Retrieval-Augmented Generation

Lara J. Martin (she/they)

https://laramartin.net/interactive-fiction-class

Slides from ACL 2023 Tutorial by Akari Asai, Sewon Min, Zexuan Zhong, & Dr. Danqi Chen

Learning Objectives

Define the Story Cloze Test and determine its place in guided story generation

Understand the reasons why RAG was created

Explore how the retrieval component interacts with the LLM in RAG

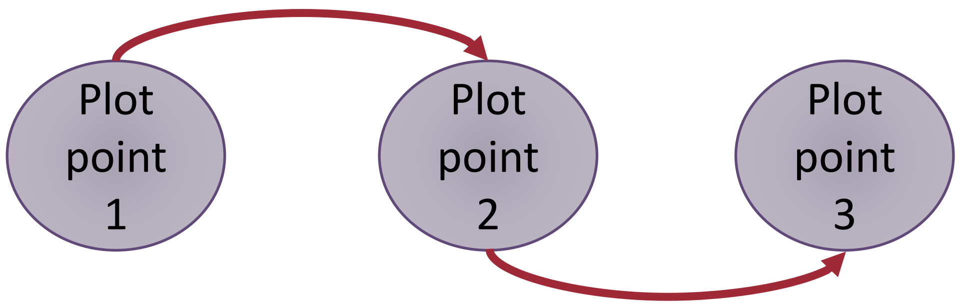
Extract implementation details from papers and find different ways RAG is implemented

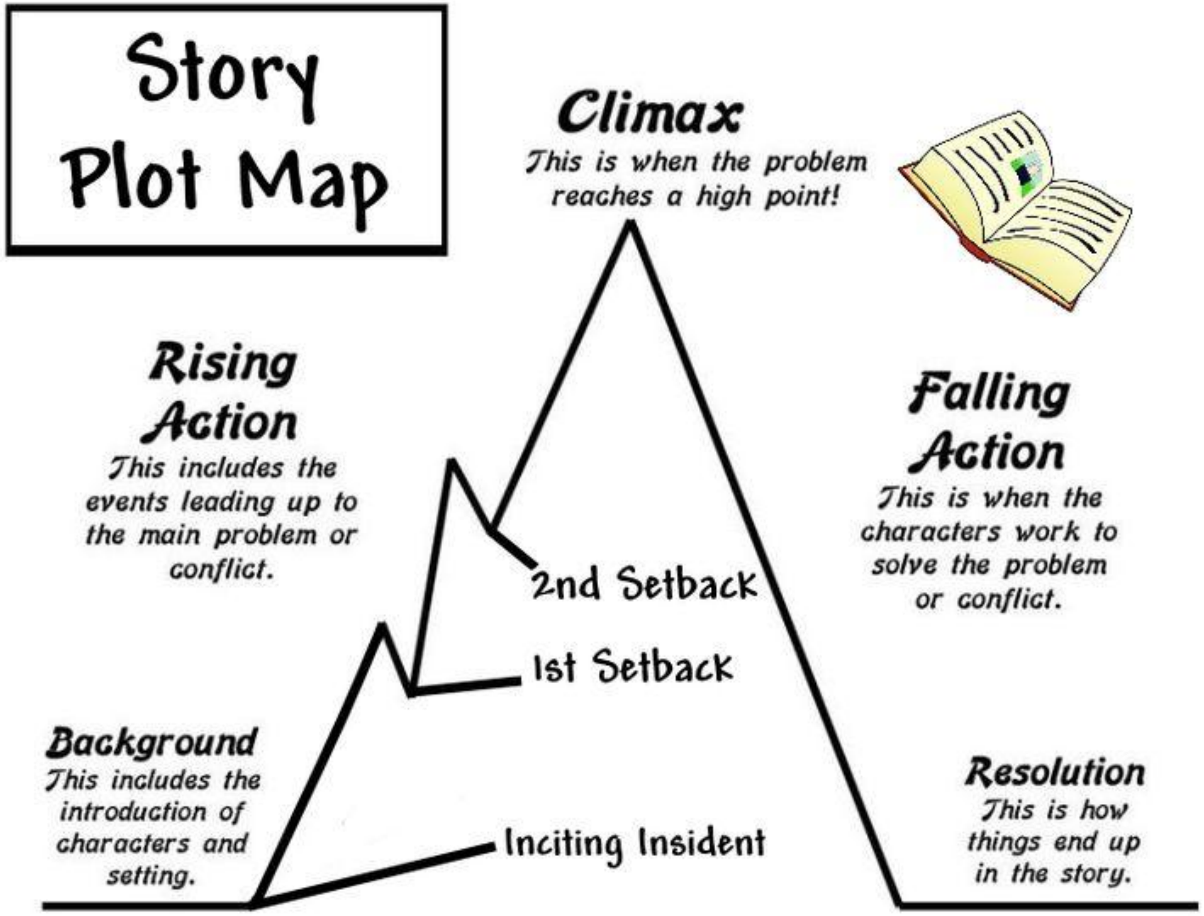
Compare plot-guided generation to retrieval-augmented generation for stories

Review:
Scripts, Procedures, and Plots…oh my!

Schank & Abelson believe that everyone has scripts in their heads built from common experiences

Authors often plan out **plots** before they write stories





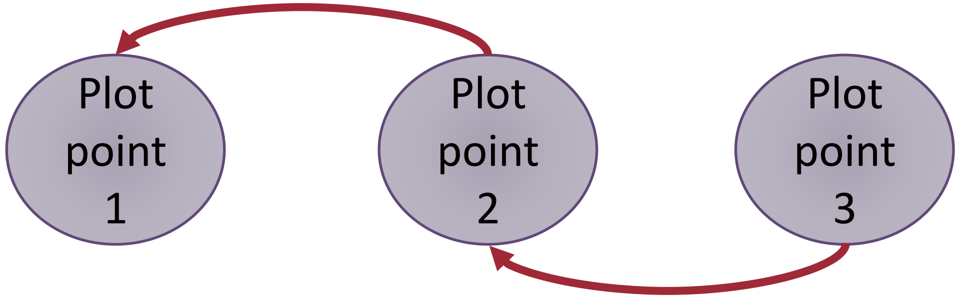
https://i.pinimg.com/736x/57/f7/03/57f703afc709080bddc2c3cfed8dd061.jpg

Review:
Scripts, Procedures, and Plots…oh my!

Schank & Abelson believe that everyone has scripts in their heads built from common experiences

Authors often plan out plots before they write stories

Stories that aren't planned out either have to "**reincorporate**"[1] ideas or the stories feel unfinished



[1] The idea of *reincorporation* is explored in the book Impro by Keith Johnstone

Review: Ways of Extracting Plot Points

Most salient keywords

Event representations

Verb-Noun Sets

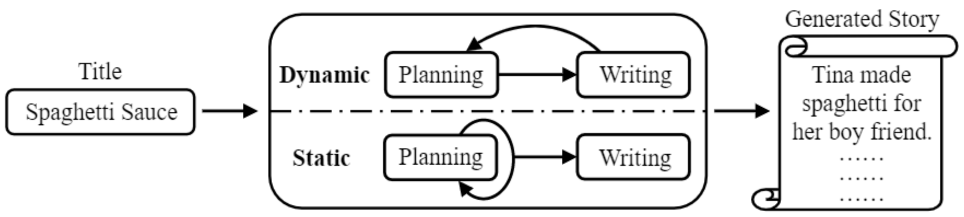
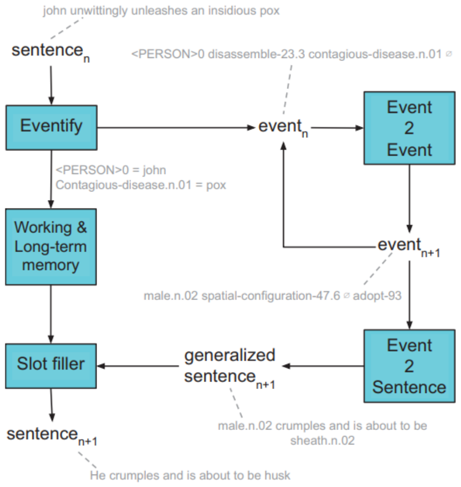
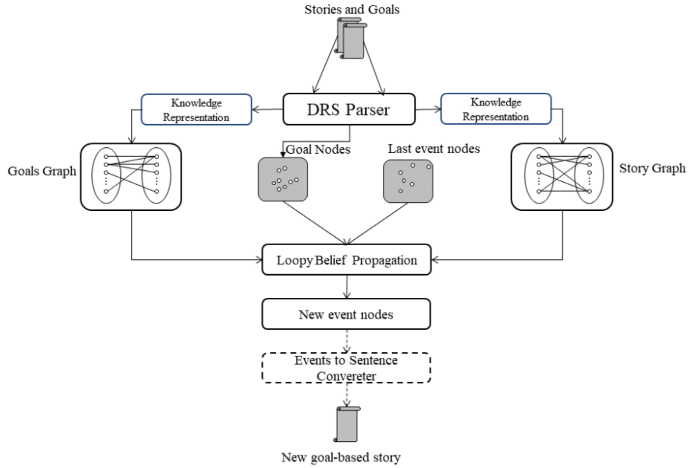


Figure 1: An overview of our system.





Review: Generating with Plot Points

Co-generated vs conditioned (prompted) with plot

Generate event & then translate to natural language

Graph algorithms (Loopy Belief Propagation)

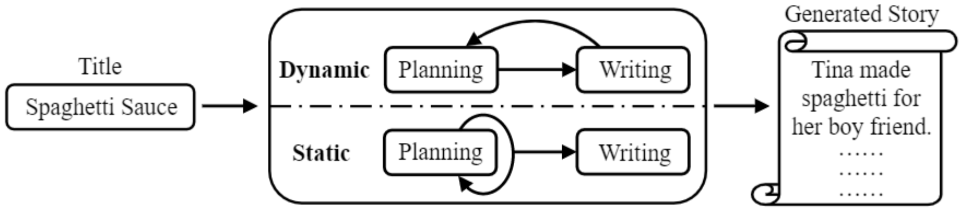
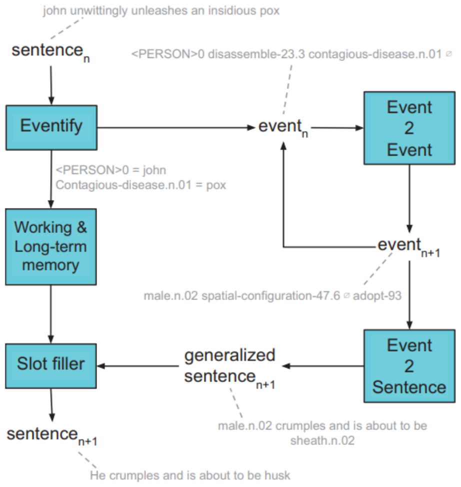
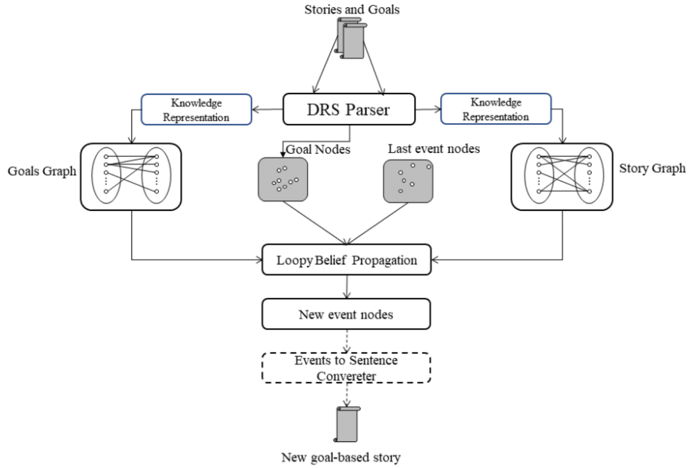


Figure 1: An overview of our system.





The Story Cloze Test

What is a Cloze Test?

- Something is removed from a text; try to guess what's missing
- Used for reading comprehension, grammar, etc. (with humans)

Unsupervised Learning of Narrative Event Chains

Nathanael Chambers and **Dan Jurafsky**
Department of Computer Science
Stanford University
Stanford, CA 94305
{natec,jurafsky}@stanford.edu

Abstract

Hand-coded *scripts* were used in the 1970-80s as knowledge backbones that enabled inference and other NLP tasks requiring deep semantic knowledge. We propose unsupervised induction of similar schemata called *narrative event chains* from raw newswire text.

A narrative event chain is a partially ordered set of events related by a common protagonist. We describe a three step process to learning narrative event chains. The first uses unsupervised distributional methods to learn narrative relations between events sharing coreferring arguments. The second applies a tem

tate learning, and thus this paper addresses the three tasks of chain induction: *narrative event induction*, *temporal ordering of events* and *structured selection* (pruning the event space into discrete sets).

Learning these prototypical schematic sequences of events is important for rich understanding of text. Scripts were central to natural language understanding research in the 1970s and 1980s for proposed tasks such as summarization, coreference resolution and question answering. For example, Schank and Abelson (1977) proposed that understanding text about restaurants required knowledge about the Restaurant Script, including the participants (Cus-
er, Waiter, Cook, Tables, etc.) the e... consti-

Narrative Cloze Test

Evaluate “event relatedness”

Find which events could be missing from a narrative chain

Uses verbs only

N. Chambers and D. Jurafsky, “Unsupervised Learning of Narrative Event Chains,” in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2008, pp. 789–797, doi: 10.1.1.143.1555.

Narrative Cloze Test

Known events:
(pleaded subj), (admits subj), (convicted obj)

Likely Events:

sentenced obj	0.89	indicted obj	0.74
paroled obj	0.76	fined obj	0.73
fired obj	0.75	denied subj	0.73

Figure 1: Three narrative events and the six most likely events to include in the same chain.

X pleaded _

X admits _

_ convicted X

N. Chambers and D. Jurafsky, “Unsupervised Learning of Narrative Event Chains,” in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2008, pp. 789–797, doi: 10.1.1.143.1555.

Finish the story

Gina was worried the cookie dough in the tube would be gross.

She was very happy to find she was wrong.

The cookies from the tube were as good as from scratch.

Gina intended to only eat 2 cookies and save the rest.

A. Gina liked the cookies so much she ate them all in one sitting. ✓
B. Gina gave the cookies away at her church.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 839–849. http://www.aclweb.org/anthology/N16-1098

Story Cloze Test

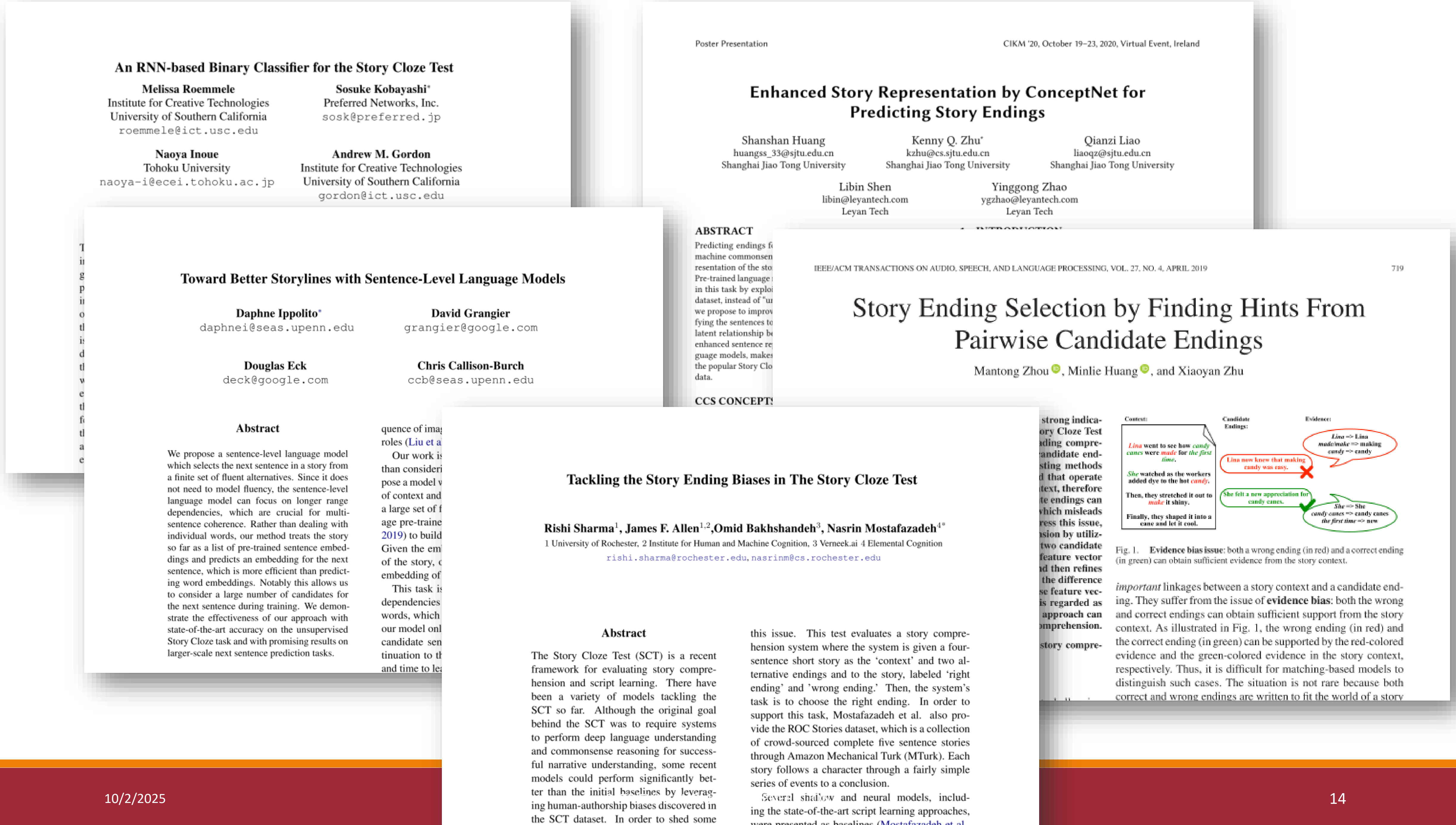
Predict/select the most likely story *ending*

- Given the first 4 sentences of the story

Full sentences

Multiple choice evaluation

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 839–849. http://www.aclweb.org/anthology/N16-1098



An RNN-based Binary Classifier for the Story Cloze Test

Melissa Roemmele, Institute for Creative Technologies, University of Southern California, roemmele@ict.usc.edu
Sosuke Kobayashi*, Preferred Networks, Inc., sosk@preferred.jp
Naoya Inoue, Tohoku University, naoya-i@ecei.tohoku.ac.jp
Andrew M. Gordon, Institute for Creative Technologies, University of Southern California, gordon@ict.usc.edu

Poster Presentation — CIKM '20, October 19–23, 2020, Virtual Event, Ireland

Enhanced Story Representation by ConceptNet for Predicting Story Endings

Shanshan Huang, huangss_33@sjtu.edu.cn, Shanghai Jiao Tong University
Kenny Q. Zhu*, kzhu@cs.sjtu.edu.cn, Shanghai Jiao Tong University
Qianzi Liao, liaoqz@sjtu.edu.cn, Shanghai Jiao Tong University
Libin Shen, libin@leyantech.com, Leyan Tech
Yinggong Zhao, ygzhao@leyantech.com, Leyan Tech

Toward Better Storylines with Sentence-Level Language Models

Daphne Ippolito*, daphnei@seas.upenn.edu
David Grangier, grangier@google.com
Douglas Eck, deck@google.com
Chris Callison-Burch, ccb@seas.upenn.edu

Abstract

We propose a sentence-level language model which selects the next sentence in a story from a finite set of fluent alternatives. Since it does not need to model fluency, the sentence-level language model can focus on longer range dependencies, which are crucial for multi-sentence coherence. Rather than dealing with individual words, our method treats the story so far as a list of pre-trained sentence embeddings and predicts an embedding for the next sentence, which is more efficient than predicting word embeddings. Notably this allows us to consider a large number of candidates for the next sentence during training. We demonstrate the effectiveness of our approach with state-of-the-art accuracy on the unsupervised Story Cloze task and with promising results on larger-scale next sentence prediction tasks.

IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 27, NO. 4, APRIL 2019 — 719

Story Ending Selection by Finding Hints From Pairwise Candidate Endings

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu

Fig. 1. **Evidence bias issue**: both a wrong ending (in red) and a correct ending (in green) can obtain sufficient evidence from the story context.

important linkages between a story context and a candidate ending. They suffer from the issue of **evidence bias**: both the wrong and correct endings can obtain sufficient support from the story context. As illustrated in Fig. 1, the wrong ending (in red) and the correct ending (in green) can be supported by the red-colored evidence and the green-colored evidence in the story context, respectively. Thus, it is difficult for matching-based models to distinguish such cases. The situation is not rare because both correct and wrong endings are written to fit the world of a story

Tackling the Story Ending Biases in The Story Cloze Test

Rishi Sharma[1], James F. Allen[1,2], Omid Bakhshandeh[3], Nasrin Mostafazadeh[4*]
1 University of Rochester, 2 Institute for Human and Machine Cognition, 3 Verneek.ai 4 Elemental Cognition
rishi.sharma@rochester.edu, nasrinm@cs.rochester.edu

Abstract

The Story Cloze Test (SCT) is a recent framework for evaluating story comprehension and script learning. There have been a variety of models tackling the SCT so far. Although the original goal behind the SCT was to require systems to perform deep language understanding and commonsense reasoning for successful narrative understanding, some recent models could perform significantly better than the initial baselines by leveraging human-authorship biases discovered in the SCT dataset. In order to shed some

this issue. This test evaluates a story comprehension system where the system is given a four-sentence short story as the 'context' and two alternative endings and to the story, labeled 'right ending' and 'wrong ending.' Then, the system's task is to choose the right ending. In order to support this task, Mostafazadeh et al. also provide the ROC Stories dataset, which is a collection of crowd-sourced complete five sentence stories through Amazon Mechanical Turk (MTurk). Each story follows a character through a fairly simple series of events to a conclusion.

Several shallow and neural models, including the state-of-the-art script learning approaches,

Think Pair Share

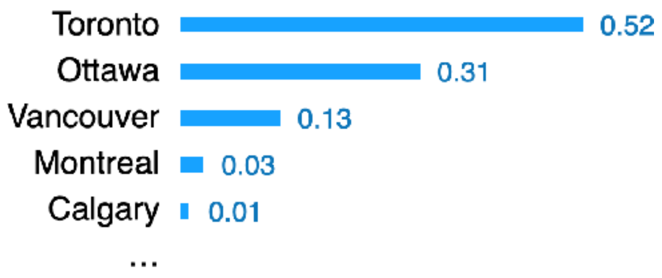
The Story Cloze Test was created for evaluating systems' performance on understanding stories.

How could you use it instead for *generation*?

Retrieval-Augmented Generation

Retrieval-based language models (LMs)
Retrieval-based LMs = Retrieval + LMs

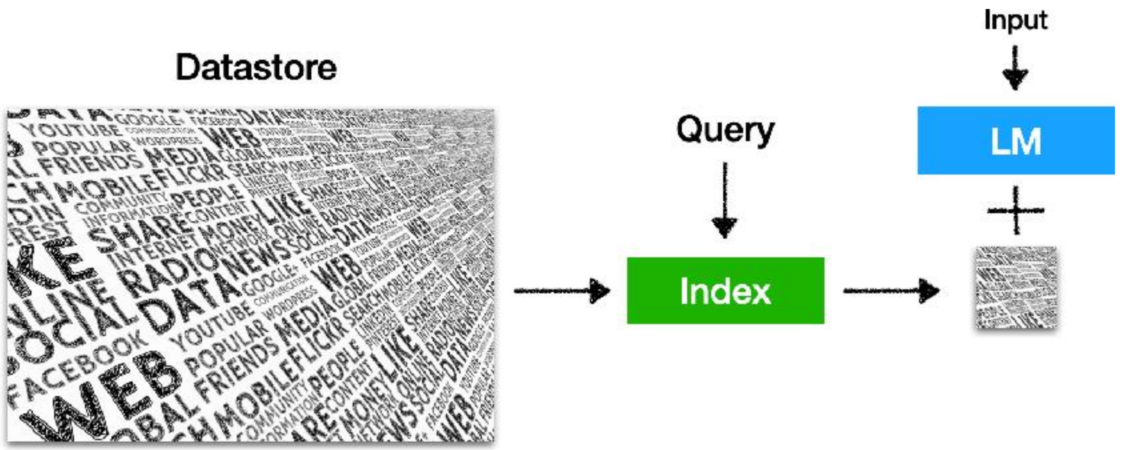
- It is a **language model** $P(x_n \mid x_1, x_2, \cdots, x_{n-1})$



The capital city of Ontario is ___

(can be broadly extended to masked language models or encoder-decoder models)

- It retrieves from an **external datastore** (at least during inference time)



Retrieval for knowledge-intensive NLP tasks

Representative tasks: open-domain QA, fact checking, entity linking, ..

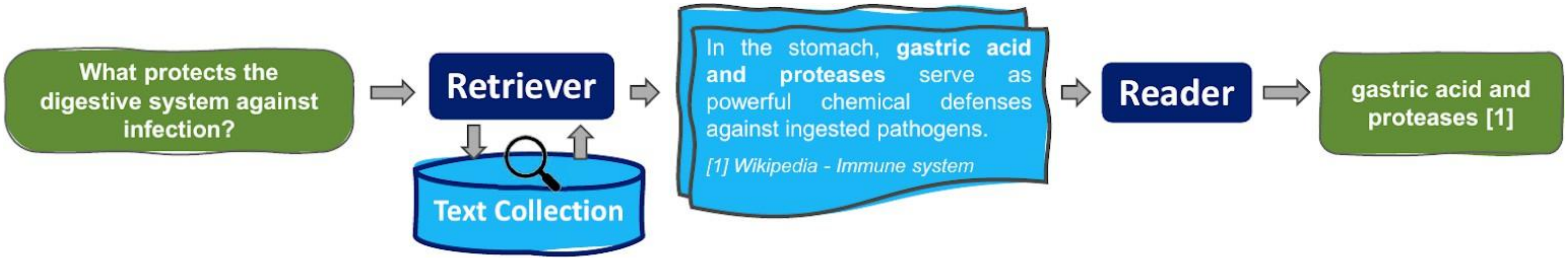


Image: http://ai.stanford.edu/blog/retrieval-based-NLP/

Drives a lot of research on better algorithms for **dense retrieval**, e.g., **DPR** (Karpukhin et al., 2020), **ColBERT** (Khattab and Zaharia, 2020), **ANCE** (Xiong et al., 2021), **Contriever** (Izacard et al., 2022), …

Why retrieval LMs?

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

Why retrieval-based LMs?

List 5 important papers authored by Geoffrey Hinton

ChatGPT

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**, ..

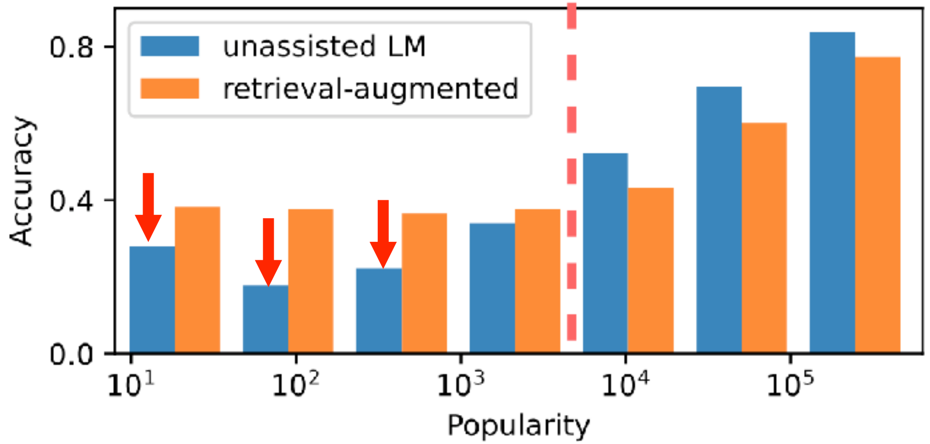
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ..

…

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,…

What is Kathy Saltzman's occupation?



(Mallen et al., 2023)

GPT-3 davinci-003: 20%-30% accuracy

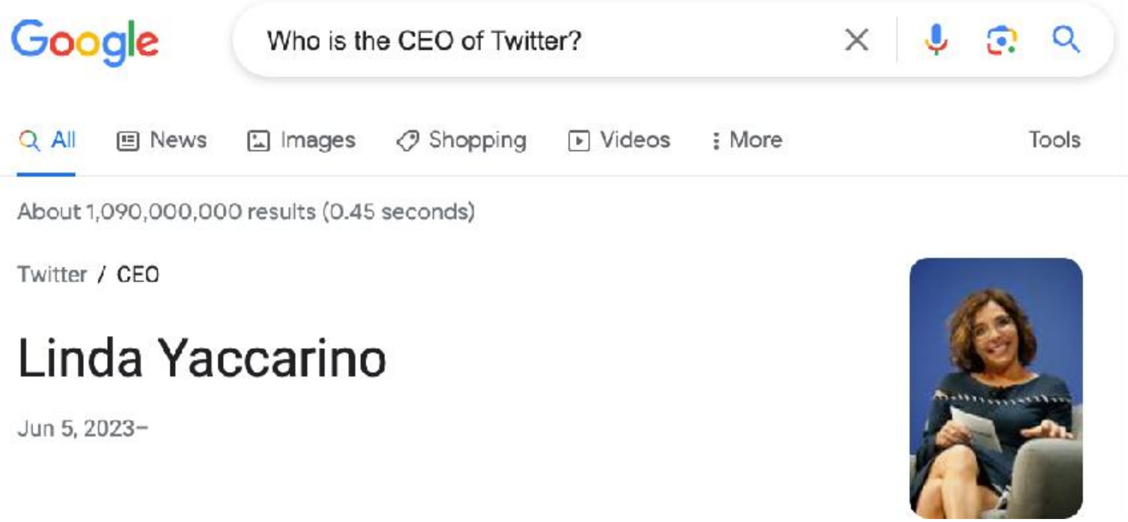
LLMs’ knowledge is easily outdated and hard to update

Why retrieval-based LMs?

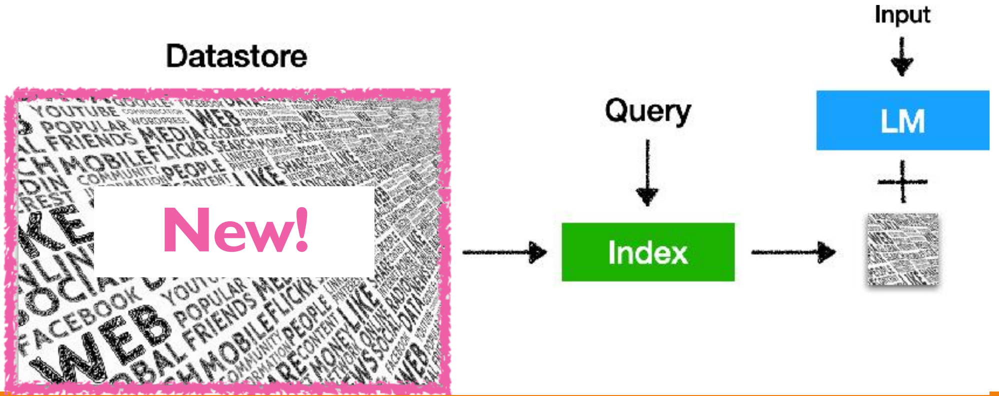
Who is the CEO of Twitter?

ChatGPT

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



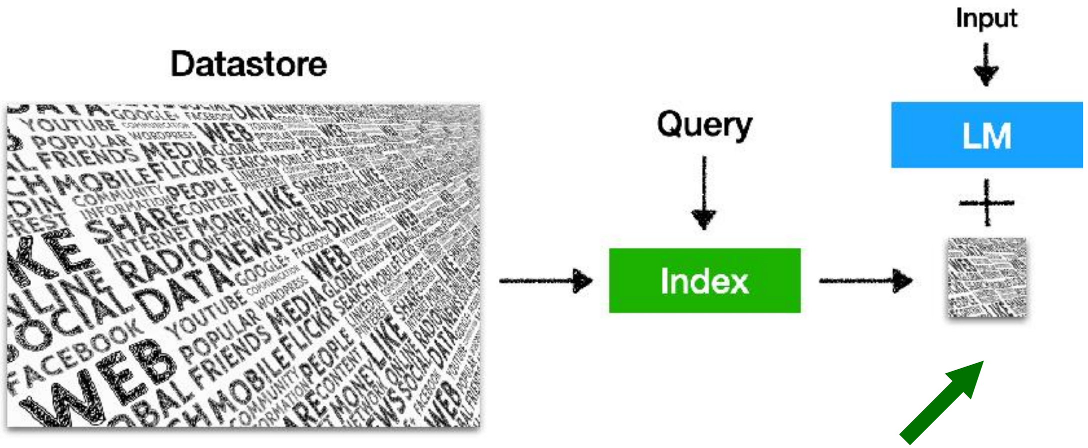
- Existing **knowledge editing** methods are still NOT scalable (active research!)
- The datastore can be easily **updated** and **expanded** - even without retraining!



LLMs’ output is challenging to interpret and verify

Why retrieval-based LMs?

Generating text with citations

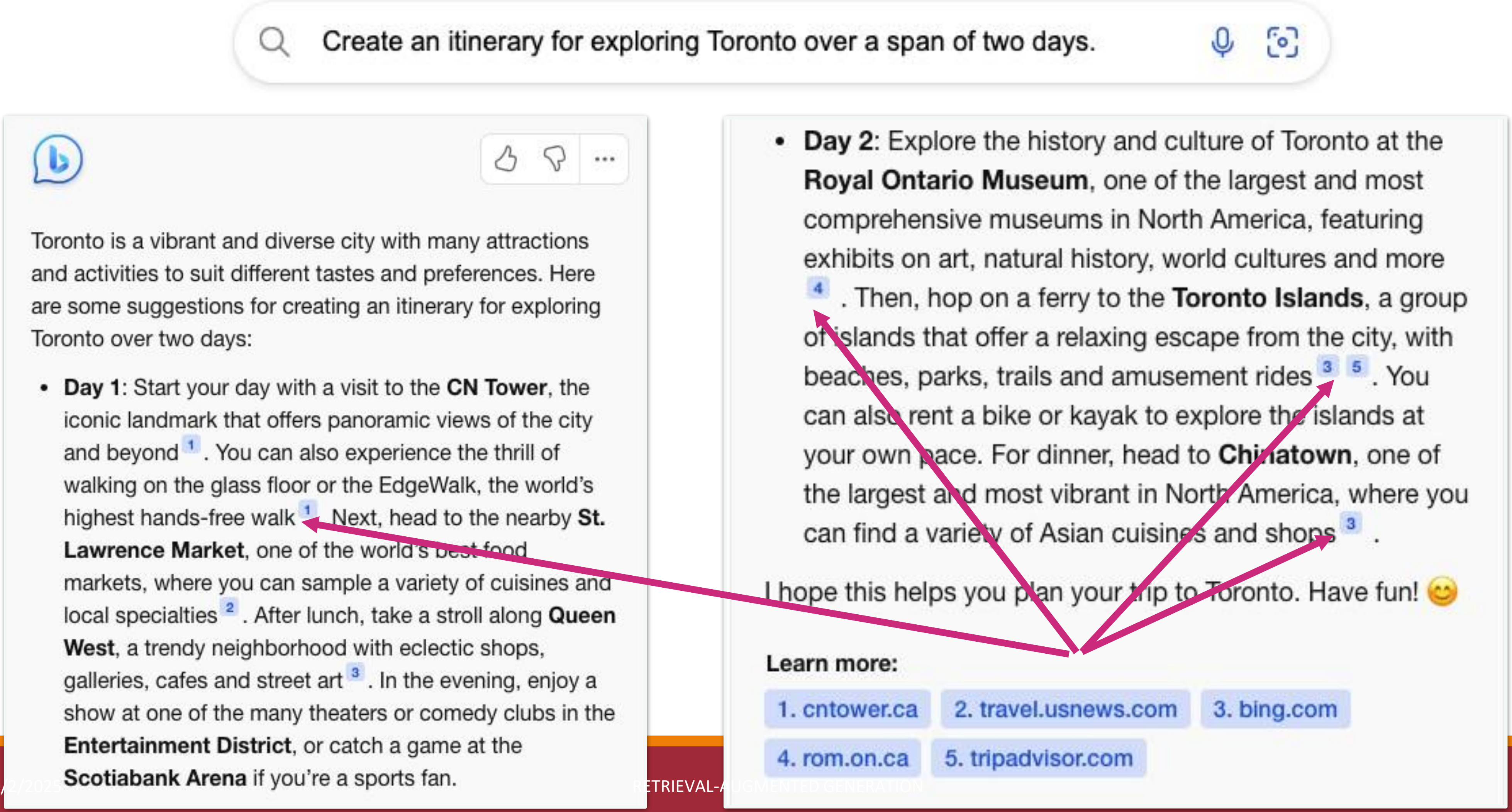


Can **trace knowledge source** from retrieval results - better **interpretability** & **control**

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

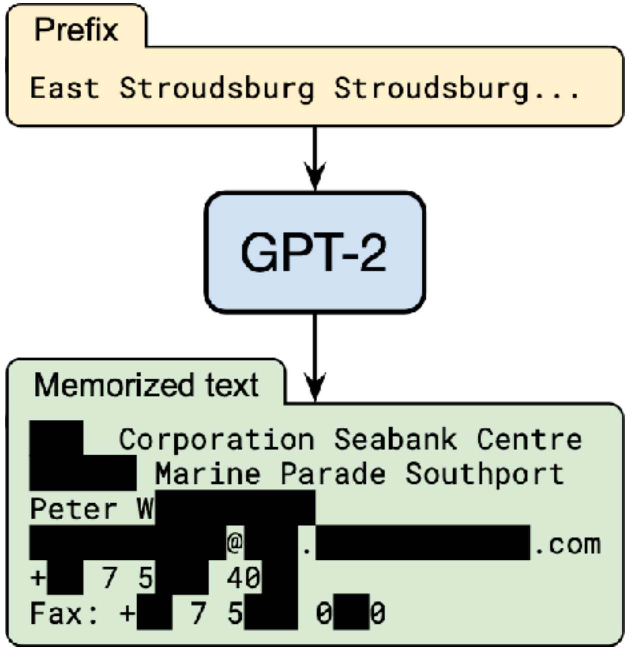
LLMs' output is challenging to interpret and verify

Why retrieval-based LMs?



LLMs are shown to easily leak private training data

Why retrieval-based LMs?



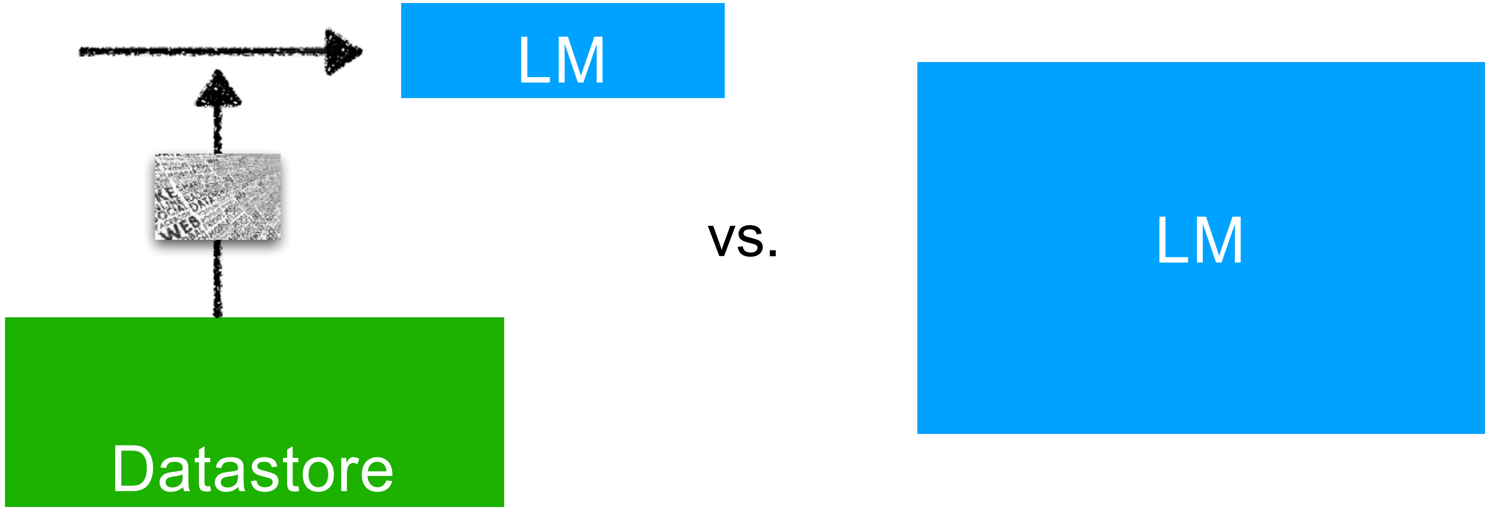
Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	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

Individualization on private data by storing it in the datastore

Carlini, N., *et al.* (2020). Extracting Training Data from Large Language Models. https://arxiv.org/abs/2012.07805

LLMs are *large* and expensive to train and run

Why retrieval-based LMs?



Long-term goal: can we possibly reduce the **training** and **inference costs,** and scale down the size of LLMs?

e.g., RETRO (Borgeaud et al., 2021): “obtains comparable performance to GPT-3 on the Pile, despite using **25x fewer parameters**”

A Retrieval-based LM: Definition

A language model (LM) that uses
an external datastore at test time

Typical LMs

The capital city of Ontario is **Toronto**

LM

Training time

The capital city of Ontario is _____

LM

Test time / Inference

Retrieval-based LMs

The capital city of Ontario is **Toronto**

LM

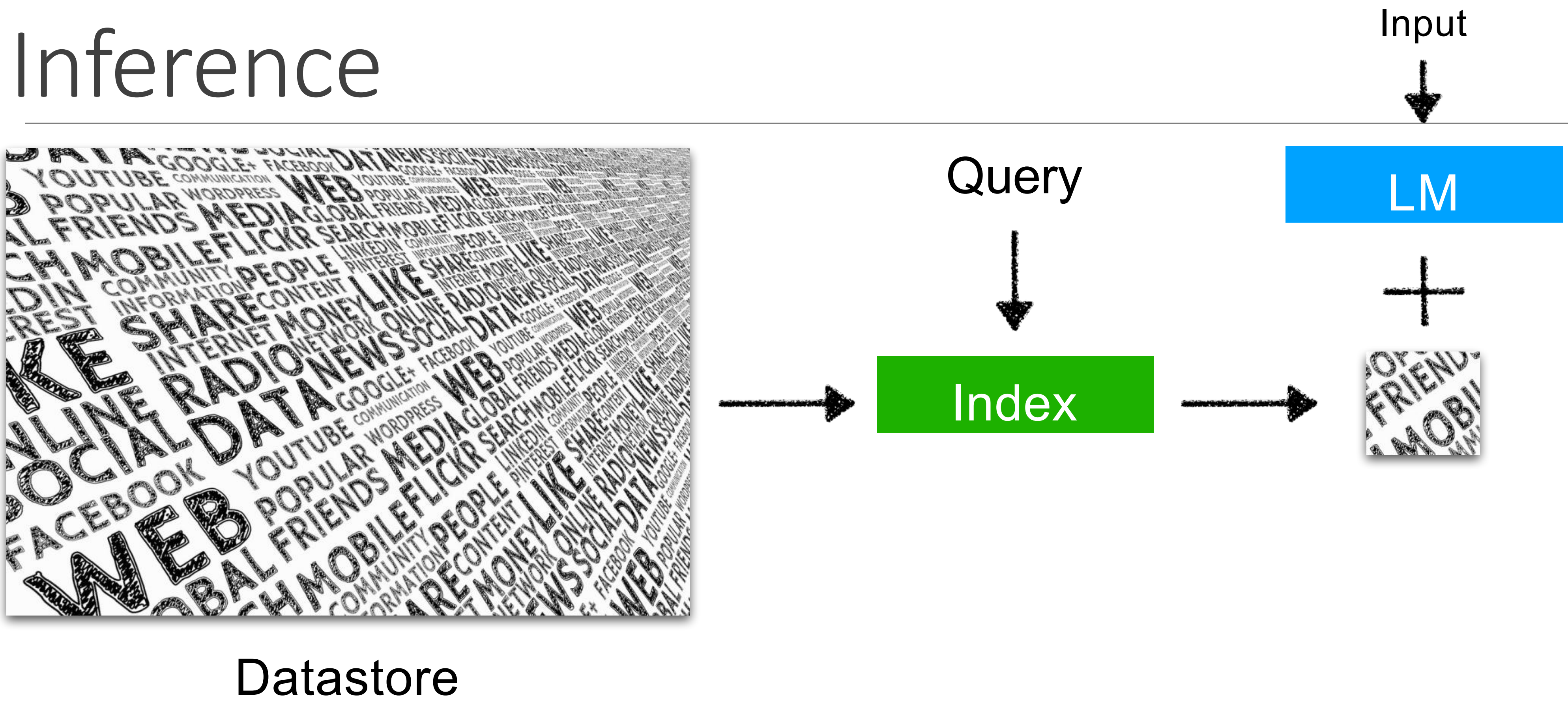
Training time

The capital city of Ontario is _____

LM

Test time / Inference

Inference



Inference: Datastore

Input

LM

+

Query

Index

More recently people have used structured data

Datastore

Raw text corpus

(originally)

At least billions~trillions of tokens

Not labeled datasets

Not structured data (knowledge bases)

Inference: Index

Input

Retrieval input
(not necessarily input to the LM)

Query

LM

Datastore

Index

Find a small subset of elements in a datastore that are the most similar to the query

Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Example

$$\mathrm{sim}(i,j) = \mathrm{tf}_{i,j} \times \log \frac{N}{\mathrm{df}_i}$$

# of total docs

# of docs containing

# of occurrences of in j

Remember cosine similarity from our discussion of word embeddings

$$\frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{b}|}$$

Example

$$\mathrm{sim}(i,j) = \mathrm{Encoder}(i) \cdot \mathrm{Encoder}(j)$$

Maps the text into an -dimensional vector

Inference: Index

Goal: find a small subset of elements in a datastore that are the most similar to the query

sim: a similarity score between two pieces of text

Can be a totally separate research area on how to do this fast & accurate

Index: given q , return $\underbrace{\text{argTop-}k_{d \in \mathcal{D}} \text{sim}(q, d)}_{k \text{ elements from a datastore}}$ through fast nearest neighbor search

Software: FAISS, Distributed FAISS, SCaNN, etc...

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive	Comments
Exact Search for L2	`IndexFlatL2`	`"Flat"`	`d`	`4*d`	yes	brute-force
Exact Search for Inner Product	`IndexFlatIP`	`"Flat"`	`d`	`4*d`	yes	also for cosine (normalize vectors beforehand)
Hierarchical Navigable Small World graph exploration	`IndexHNSWFlat`	`"HNSW,Flat"`	`d`, `M`	`4*d + x * M * 2 * 4`	no	
Inverted file with exact post-verification	`IndexIVFFlat`	`"IVFx,Flat"`	`quantizer`, `d`, `nlists`, `metric`	`4*d + 8`	no	Takes another index to assign vectors to inverted lists. The 8 additional bytes are the vector id that needs to be stored.
Locality-Sensitive Hashing (binary flat index)	`IndexLSH`	-	`d`, `nbits`	`ceil(nbits/8)`	yes	optimized by using random rotation instead of random projections
Scalar quantizer (SQ) in flat mode	`IndexScalarQuantizer`	`"SQ8"`	`d`	`d`	yes	4 and 6 bits per component are also implemented.
Product quantizer (PQ) in flat mode	`IndexPQ`	`"PQx"`, `"PQ"M"x"nbits`	`d`, `M`, `nbits`	`ceil(M * nbits / 8)`	yes	
IVF and scalar quantizer	`IndexIVFScalarQuantizer`	"IVFx,SQ4" "IVFx,SQ8"	`quantizer`, `d`, `nlists`, `qtype`	SQfp16: 2 * `d` + 8, SQ8: `d` + 8 or SQ4: `d/2` + 8	no	Same as the `IndexScalarQuantizer`
IVFADC (coarse quantizer+PQ on residuals)	`IndexIVFPQ`	`"IVFx,PQ"y"x"nbits`	`quantizer`, `d`, `nlists`, `M`, `nbits`	`ceil(M * nbits/8)+8`	no	
IVFADC+R (same as IVFADC with re-ranking based on codes)	`IndexIVFPQR`	`"IVFx,PQy+z"`	`quantizer`, `d`, `nlists`, `M`, `nbits`, `M_refine`, `nbits_refine`	`M+M_refine+8`	no	

Exact Search

Approximate Search
(Relatively easy to scale to ~1B elements)

More info: https://github.com/facebookresearch/faiss/wiki

Inference: Search

Input

LM

+

Query

Index

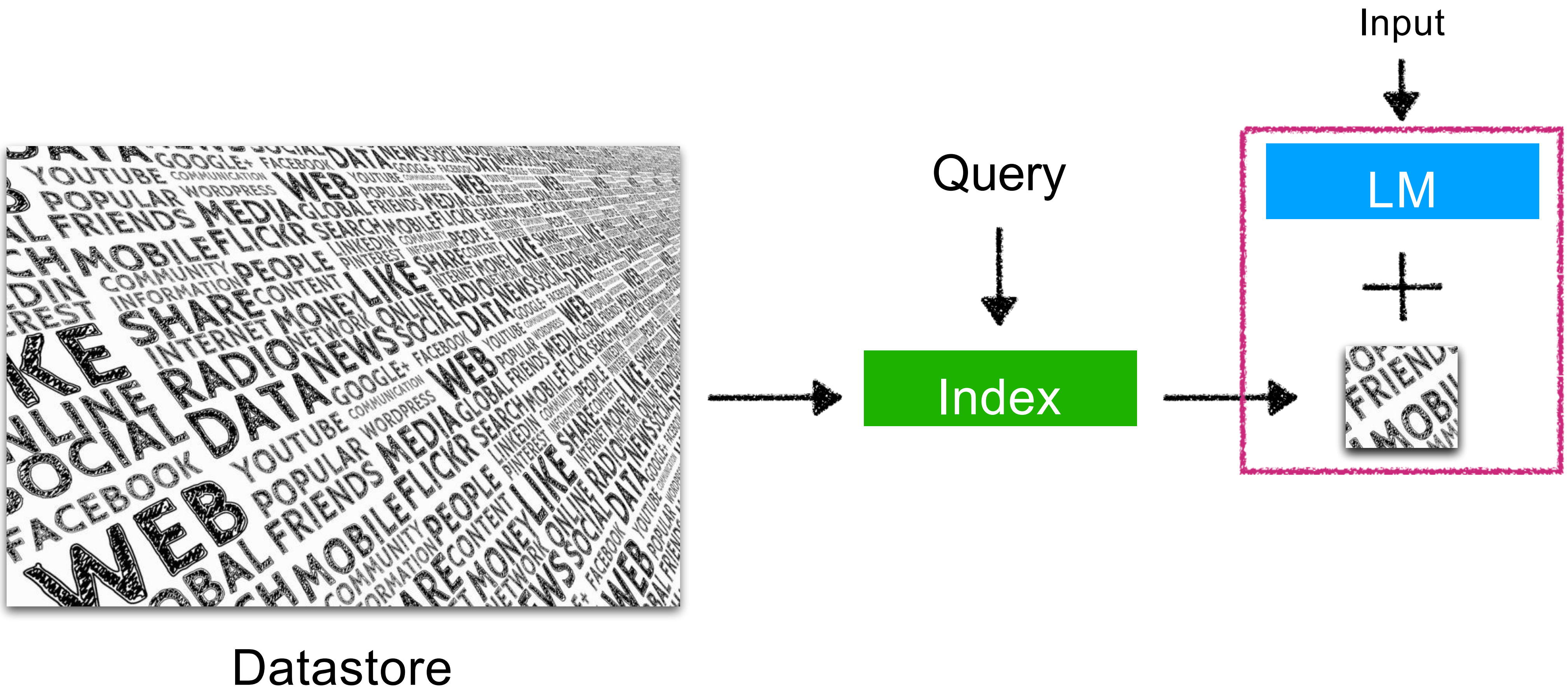
fast nearest neighbor search

In this tutorial, we assume we can do it fast & accurate

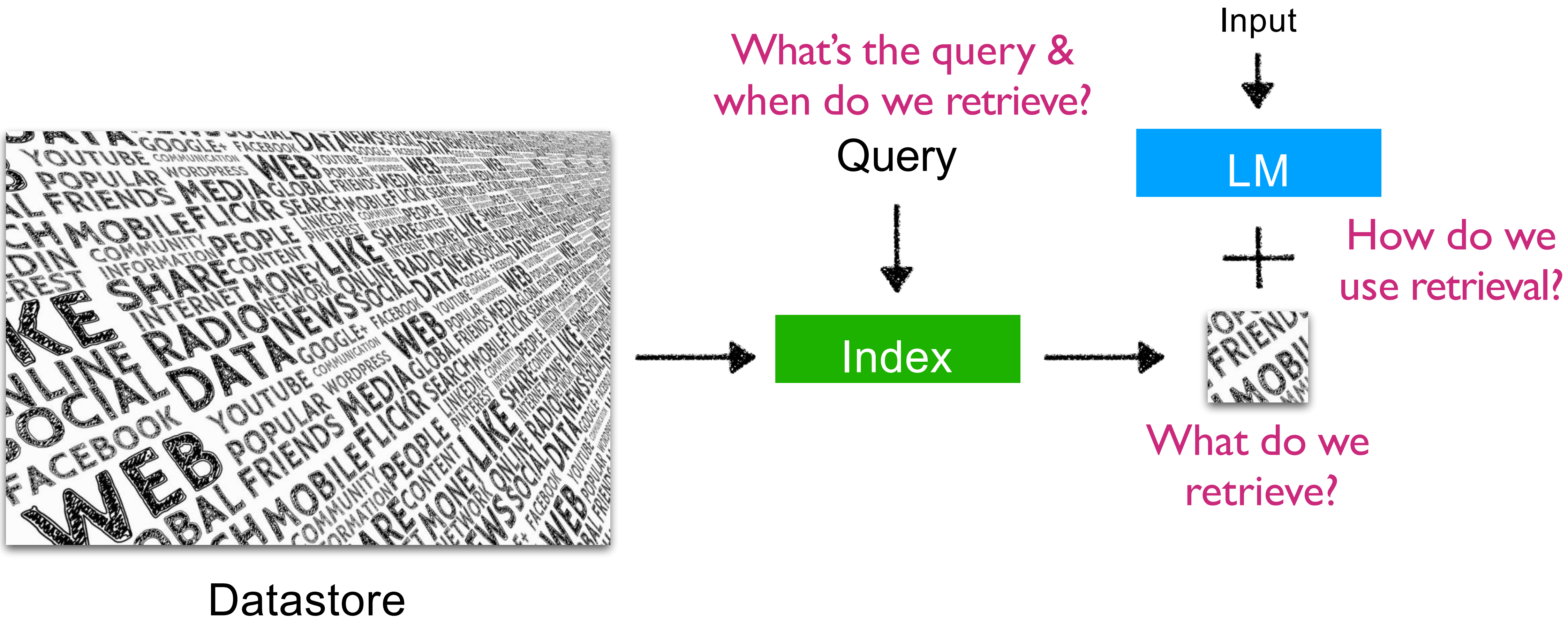
Datastore

Retrieval

Inference: Search



Variations of RAG



Variations of RAG

What to retrieve?

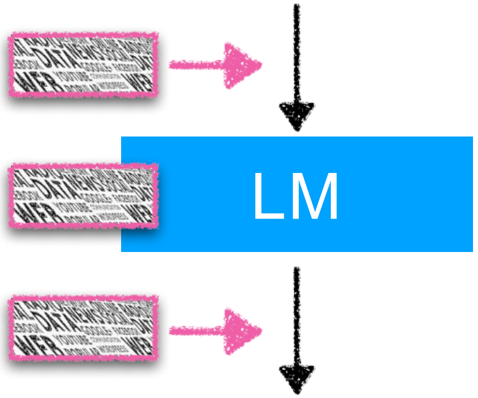
Query

Text chunks (passages)?

Tokens?

Something else?

How to use retrieval?



When to retrieve?

In-Class Activity

This is also an exercise for reading academic papers to look for specific details

Skim the paper assigned to you

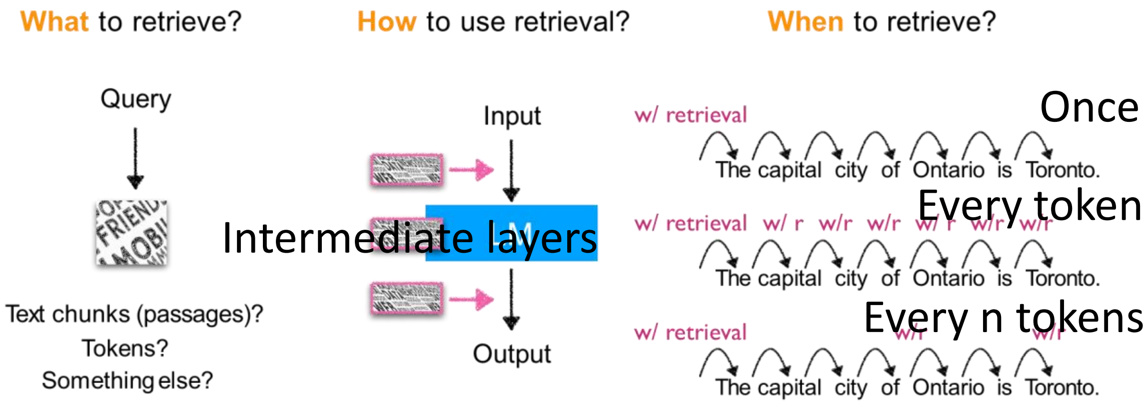
In your paper, find the answers to these questions

What to retrieve?

How to use retrieval?

When to retrieve?

Share what you learned with your table



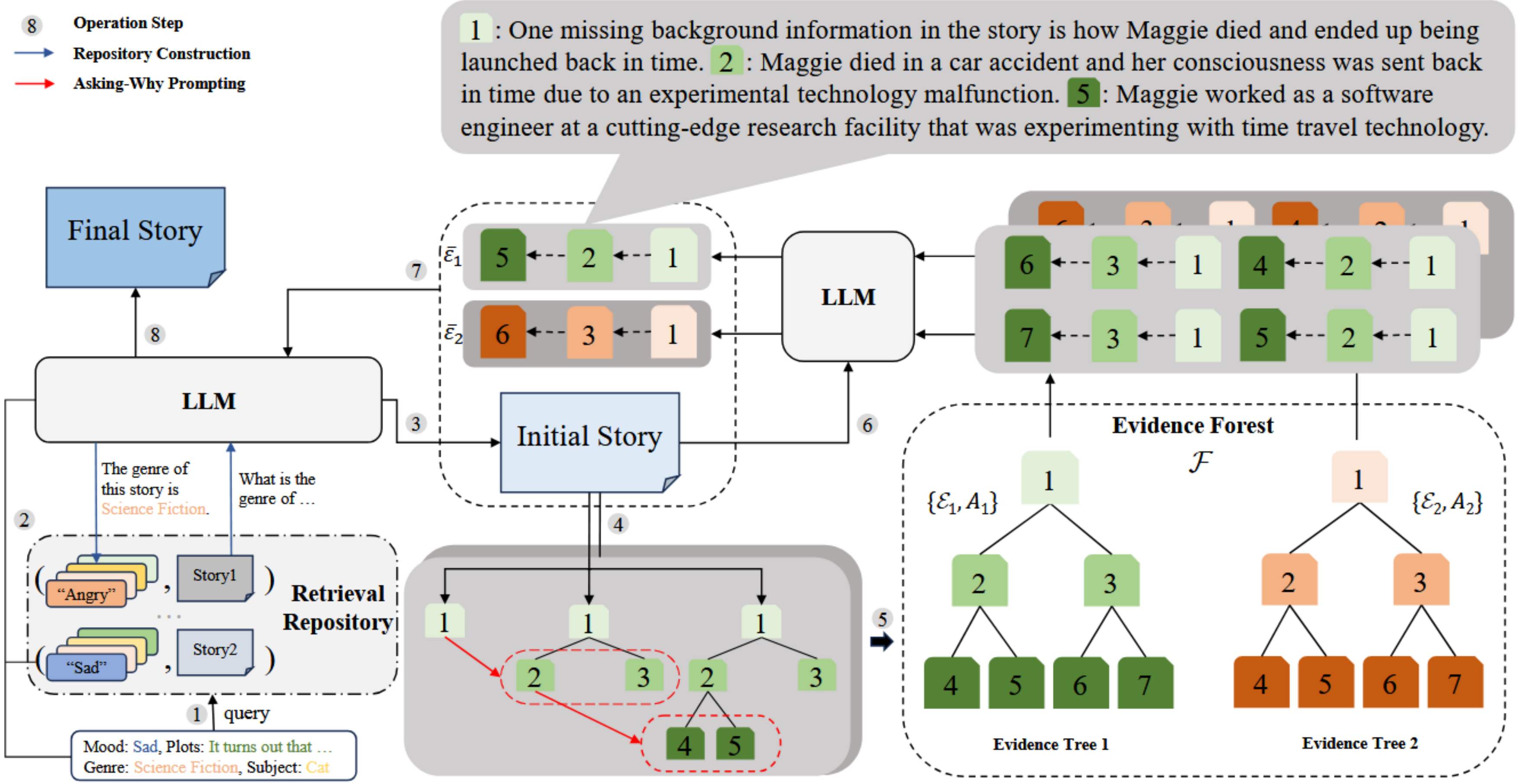
Say Anything (2009)

User written sentence	*Matching sentence*	*Computer generated sentence*
It all started easy enough.	It all started out normal enough.	Got up, had lunch, got ready went to work etc.
But then I noticed something very strange.	Then, as I flipped through the test, I realized something very strange.	Mr. Sipiera gave me points on the blank page.
Why would he do that? What was he up to?	Why would he do that?	Because he was planning to kill Sonya for her knowledge of the kidnapping plot.
I decided to get to Sonya first and run away with her to a Greek island.	We left pretty much after this and took a ferry back to Hong Kong island.	Our original plan to go to Lamma island was ditched, as we had inadvertently spent half the day on Lantau island.
But we settled for Corfu, which has a beautiful little hotel on the edge of the old town.	But we settled for the simpler name.	We had dogs (and cats) when I was growing up but my wife and sons have never had a dog before.
Corfu was full of cats and dogs, and I ended up having to look at my feet a lot when walking around town.	and i ended up having to save the day.	It's too long to list the details of the whole operation.
Sonya would take on the secret identity of a scuba instructor, and I'd be a fisherman.	"Before I even got so far as having a proper breakdown in realizing that fantasy is not reality, I found that I had the secret identity of a beetle."	"But that could have been very useful, right?
It meant we could spend a lot of secret time together on the jetty beyond the coastline.	She does spend a lot of time at the bar where they all hang out.	She has told me some of the men she hangs with have come on to her.
I told her I'd kill them just like that old math teacher.	I told her I'd come help her clean and whatever else she needs.	Once she's gone, I'm pretty much on my own.
Fishing all day, looking out for the police.	This is bad for the police.	If they stopped to survey the situation they could only see what was there -- none of the men were armed.
It was a great secret life.		

Before neural methods, retrieval was used for generating stories

Swanson, R. and Gordon, A. (2009) Open Domain Collaborative Storytelling With Say Anything. Third International Conference on Weblogs and Social Media, San Jose, CA, May 17-20, 2009.

GROVE (2023)



Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. GROVE: A Retrieval-augmented Complex Story Generation Framework with A Forest of Evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998, Singapore. Association for Computational Linguistics.

BERALL (2024) (to be presented on Tuesday)