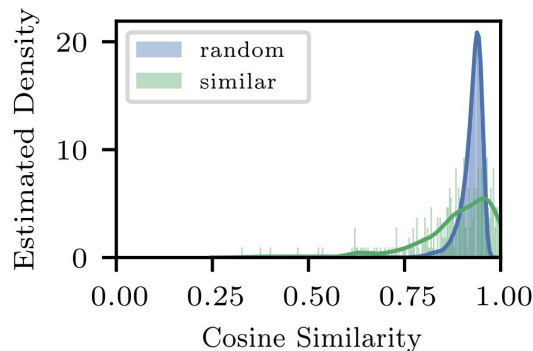
**BERT embedding similarity performances on STS tasks [7]**

Much worse performance than GloVe Embeddings.

- [CLS] is near random.
- Mean-pooling over tokens is better but still much worse than word embeddings

# Mismatch of Embeddings and LLMs: Anisotropy/Non-Uniformity

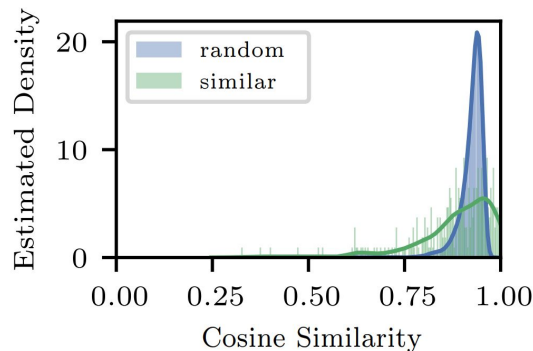
The sequence embedding space of many pretrained LLMs are highly non-uniform



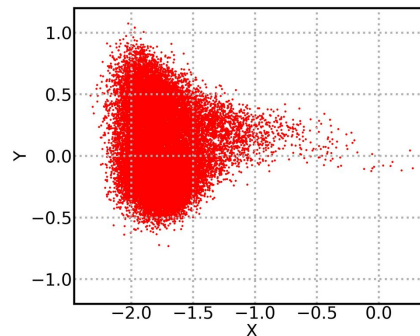
**Similarity of RoBERTa  $\overrightarrow{[CLS]}$  on  
semantically similar and random pairs  
from STS-S [8]**

# Mismatch of Embeddings and LLMs: Anisotropy/Non-Uniformity

The sequence embedding space of many pretrained LLMs are highly non-uniform



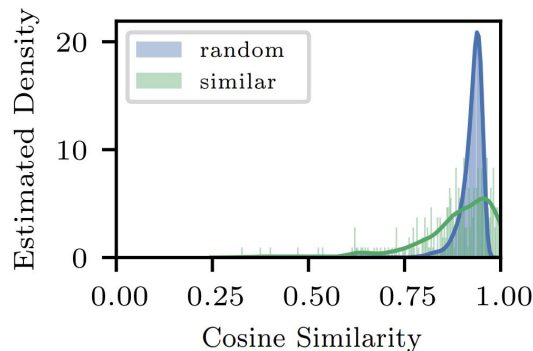
**Similarity of RoBERTa  $\overline{[CLS]}$  on semantically similar and random pairs from STS-S [8]**



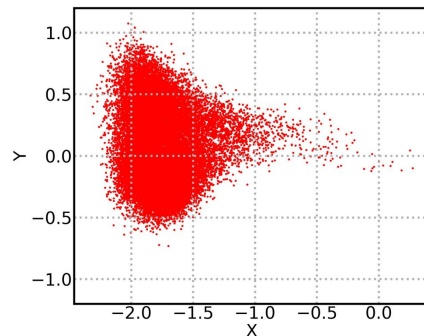
**SVD 2-D mapping of word embeddings from Transformer trained on EN→DE [9]**

# Mismatch of Embeddings and LLMs: Anisotropy/Non-Uniformity

The sequence embedding space of many pretrained LLMs are highly non-uniform



**Similarity of RoBERTa  $\overline{[CLS]}$  on semantically similar and random pairs from STS-S [8]**



**SVD 2-D mapping of word embeddings from Transformer trained on EN→DE [9]**

Most rare tokens are pushed to a narrow cone in the space, and  $[CLS]$  is a rare token in learning

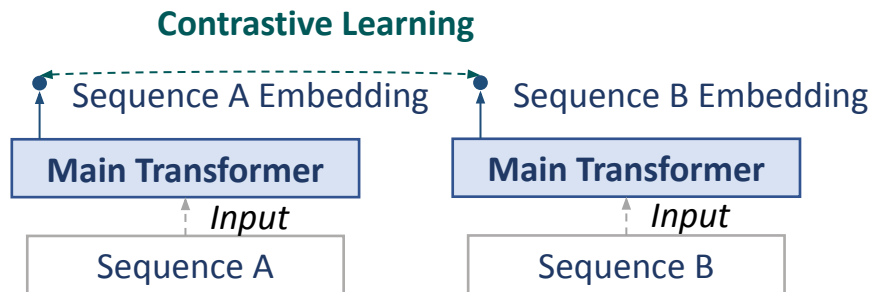
- Every training signal pushes all negatives away from the positive
- Rare tokens (without much or any positive pulls) are pushed away from all positives, into a narrow cone



# Anisotropy Solution: Sequence Contrastive Learning

Pretraining sequence representations with Sequence Contrastive Learning (SCL) [8]

Adding pretraining task:  $L_{\text{SCL}} = \mathbb{E} \left( \frac{\exp(\cos(\mathbf{s}, \mathbf{s}^+))}{\exp(\cos(\mathbf{s}, \mathbf{s}^+)) + \sum_{\mathbf{s}^-} \exp(\cos(\mathbf{s}, \mathbf{s}^-))} \right)$



# Anisotropy Solution: Sequence Contrastive Learning

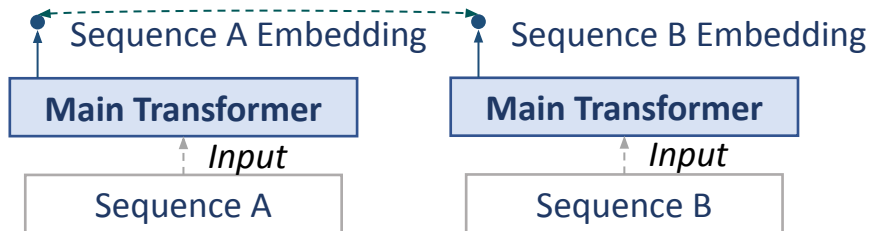
Pretraining sequence representations with Sequence Contrastive Learning (SCL) [8]

Adding pretraining task:  $L_{\text{SCL}} = \mathbb{E} \left( \frac{\exp(\cos(\mathbf{s}, \mathbf{s}^+))}{\exp(\cos(\mathbf{s}, \mathbf{s}^+)) + \sum_s \exp(\cos(\mathbf{s}, \mathbf{s}^-))} \right)$

Annotations:

- Embeddings of positive contrast sequence pairs**: Points to  $\cos(\mathbf{s}, \mathbf{s}^+)$
- Embeddings of negative sequence pairs**: Points to  $\cos(\mathbf{s}, \mathbf{s}^-)$

## Contrastive Learning

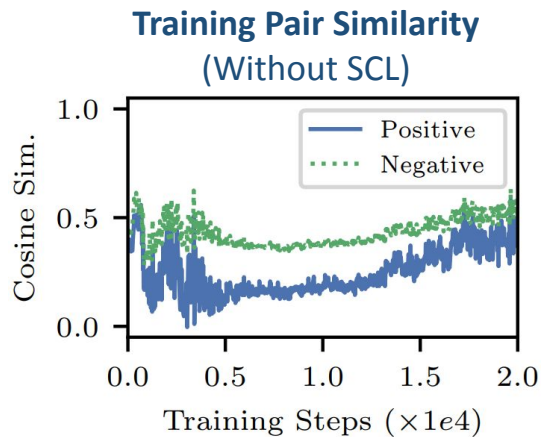


## Construction of positive contrast sequence pairs:

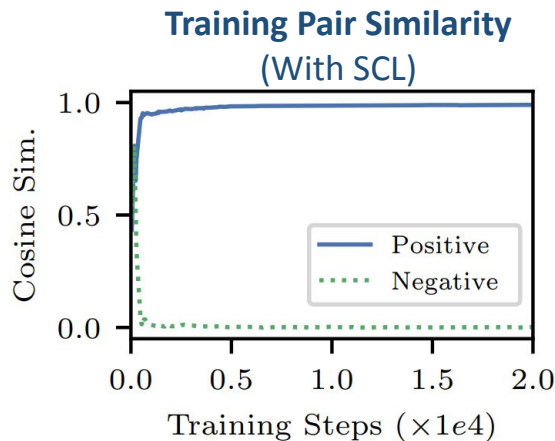
- *Data augmentation*: cropping [11], random replacement, back translation, different dropout (SimCSE), etc.
- *Unsupervised pairs*: co-occurrence in doc (co-doc), etc.
- *Supervisions*: Web QA pairs, search query-clicked docs...

# Anisotropy Solution: Sequence Contrastive Learning

Recalibration of the embedding space, e.g., using cropped sequence pairs (90% overlap)



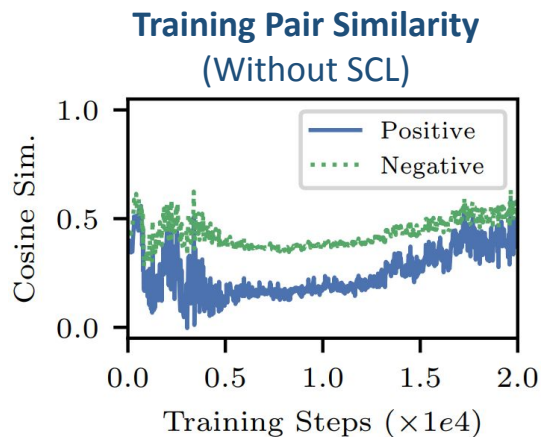
Failed without SCL  
(Although 90% overlap!)



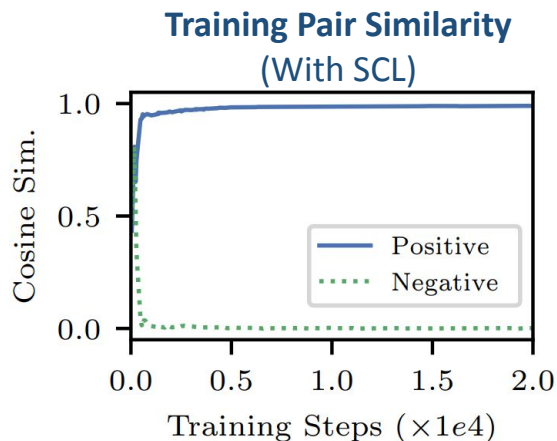
Easy-to-Learn Task  
(90% overlap, after all)

# Anisotropy Solution: Sequence Contrastive Learning

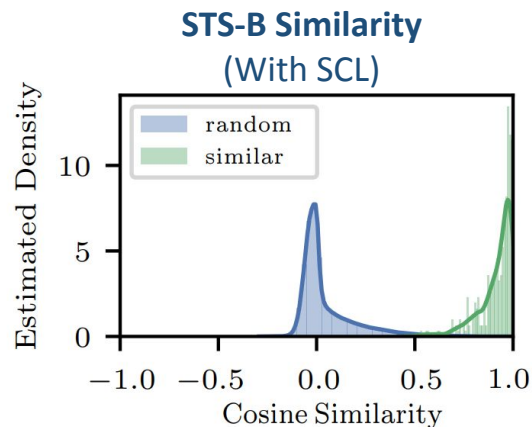
Recalibration of the embedding space, e.g., using cropped sequence pairs (90% overlap)



Failed without SCL  
(Although 90% overlap!)



Easy-to-Learn Task  
(90% overlap, after all)



Effective Calibration  
& Good Zero-Shot Ability

Decent zero-shot performance on many sequence similarity tasks and non-random performance on retrieval

# Deeper Look into Contrastive Learning

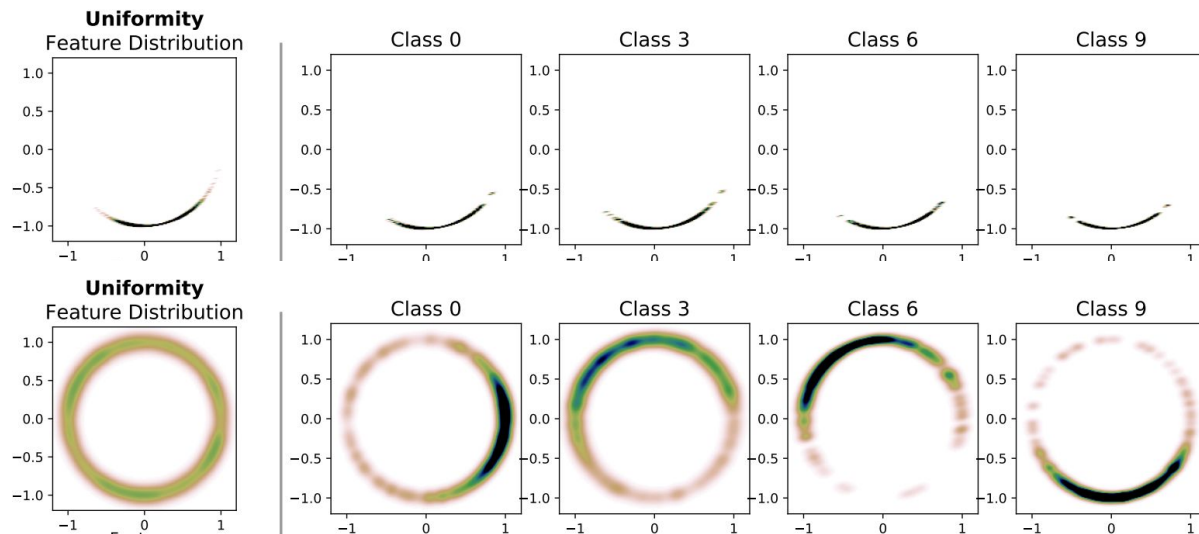
Two forces in contrastive learning: Alignment and Uniformity [10]

$$L_{\text{SCL}} = \mathbb{E} \left( \frac{\exp(\cos(\mathbf{s}, \mathbf{s}^+))}{\exp(\cos(\mathbf{s}, \mathbf{s}^+)) + \sum_{\mathbf{s}^-} \exp(\cos(\mathbf{s}, \mathbf{s}^-))} \right)$$
$$\sim \underbrace{\cos(\mathbf{s}, \mathbf{s}^+)}_{\text{Align positive pairs together}} + \underbrace{\log(\exp(\cos(\mathbf{s}, \mathbf{s}^+)) + \sum_{\mathbf{s}^-} \exp(\cos(\mathbf{s}, \mathbf{s}^-)))}_{\text{Uniformly spread random pairs in the space}}$$

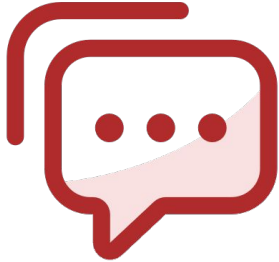
- Proof in Wang et al. [10] that, if exist, perfectly aligned/uniform encoders minimize the two terms
- Note: here negatives are sampled uniformly, not from a long tail distribution

# Deeper Look into Contrastive Learning

Two forces in contrastive learning: Alignment and Uniformity [10]



**Uniformity of image features in CIFAR-10 from random network (top) and unsupervised contrastive learning (bottom) [10]**



# Audience Q&A

① Start presenting to display the audience questions on this slide.



# Mismatch of Embeddings and LLMs: Alignment

---

LLMs are pretrained using context (word co-occurrences) information

- Predict next token given current prefix.
  - Input: I took my dog, Fido, to the park for his
  - Target: walk

Contextual information is known to disagree with many embedding tasks' needs [11]

# Mismatch of Embeddings and LLMs: Alignment

LLMs are pretrained using context (word co-occurrences) information

- Predict next token given current prefix.
  - Input: I took my dog, Fido, to the park for his
  - Target: walk

Contextual information is known to disagree with many embedding tasks' needs [11]

## Words Sharing Similar Context

Tokyo - - - - London

BMW - - - - Car

ATT - - - - Verizon

# Mismatch of Embeddings and LLMs: Alignment

LLMs are pretrained using context (word co-occurrences) information

- Predict next token given current prefix.
  - Input: I took my dog, Fido, to the park for his
  - Target: walk

Contextual information is known to disagree with many embedding tasks' needs [11]

## Words Sharing Similar Context

Tokyo - - - - London

BMW - - - - Car

ATT - - - - Verizon

## Words Indicating Relevance

Tokyo - - - - Travel

BMW - - - - Price

ATT - - - - Plans

# Mismatch of Embeddings and LLMs: Alignment

What information does unsupervised contrastive pairs bring in to align the embedding space?

Method	Sequence A	Sequence B
SimCSE	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.
Inverse Cloze Task (ICT)	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	They currently play their home games at Acrisure Stadium on Pittsburgh's North Side in the North Shore neighborhood,
Cropping Augmentation	The Steelers enjoy a large, widespread fanbase nicknamed ____	____ enjoy a large, widespread fanbase nicknamed Steeler Nation.
Co-document	The Steelers enjoy a large, widespread fanbase nicknamed Steeler Nation.	In the NFL's "modern era" (since the AFL–NFL merger in 1970) the Steelers have posted the best record in the league.

Very limited semantic signals in the alignment for search relevance

- Either strong term overlaps or loosely correlated

# Mismatch of Embeddings and LLMs: Alignment

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by

# Mismatch of Embeddings and LLMs: Alignment

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by

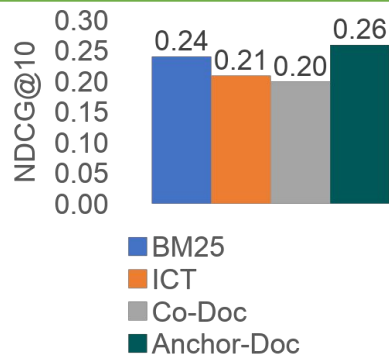
Web graph and anchor information is widely used in many web and search applications

- Determine the importance of a web page (Page Rank)
- Enrich the representation of a document , using 3<sup>rd</sup> party information (Document Expansion)
- Serve as pseudo queries for feature-based ranking models

# Mismatch of Embeddings and LLMs: Alignment

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by



**MARCO NDCG@10  
with different signals**

**Anchor-Doc the only unsupervised signal source outperforms BM25**

- Data cleaning required to filter out functional anchors, e.g., “homepage”

**A widely useful information in standard web search**

- Page Rank, Document Expansion, etc.

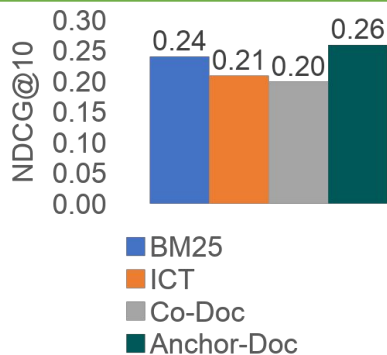
**Still a weakly supervised method, rather than a pretraining method**

- Behavior closer to weak supervision/transfer learning, not pretraining

# Mismatch of Embeddings and LLMs: Alignment

Leverage Anchor Texts and the document they point to pseudo query-relevant document pairs

Method	Sequence A	Sequence B
Anchor-Document	Vegetarian Society of Ireland	The Vegetarian Society of Ireland is a registered charity. Our aim is to increase awareness of vegetarianism in relation to health,
Actual Argument Retrieval Data	Becoming a vegetarian is an environmentally friendly thing to do.	Health general weight philosophy ethics You don't have to be vegetarian to be green. Many special environments have been created by



**MARCO NDCG@10  
with different signals**

**Anchor-Doc the only unsupervised signal source outperforms BM25**

- Data cleaning required to filter out functional anchors, e.g., “homepage”

**A widely useful information in standard web search**

- Page Rank, Document Expansion, etc.

**Still a weakly supervised method, rather than a pretraining method**

- Behavior closer to weak supervision/transfer learning, not pretraining

Future lecture: LLM generated synthetic queries are good weak supervisions too.



# Mismatch of Embeddings and LLMs

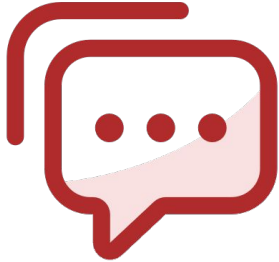
---

There are intrinsic differences between LLMs' ability and the embedding models' needs

- Adapting LLMs into embedding models require special work
- Often separated offering from LLMs companies for embeddings

The intrinsic differences make many LLM's power not seeing yet in embedding

- Scaling law is shaky
- Not much prompting or instruction following
- Nor In-context learning



# Audience Q&A

① Start presenting to display the audience questions on this slide.