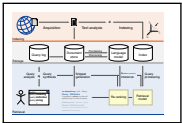
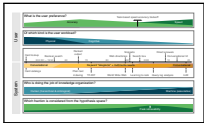
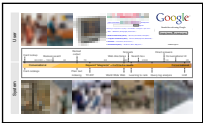


Retrieval Technologies for the Infinite Index

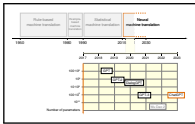
Martin Potthast
University of Kassel,
hessian.AI, and ScaDS.AI

March 5th, 2025 • BTW 2025

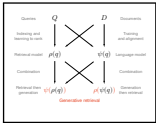
Web Search



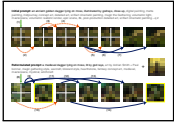
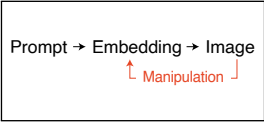
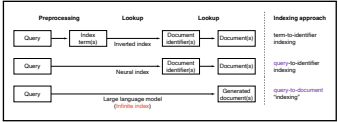
Language Models



Retrieval-Augmented Generation



The Infinite Index

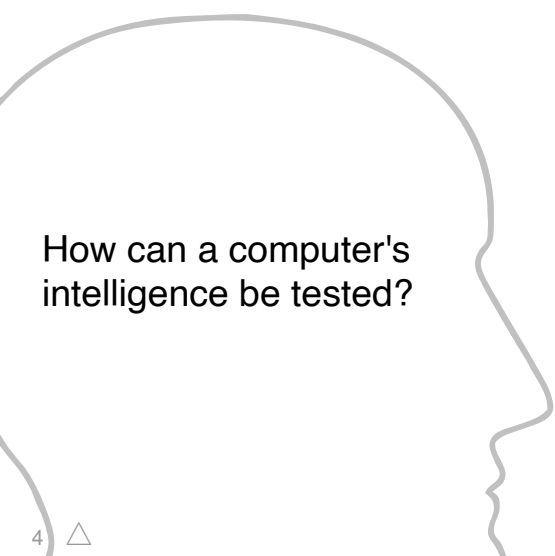




A Short History of Search Engines

Information Retrieval in a Nutshell

- ❑ **A vague request.**
Expression of a complex information need: a question
- ❑ **Billions of documents.**
Text, images, audio files, videos, ...



How can a computer's intelligence be tested?



Information Retrieval in a Nutshell

- ❑ A vague request.

Expression of a complex information need: a question, or just a few keywords.

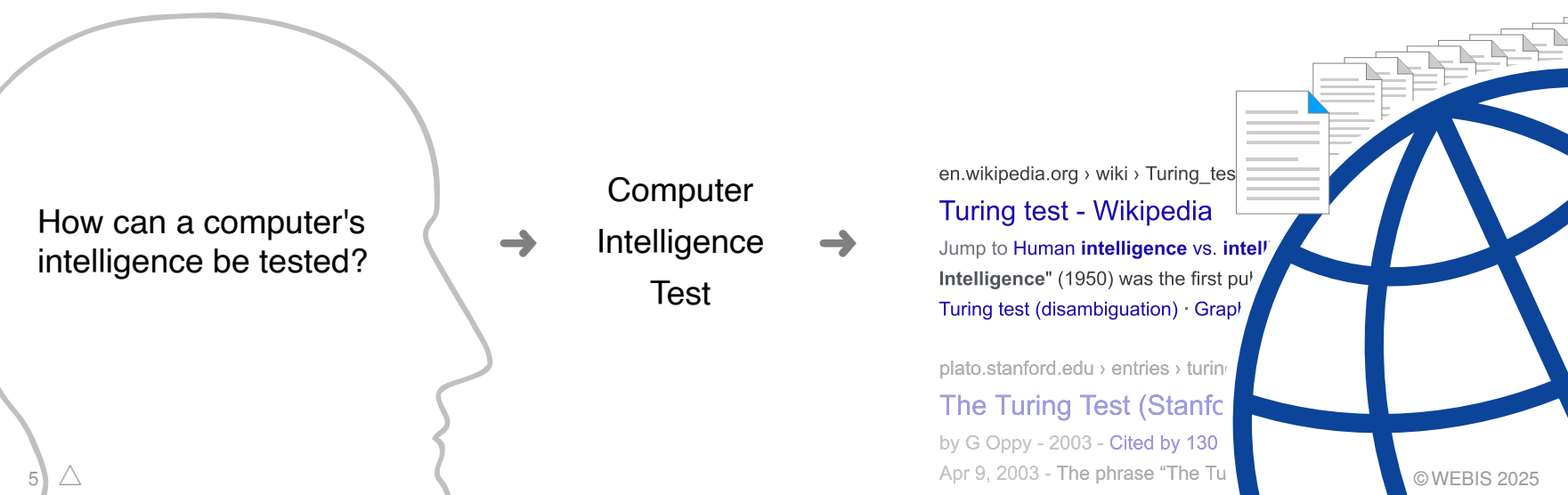
- ❑ Billions of documents.

Text, images, audio files, videos, ...

- ❑ High class imbalance.

Only a tiny fraction of all documents are relevant to the request.

→ Retrieve relevant documents in milliseconds.



How can a computer's intelligence be tested?

Computer
Intelligence
Test

en.wikipedia.org › wiki › Turing_test

Turing test - Wikipedia

Jump to **Human intelligence vs. artificial intelligence** · **"Intelligence"** (1950) was the first published Turing test (disambiguation) · Graph

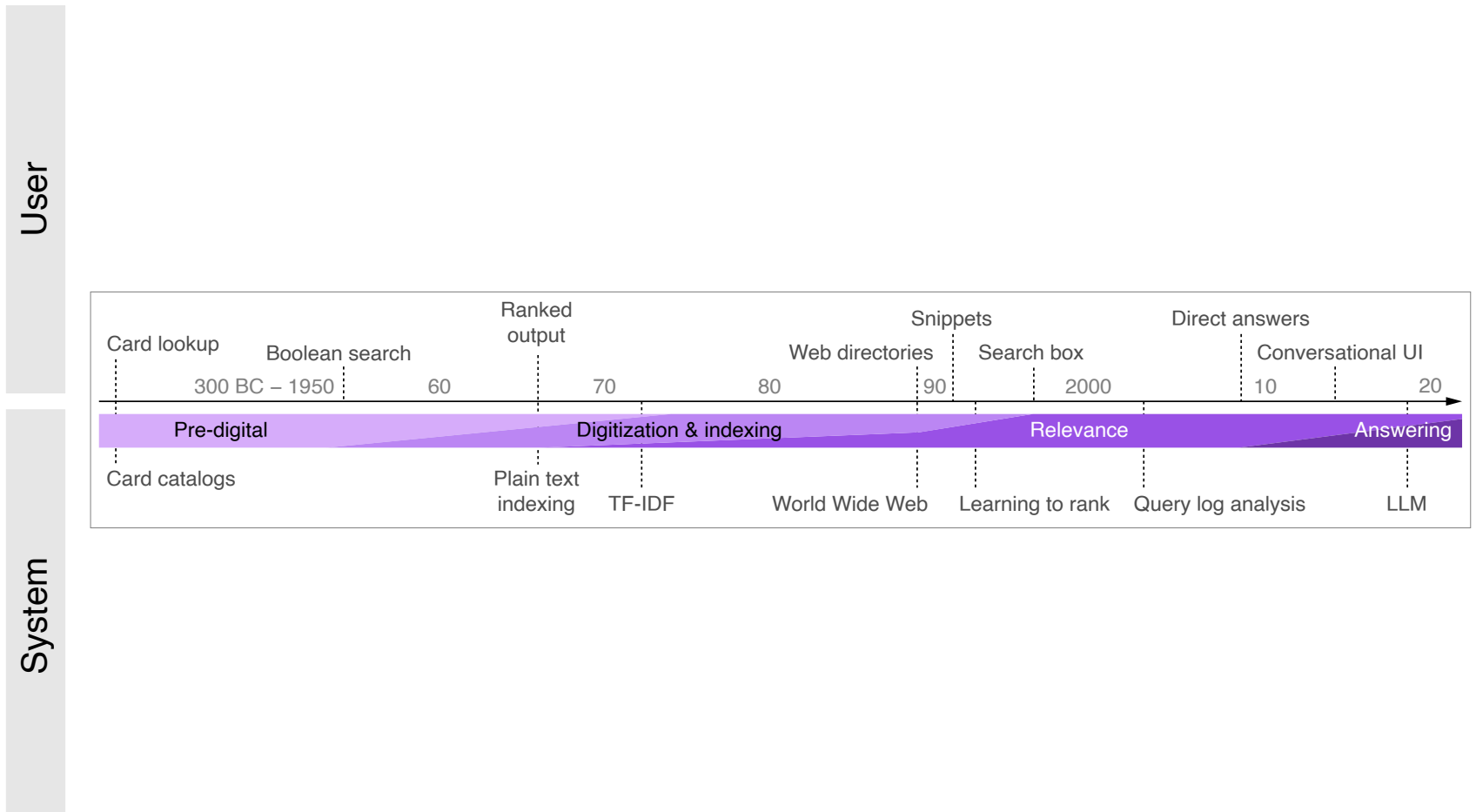
plato.stanford.edu › entries › turing_test

The Turing Test (Stanford)

by G Oppy - 2003 - Cited by 130

Apr 9, 2003 - The phrase "The Turing test"

A Short History of Search Engines

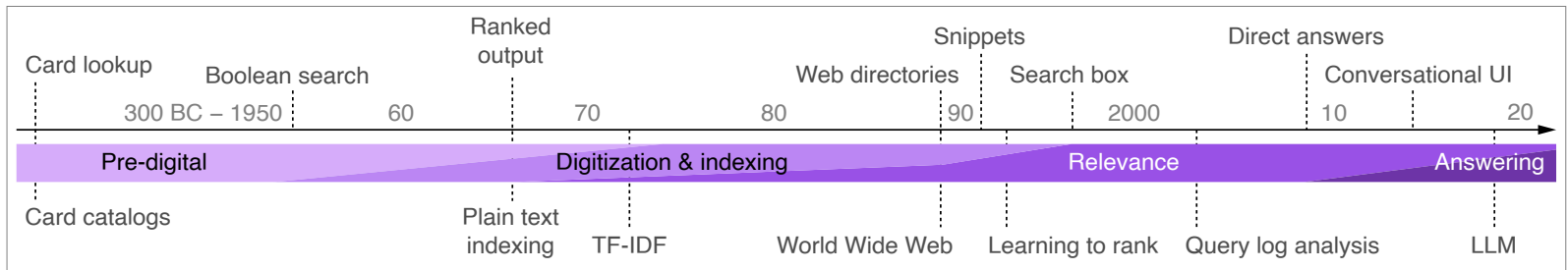


- M. Potthast, M. Hagen, B. Stein (2020). [The dilemma of the direct answer.](#)

A Short History of Search Engines

User

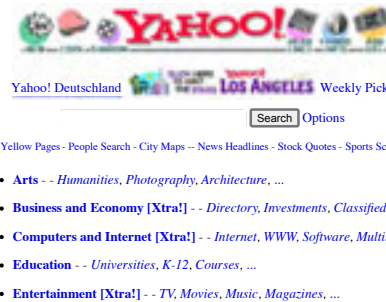
System



- M. Potthast, M. Hagen, B. Stein (2020). [The dilemma of the direct answer.](#)

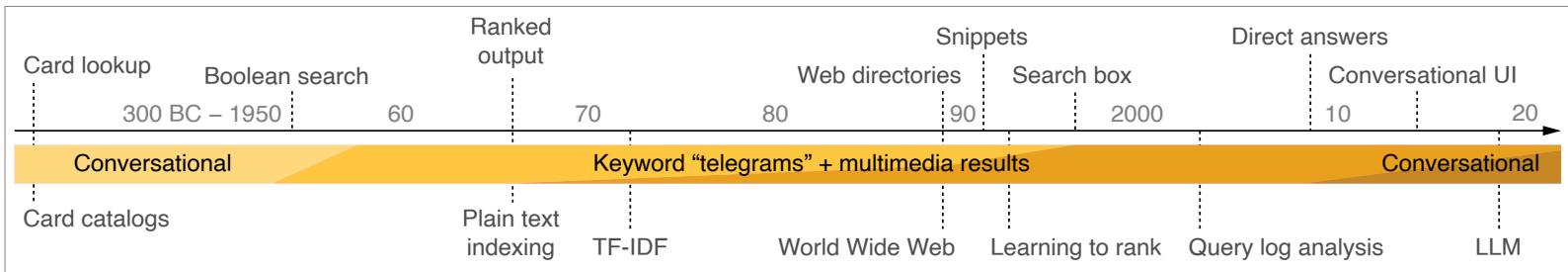
A Short History of Search Engines

User



Search the web using Google

©1999 Google Inc.



System

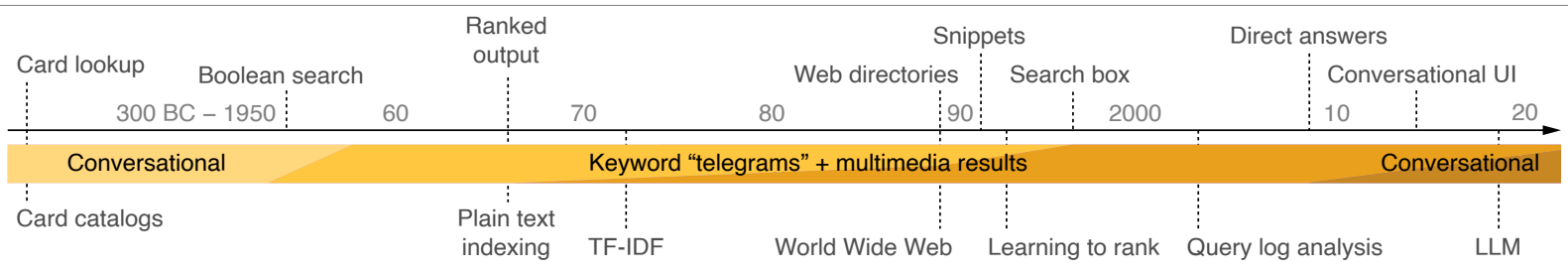


- M. Potthast, M. Hagen, B. Stein (2020). The dilemma of the direct answer.

A Short History of Search Engines

User

Of which kind is the user workload?



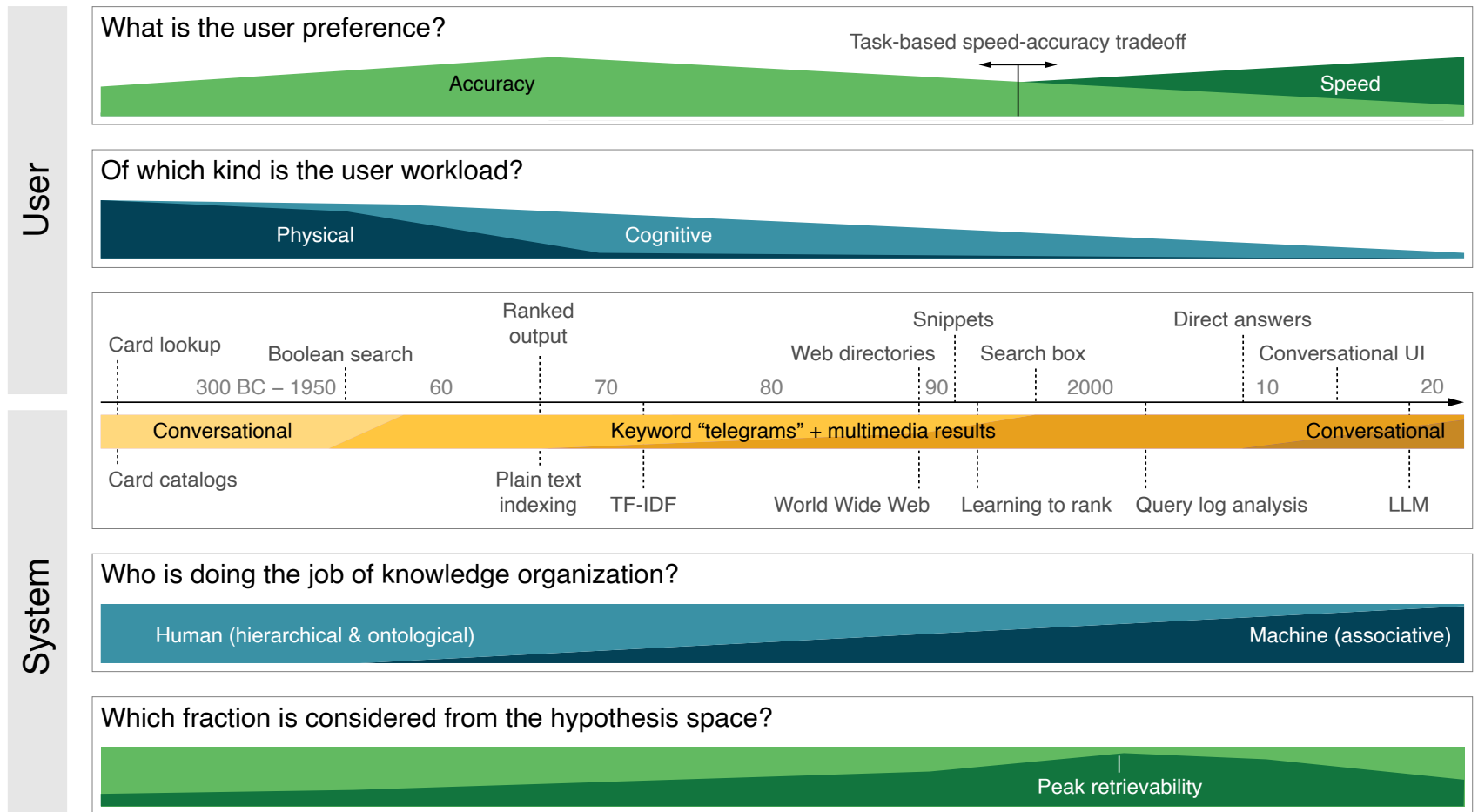
System

Who is doing the job of knowledge organization?



- M. Potthast, M. Hagen, B. Stein (2020). [The dilemma of the direct answer.](#)

A Short History of Search Engines



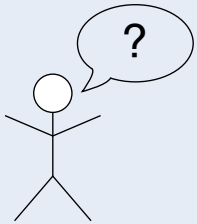
- M. Potthast, M. Hagen, B. Stein (2020). [The dilemma of the direct answer.](#)

Web Search Architecture



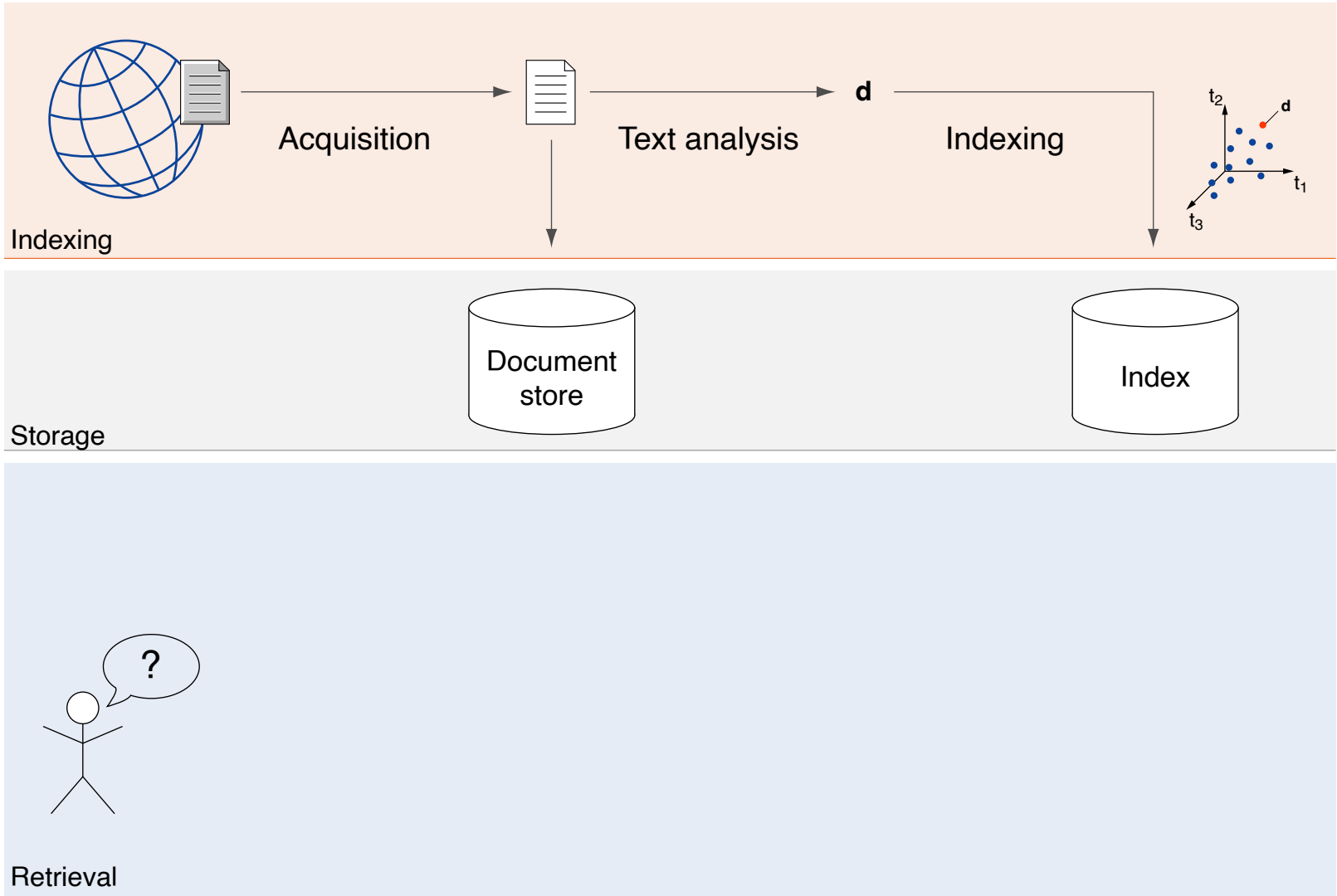
Indexing

Storage

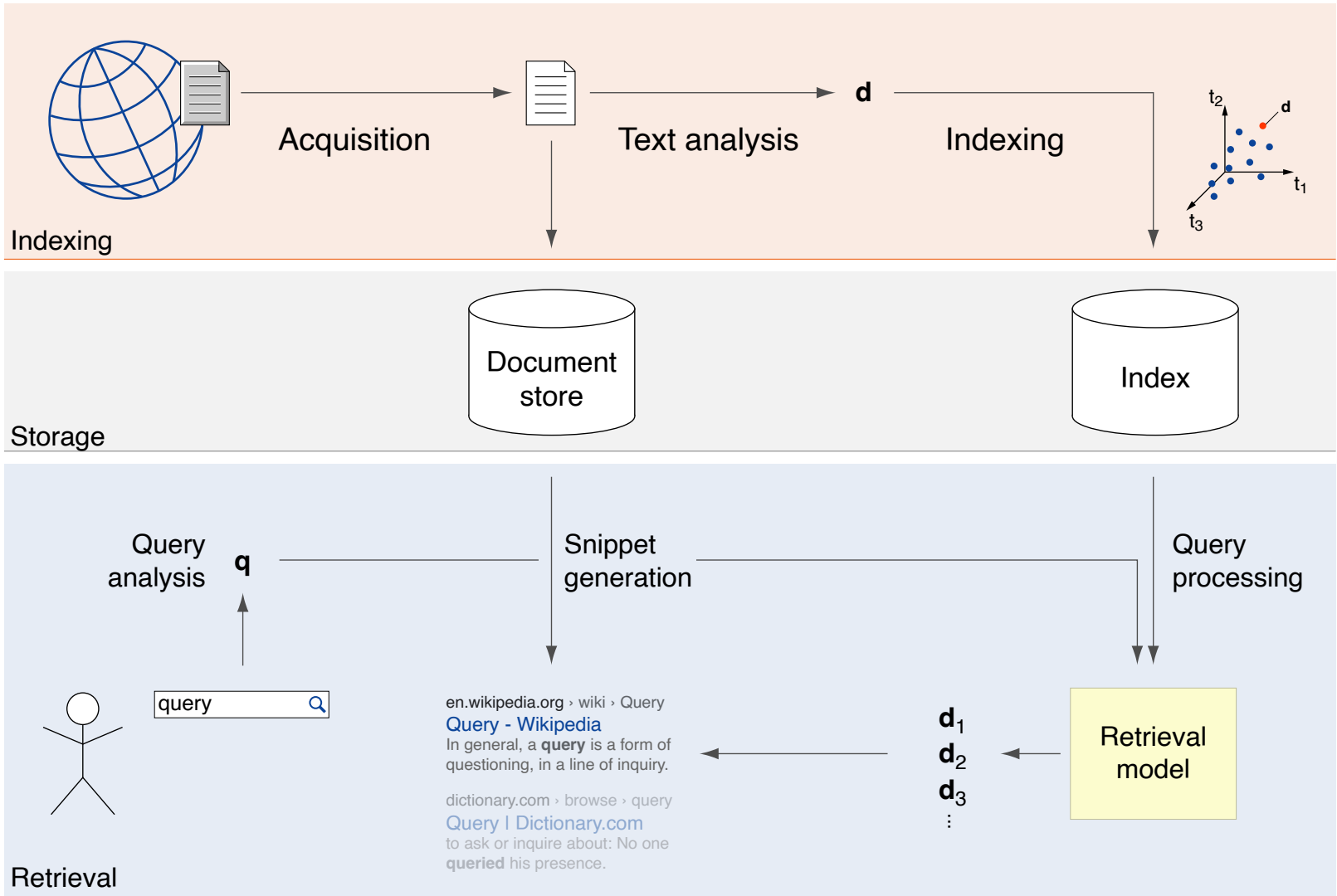


Retrieval

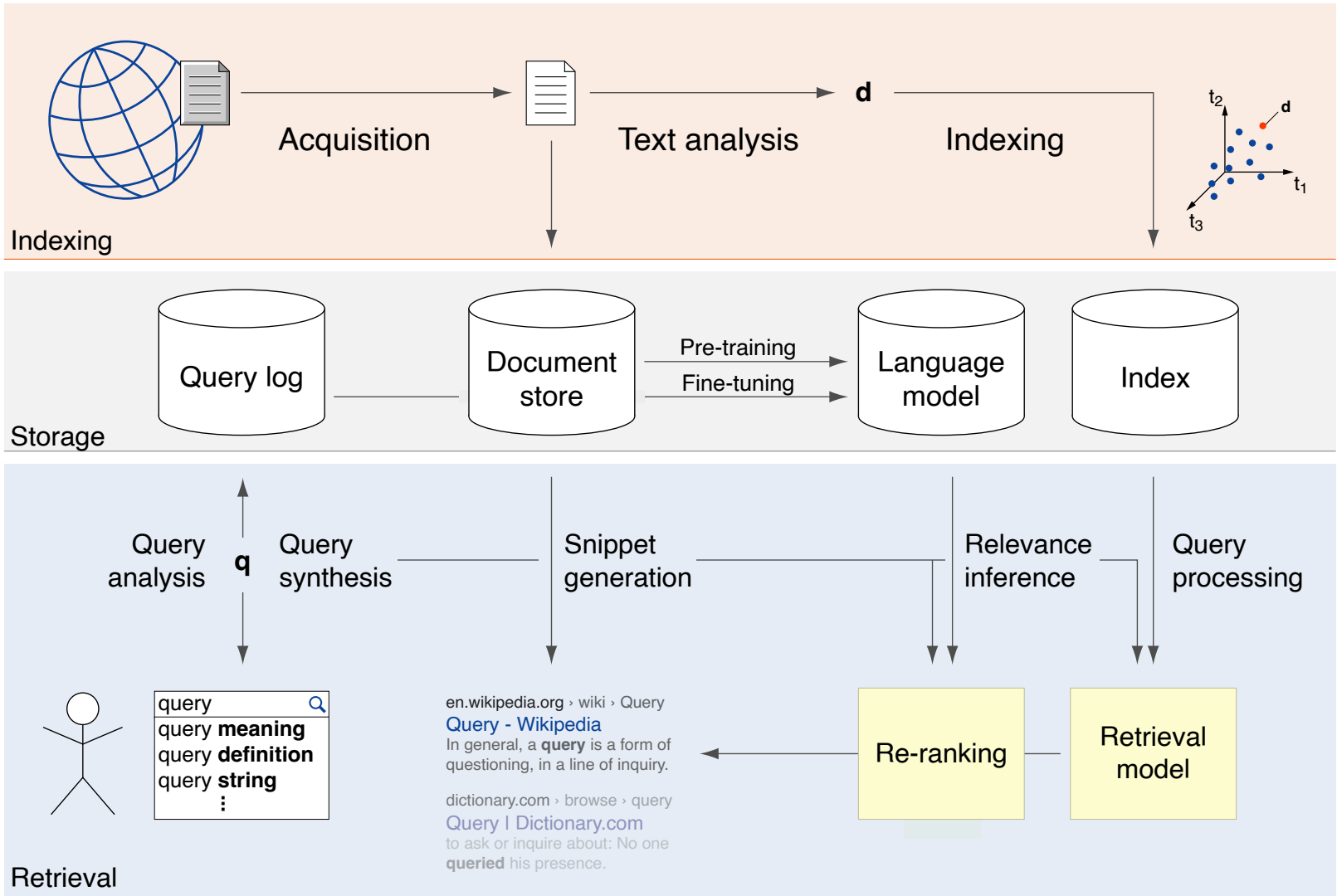
Web Search Architecture



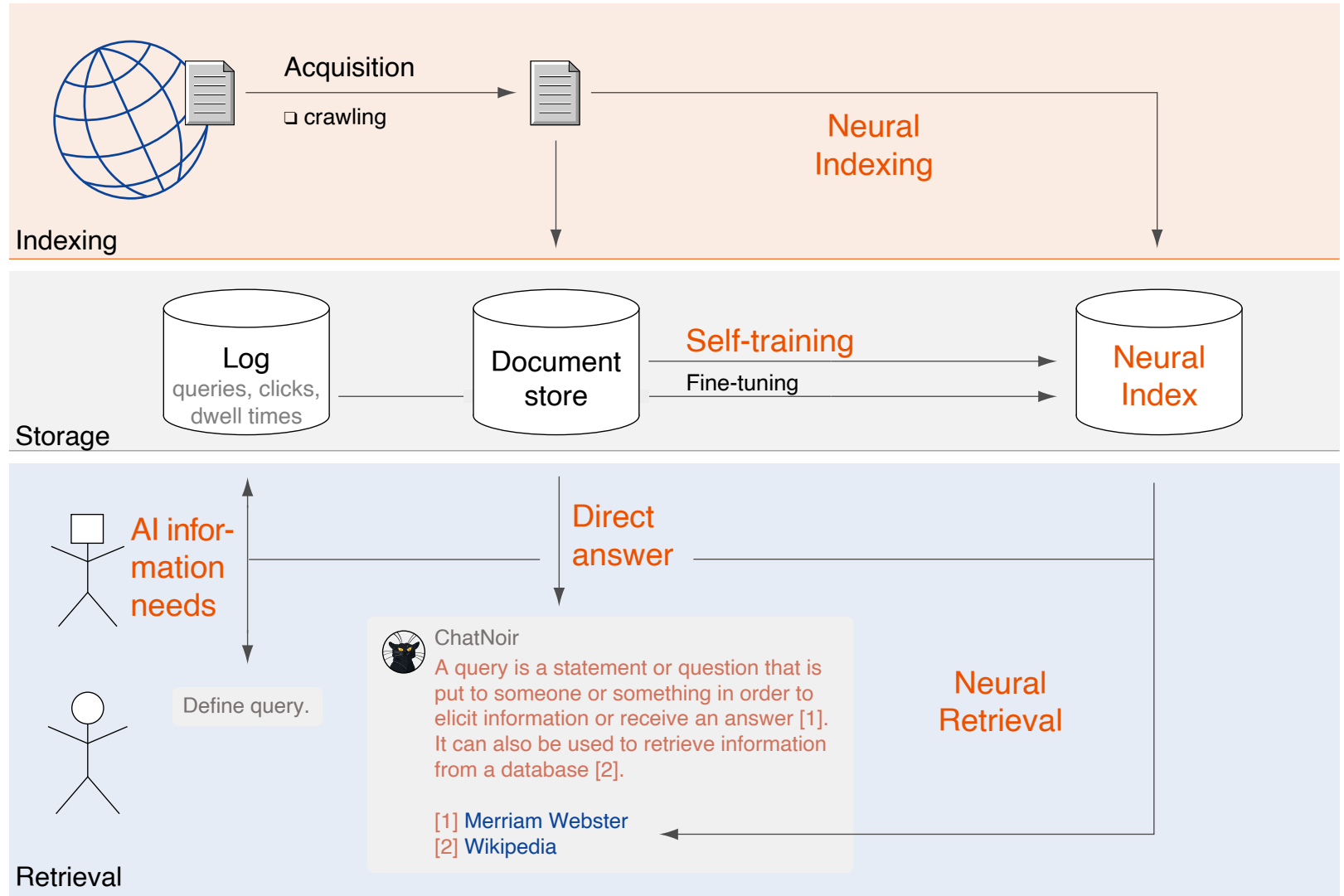
Web Search Architecture



Web Search Architecture



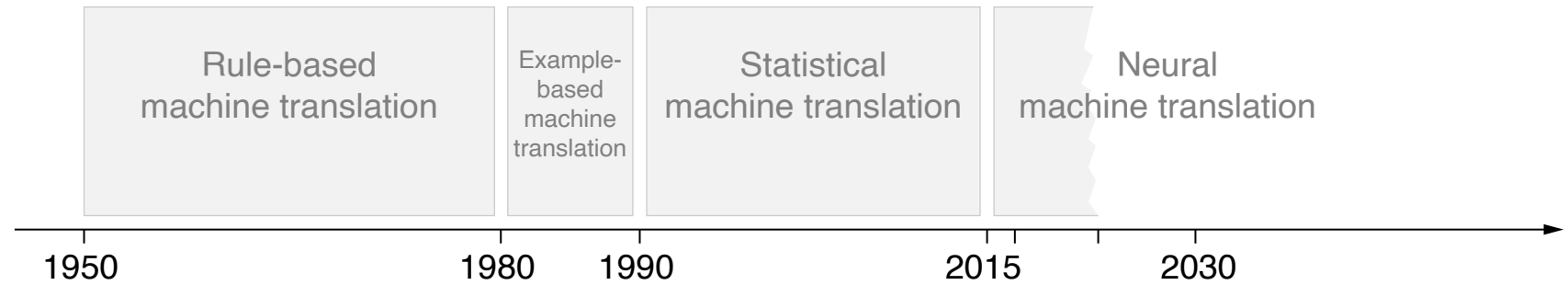
Web Search Architecture



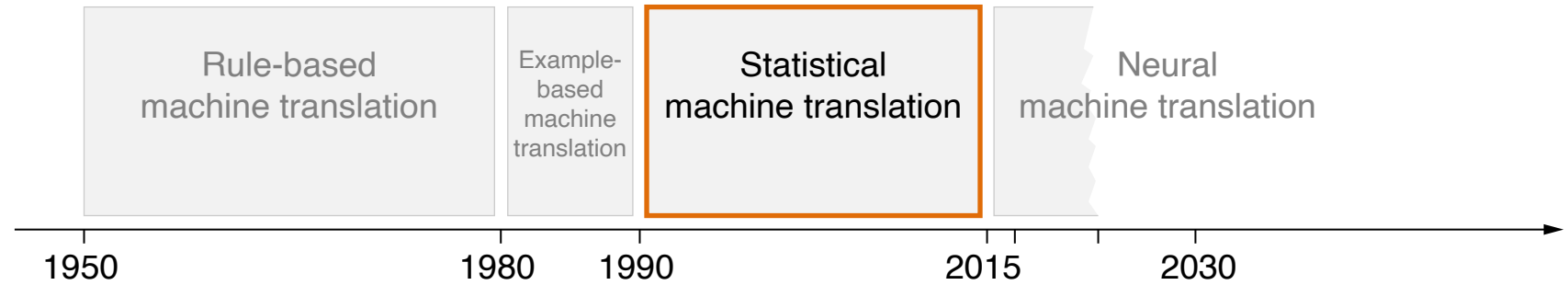


A Short History of Language Models

A Short History of Language Models

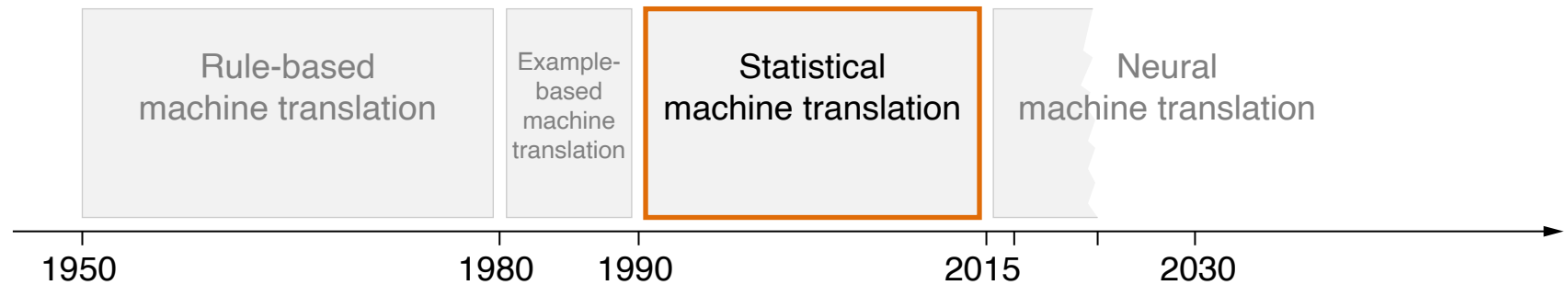


A Short History of Language Models



A statistical language model
is a probability distribution over all possible texts.

A Short History of Language Models



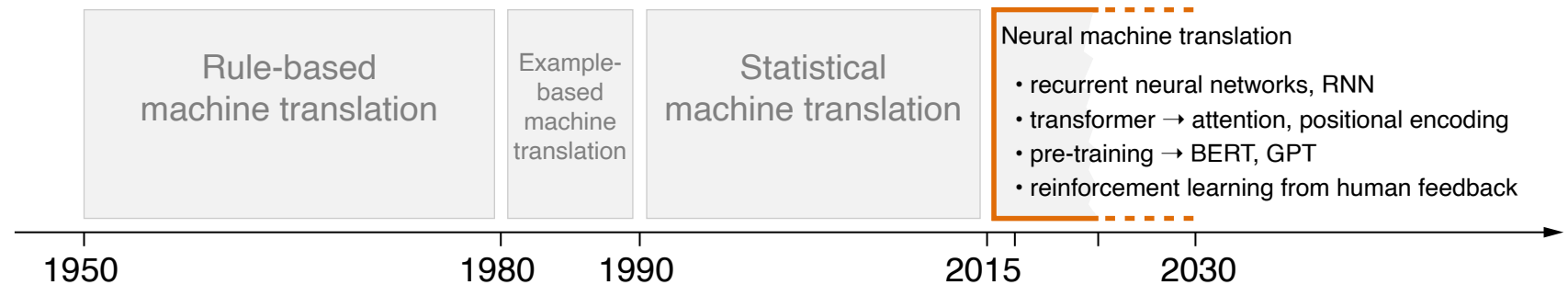
A statistical language model
is a probability distribution over all possible texts.

Illustration:

(1) i love my ? N N

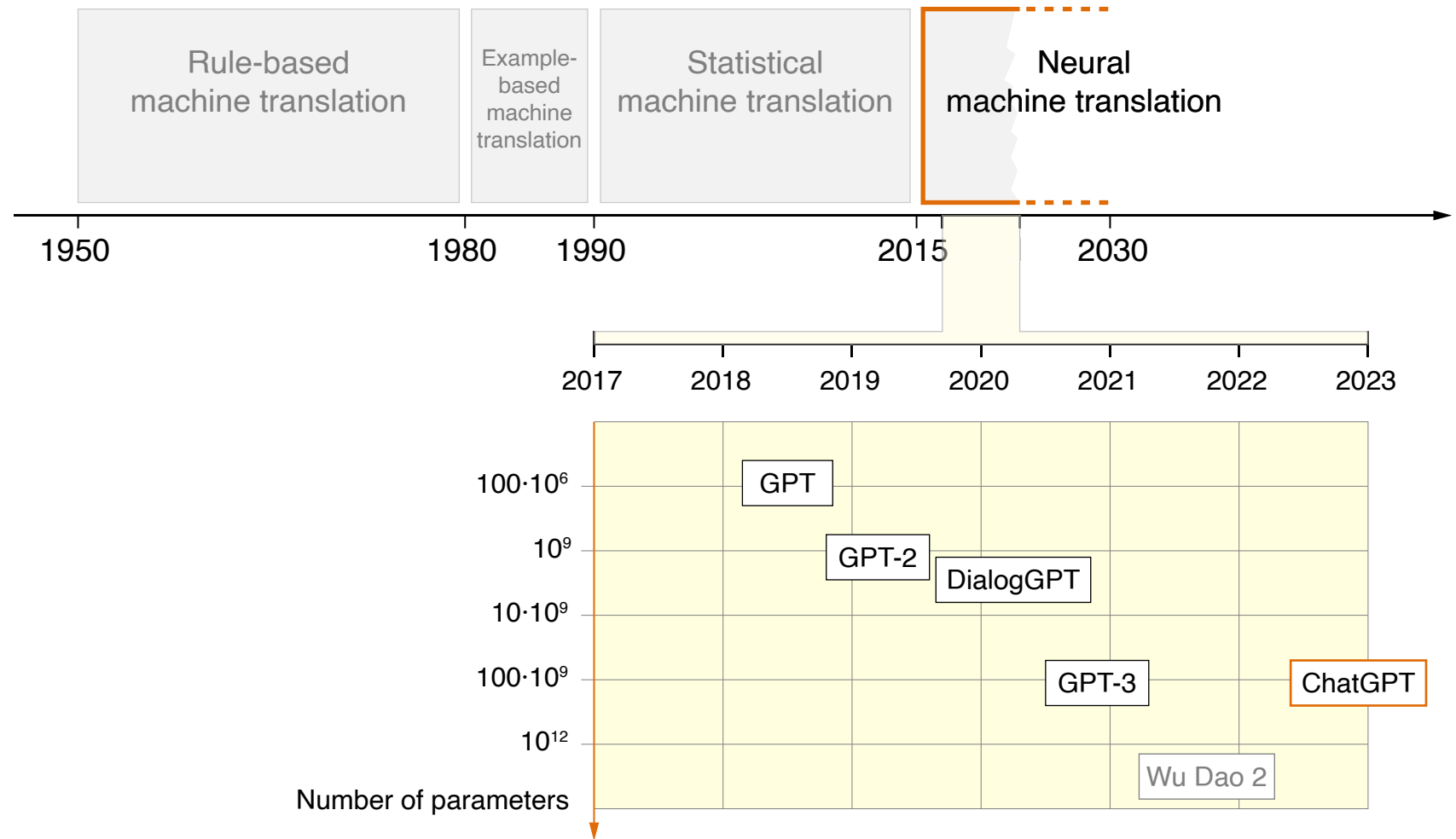
(2) see ... works. N N

A Short History of Language Models



A neural language model
approximates a statistical language model.

A Short History of Language Models



Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- ≈ 34 days on 1,024 x A100 GPUs.
- \$4.6M costs a single training run.

GPT-3 [Jun. 2020]

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- ≈ 34 days on 1,024 x A100 GPUs.
- \$4.6M costs a single training run.

GPT-3 [Jun. 2020]

World Knowledge



What city is in the northwest corner of Ohio?



Toledo is in the north-west corner of Ohio.

Common Sense



Why don't animals have three legs?



Animals don't have three legs because they would fall over.

Logical Reasoning



If I put a pencil in a box, then put another pencil in the box, what is in the box?



Two pencils.

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- ≈ 34 days on 1,024 x A100 GPUs.
- \$4.6M costs a single training run.

GPT-3 [Jun. 2020]

World Knowledge



What city is in the northwest corner of Ohio?



Toledo is in the northwest corner of Ohio.

Common Sense



Why don't animals have three legs?



Animals don't have three legs because they would fall over.

Logical Reasoning



If I put a pencil in a box, then put another pencil in the box, what is in the box?



Two pencils.

+ Learn to follow instructions and to comply with answer policies.

- (1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
- (2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
- (3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

GPT-3.5 (InstructGPT) [Jan. 2022]

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- ≈ 34 days on 1,024 x A100 GPUs.
- \$4.6M costs a single training run.

GPT-3 [Jun. 2020]

World Knowledge



What city is in the northwest corner of Ohio?



Toledo is in the northwest corner of Ohio.

Common Sense



Why don't animals have three legs?



Animals don't have three legs because they would fall over.

Logical Reasoning



If I put a pencil in a box, then put another pencil in the box, what is in the box?



Two pencils.

+ Learn to follow instructions and to comply with answer policies.

- (1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
- (2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
- (3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

GPT-3.5 (InstructGPT) [Jan. 2022]

+ Fine-tuning of GPT-3.5 to comply with even stricter guardrails.

ChatGPT [Nov. 2022]

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- ≈ 34 days on 1,024 x A100 GPUs.
- \$4.6M costs a single training run.

GPT-3 [Jun. 2020]

World Knowledge



What city is in the northwest corner of Ohio?



Toledo is in the north-west corner of Ohio.

Common Sense



Why don't animals have three legs?



Animals don't have three legs because they would fall over.

Logical Reasoning



If I put a pencil in a box, then put another pencil in the box, what is in the box?



Two pencils.

+ Learn to follow instructions and to comply with answer policies.

- (1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
- (2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
- (3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

GPT-3.5 (InstructGPT) [Jan. 2022]

+ Fine-tuning of GPT-3.5 to comply with even stricter guardrails.

ChatGPT [Nov. 2022]



Retrieval-Augmented Generation

Retrieval-Augmented Generation

How to satisfy an information need:

Query

↑
↑
←
←

↑
↑
←
←

↑
↑
←
←

↑
↑
←
←

↑
↑
←
←

Query

←
←
←
←

←
←
←
←
←
←
←
←
←
←

←
←
←
←

↑ Convenient, but uncertain veracity

← Authoritative, but tedious to analyze

Retrieval-Augmented Generation

How to satisfy an information need:

Query

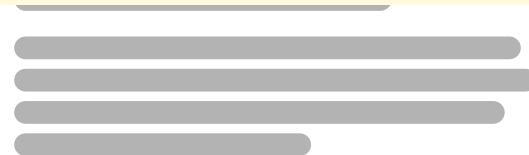
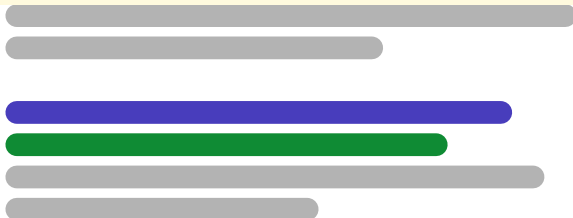


Query



The dilemma of the direct answer: [Potthast et al., 2020]

A user's choice between convenience and diligence when using an information retrieval system.

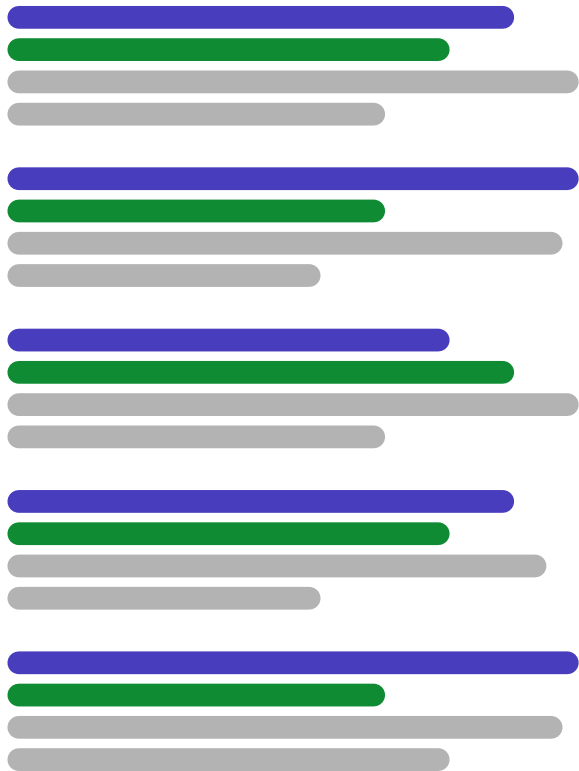


- ↑ Convenient, but uncertain veracity
- ← Authoritative, but tedious to analyze

Retrieval-Augmented Generation

How to satisfy an information need:

Query



Query

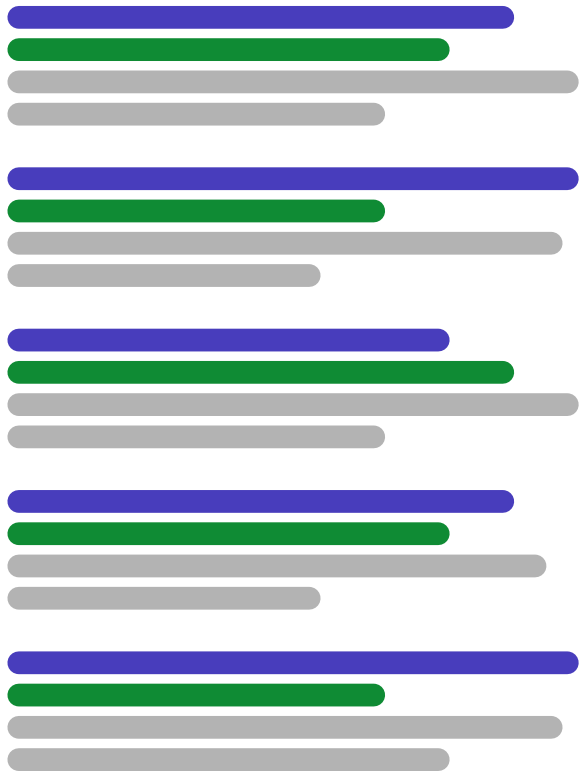


- ↑ Convenient, but uncertain veracity
- ← Authoritative, but tedious to analyze

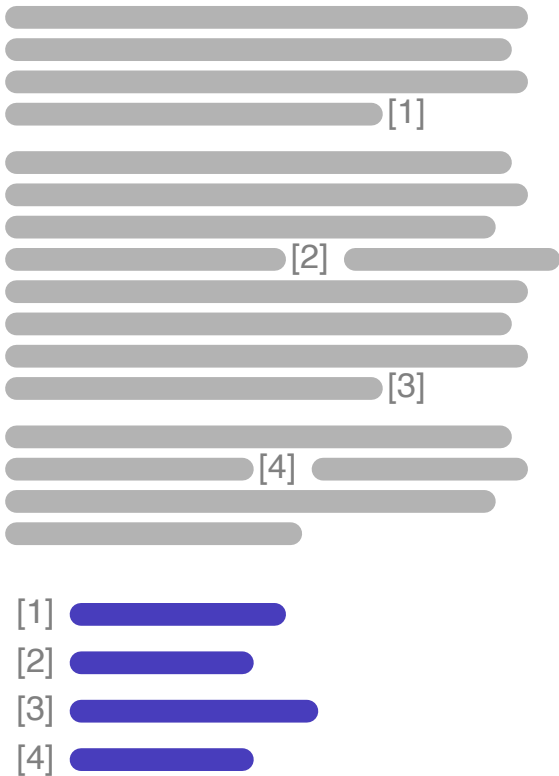
Retrieval-Augmented Generation

How to satisfy an information need:

Query



Query



Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:

Queries

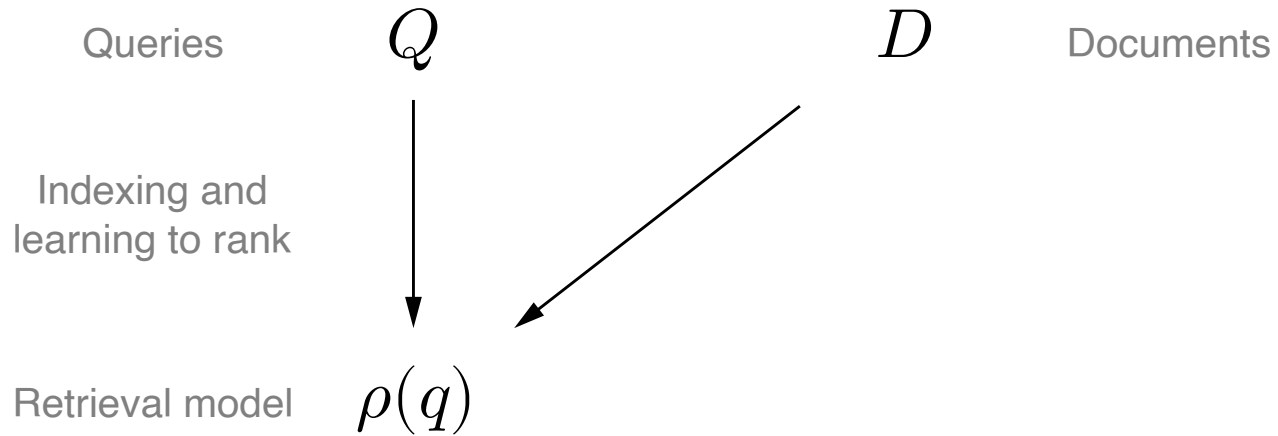
Q

D

Documents

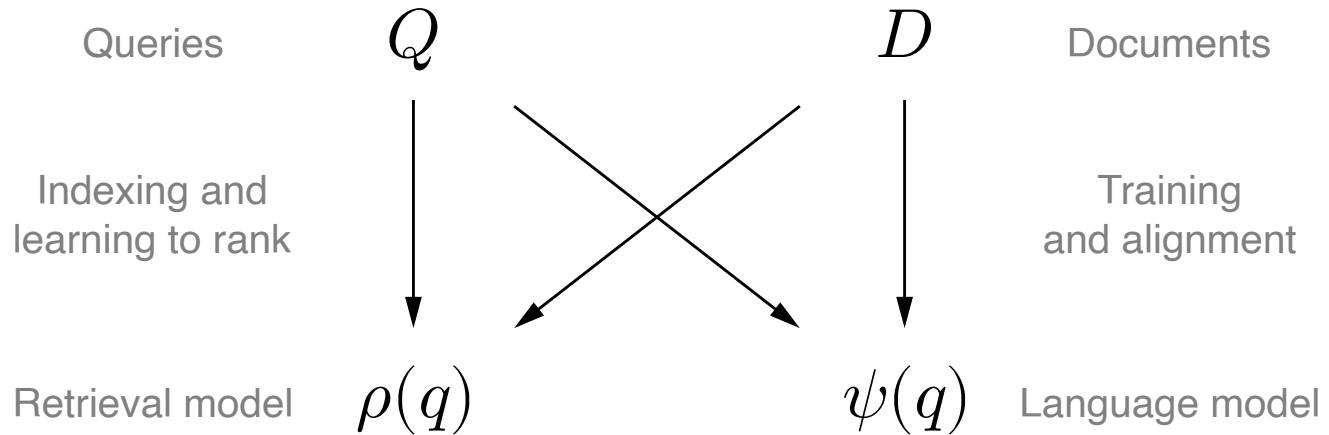
Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:



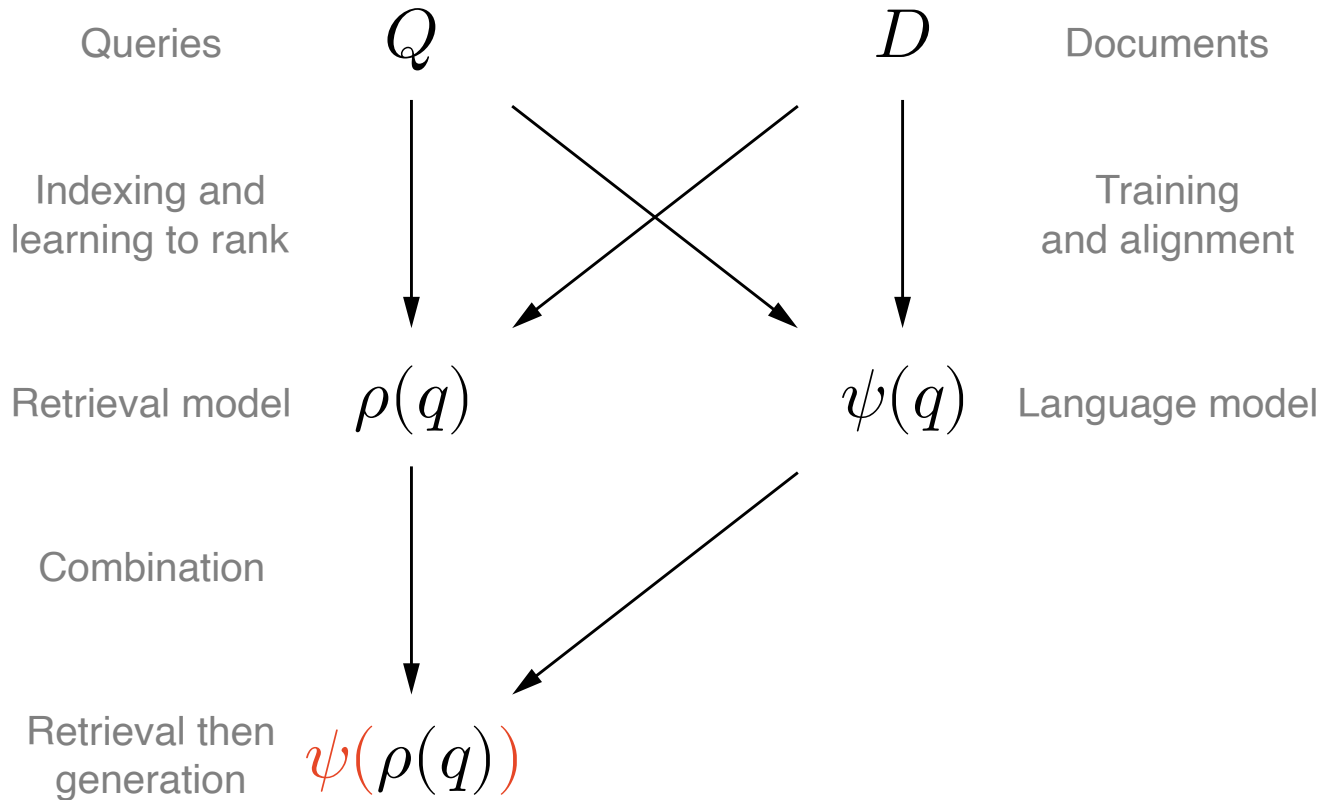
Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:



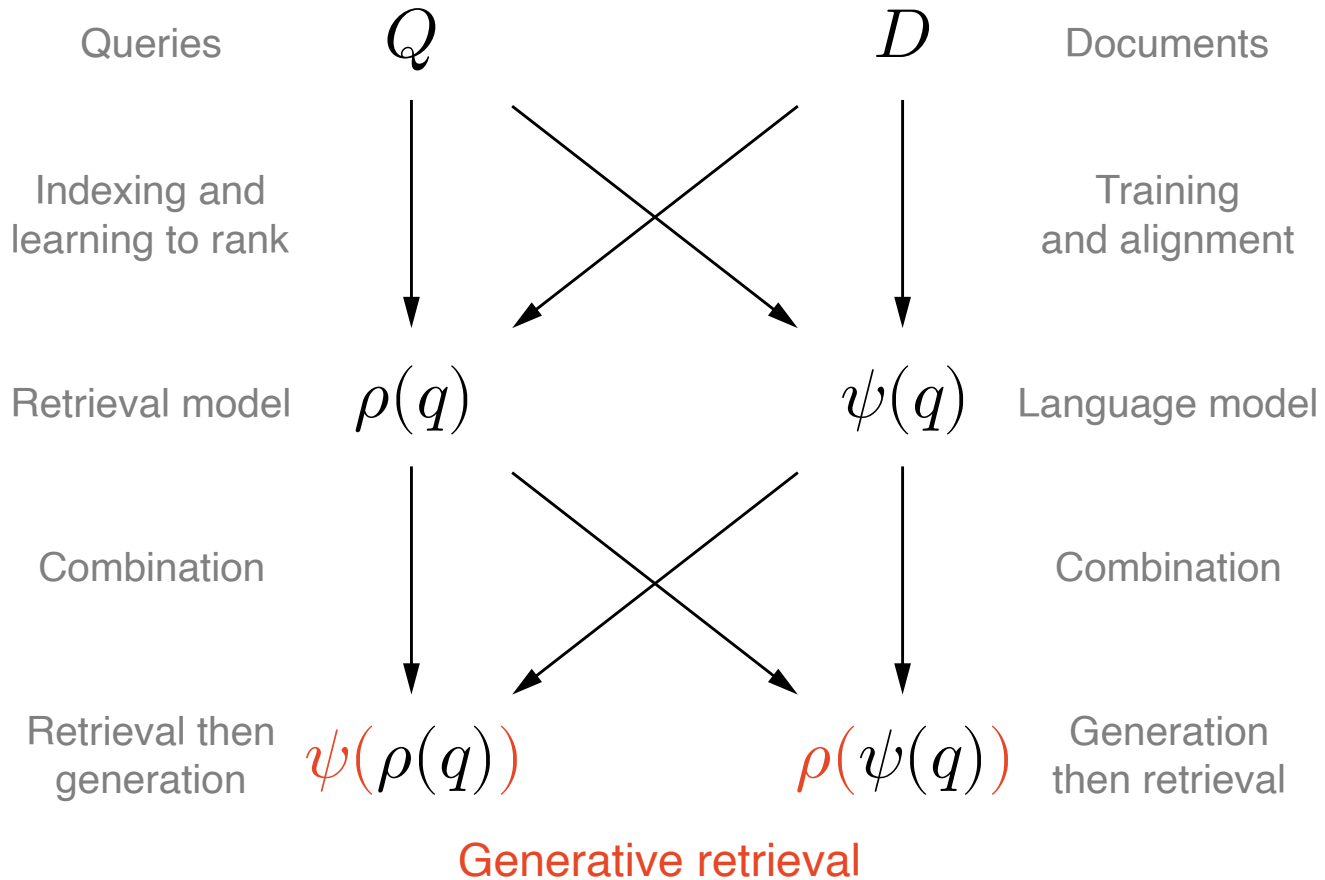
Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:



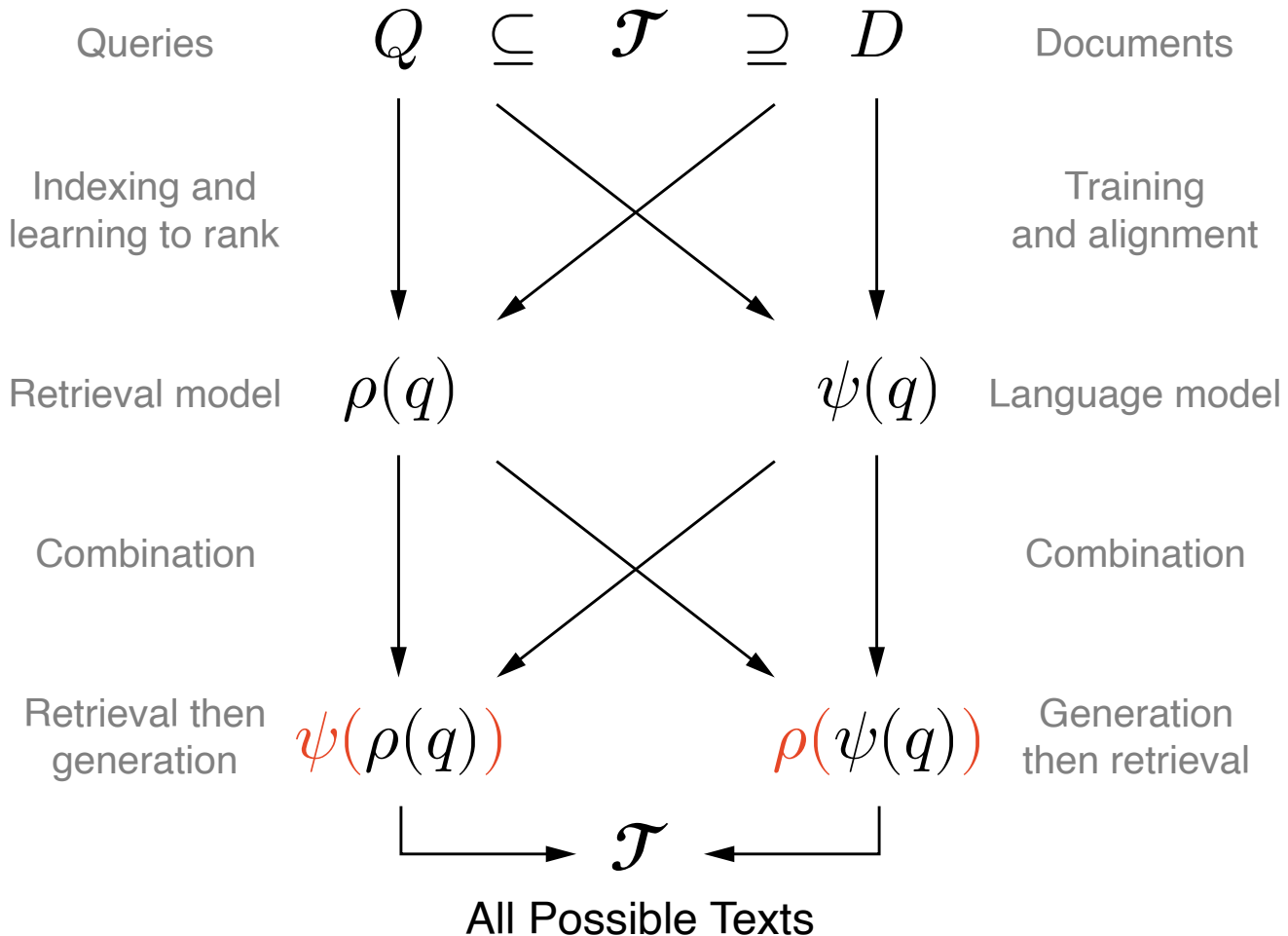
Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:



Retrieval-Augmented Generation

Prompt-level RAG combines existing systems:





The Infinite Index



The Infinite Index

Prompt-level RAG combines existing systems:

Queries

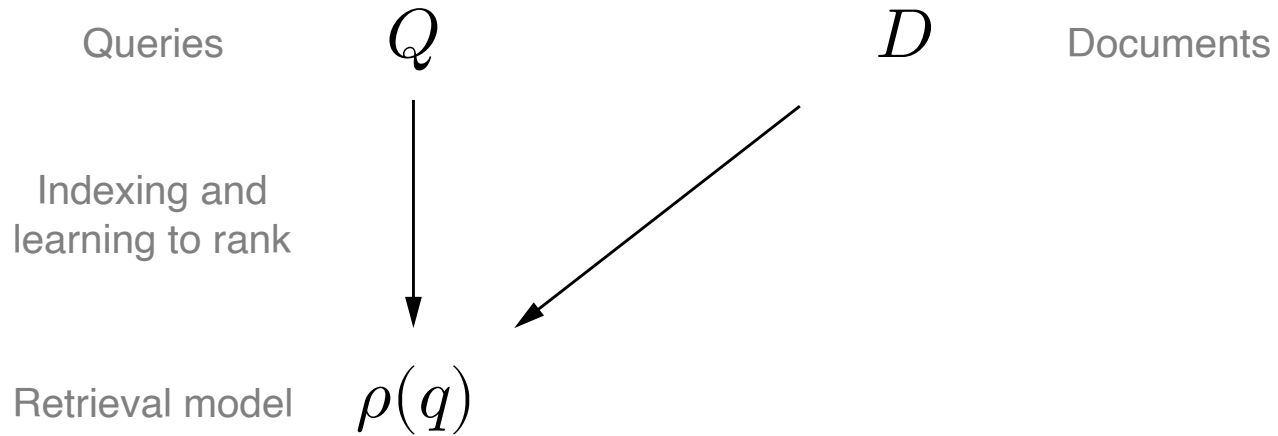
Q

D

Documents

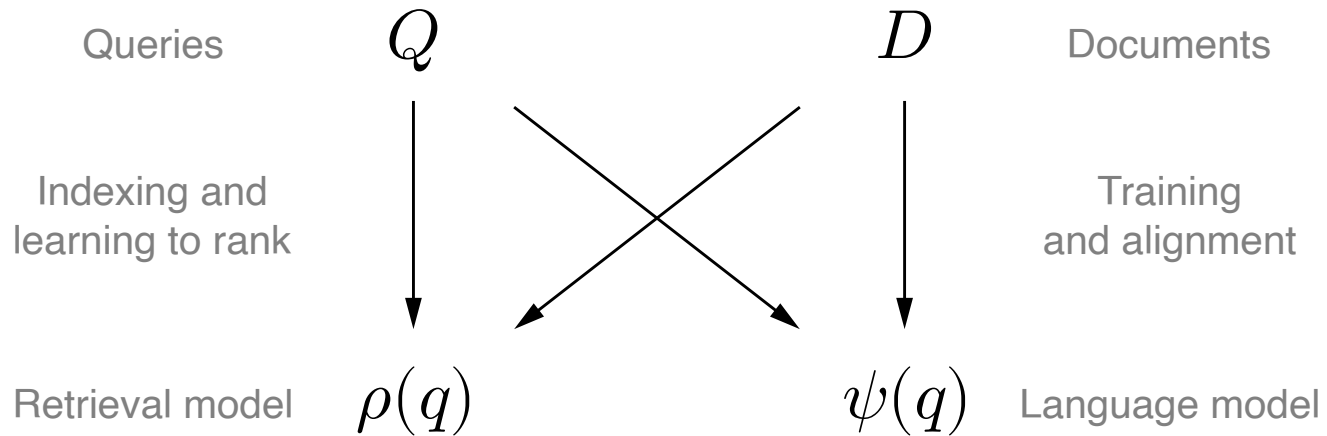
The Infinite Index

Prompt-level RAG combines existing systems:



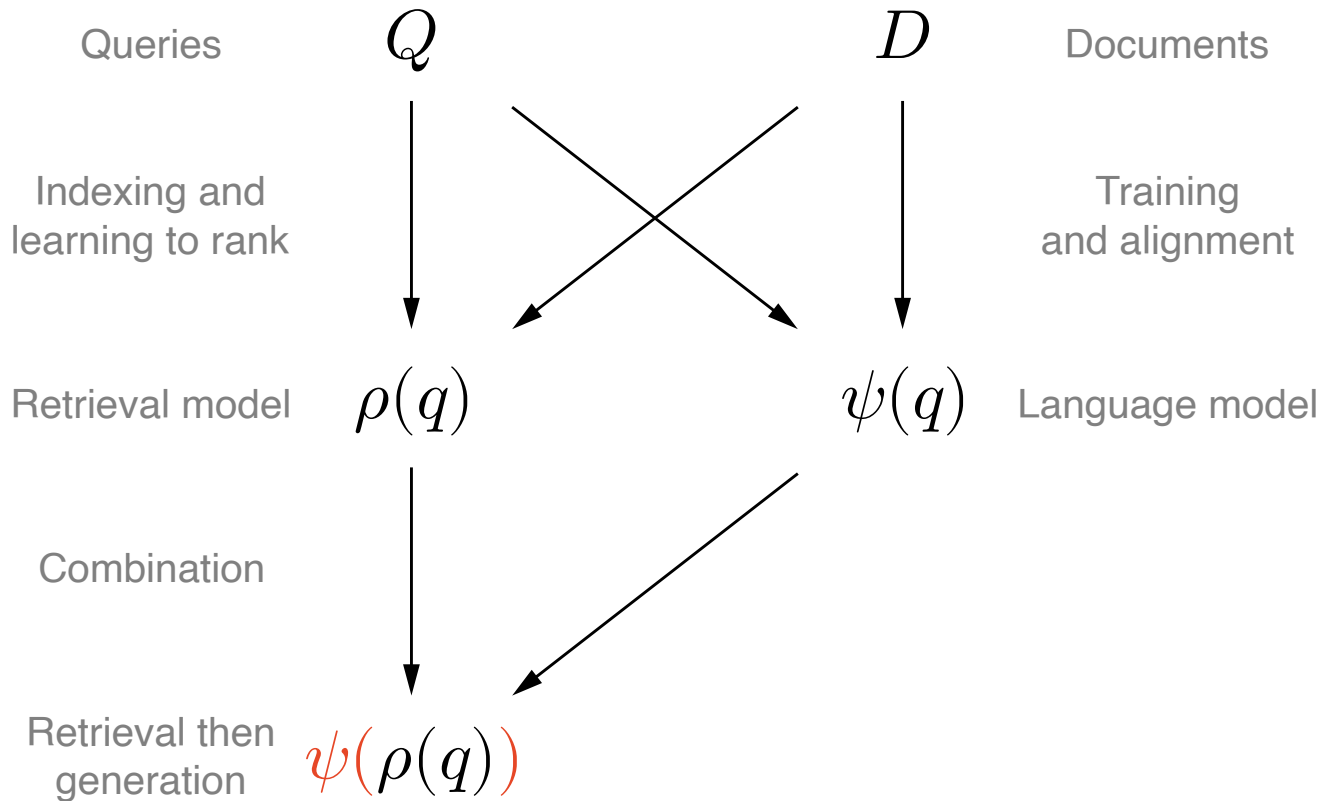
The Infinite Index

Prompt-level RAG combines existing systems:



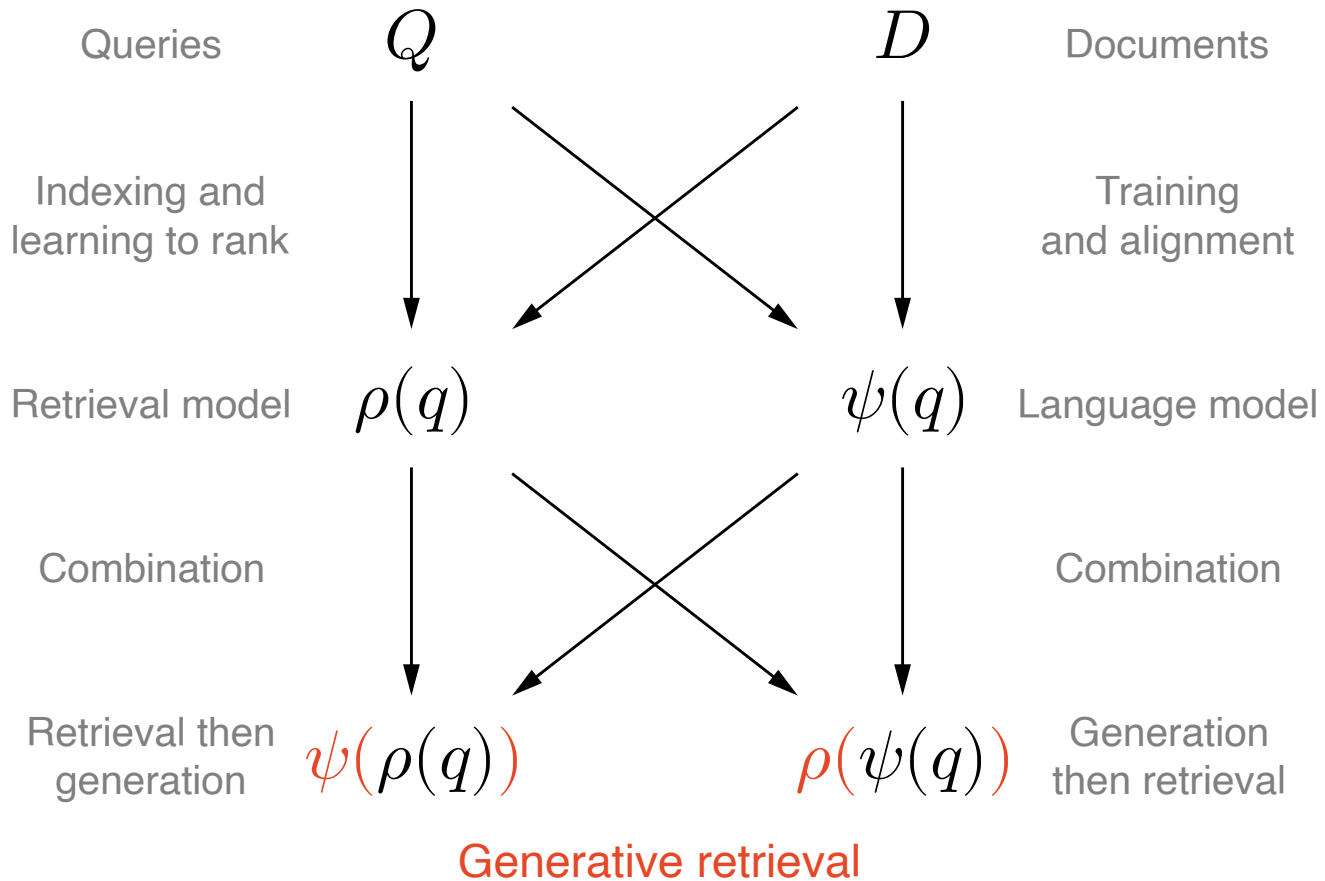
The Infinite Index

Prompt-level RAG combines existing systems:



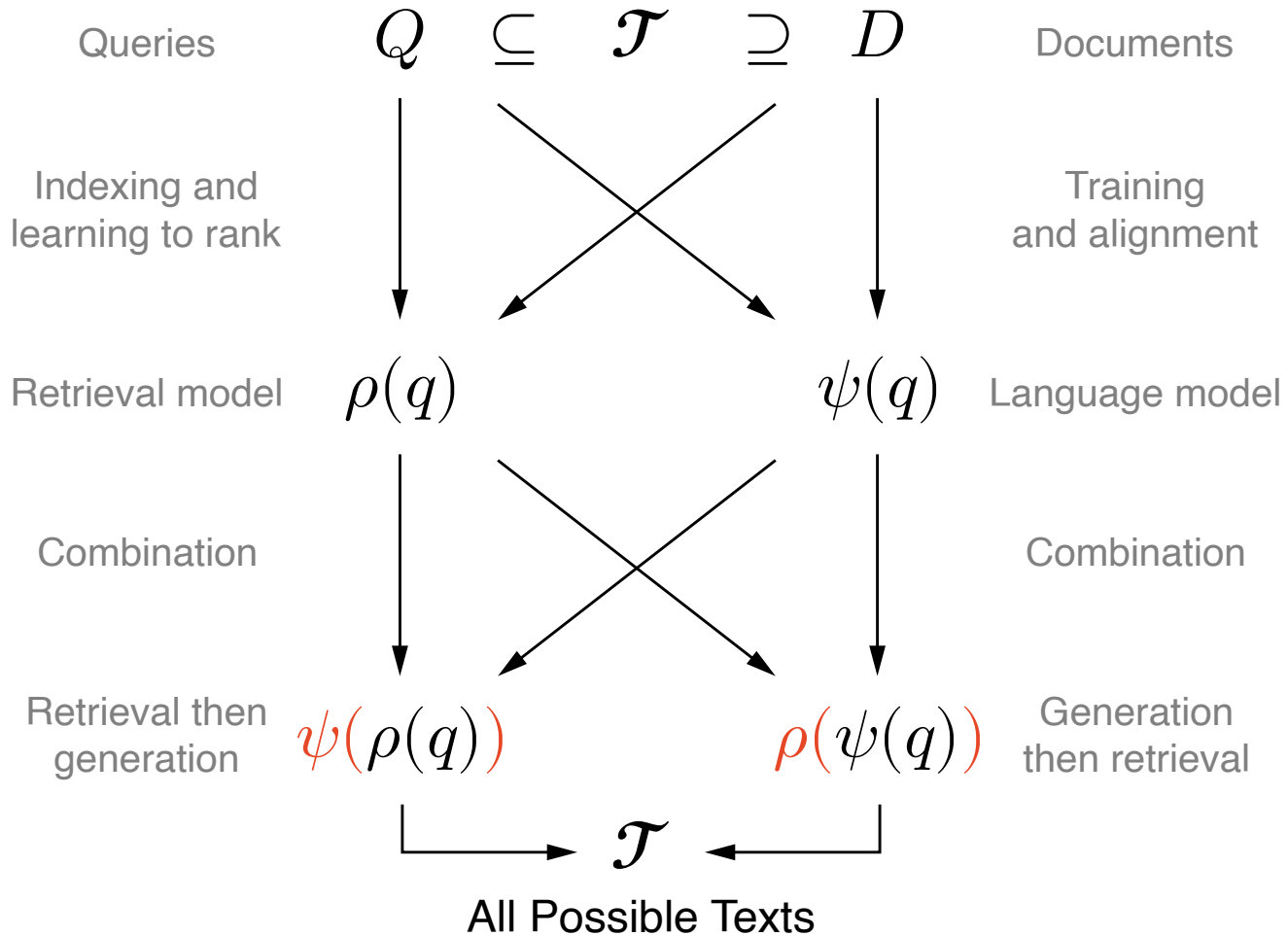
The Infinite Index

Prompt-level RAG combines existing systems:



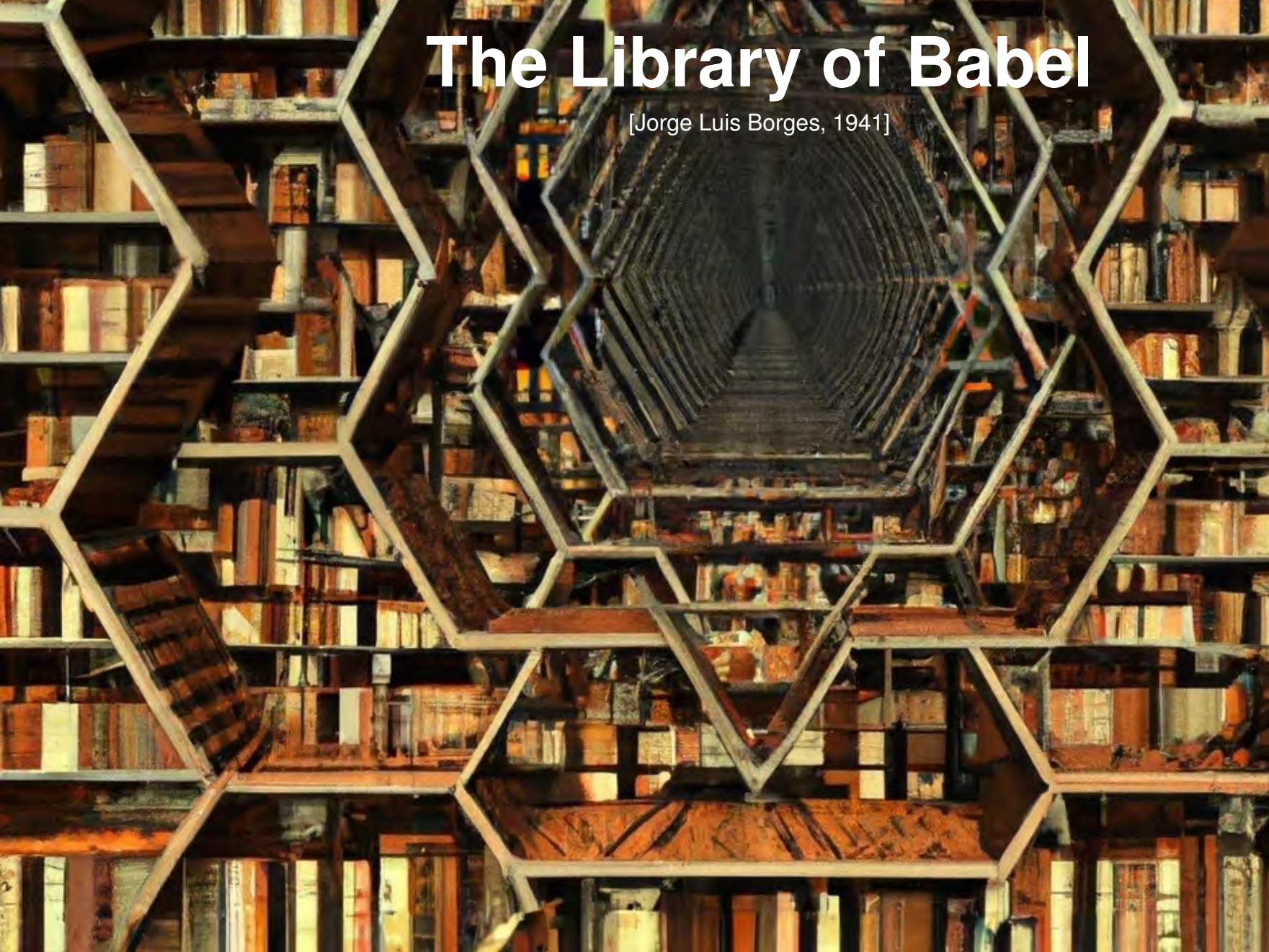
The Infinite Index

Prompt-level RAG combines existing systems:



The Library of Babel

[Jorge Luis Borges, 1941]



The Library of Babel

[Jorge Luis Borges, 1941]

- ❑ Infinite library with all possible texts from all letter combinations
- ❑ The people in it spend their lives searching for meaningful text fragments

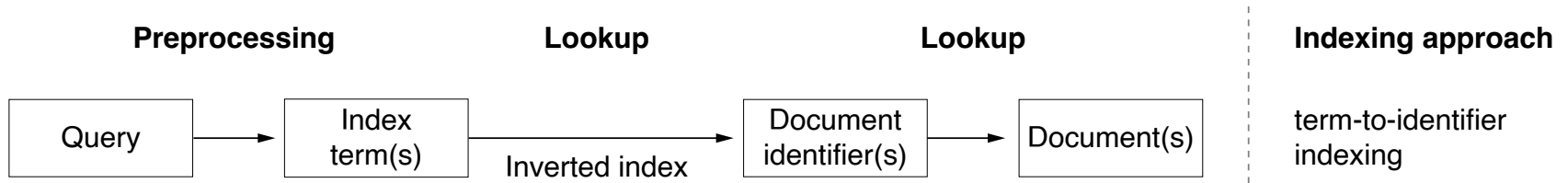
The Library of Babel

[Jorge Luis Borges, 1941]

- ❑ Infinite library with all possible texts from all letter combinations
- ❑ The people in it spend their lives searching for meaningful text fragments
- ❑ When prompted, a language model “retrieves” a relevant text [[CHIIR 2023](#)]:

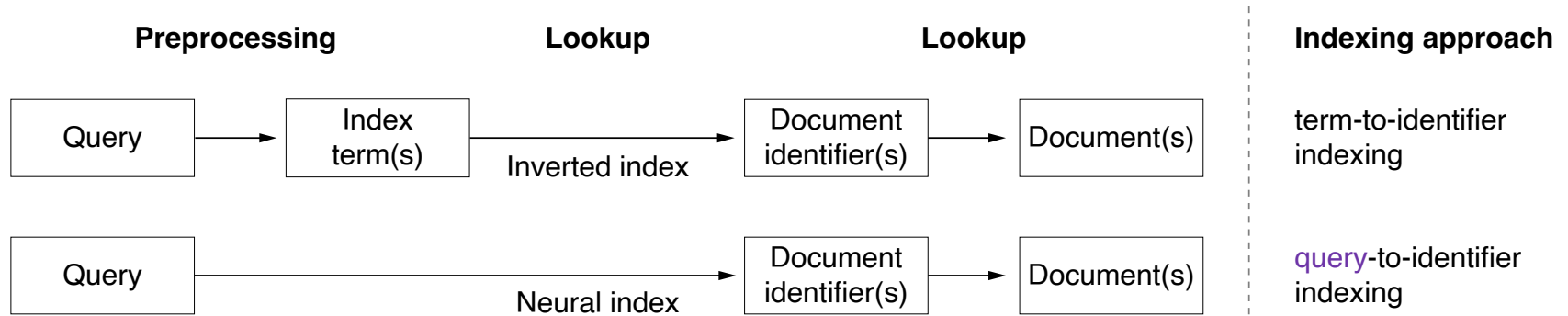
A language model is an infinite index

The Infinite Index



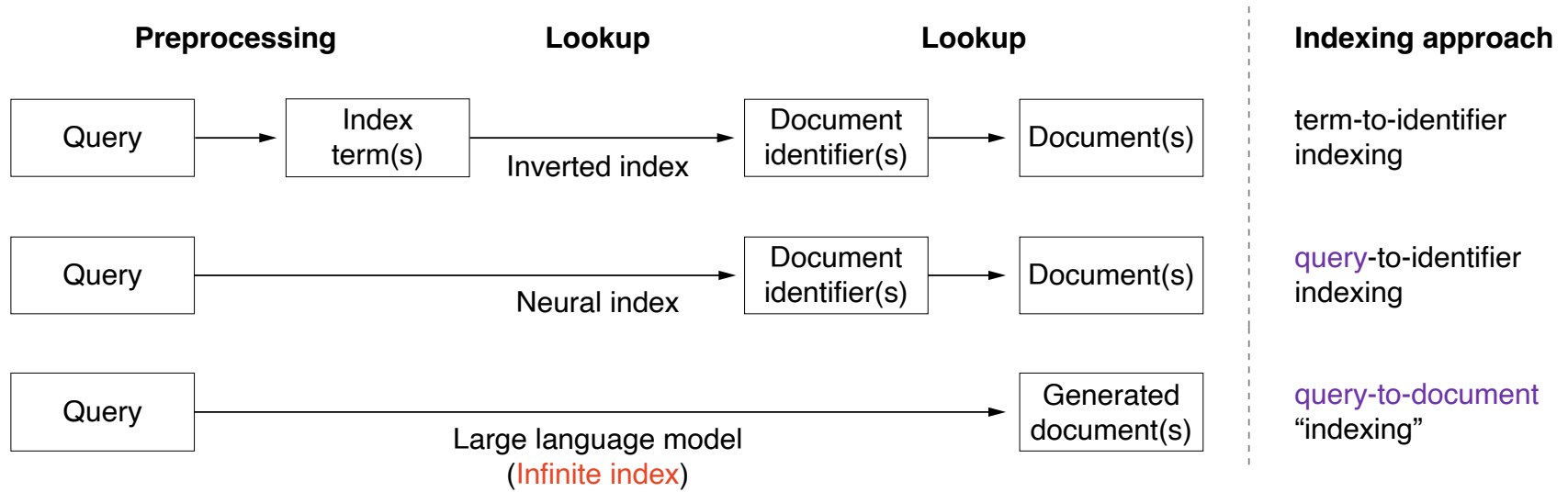
[[Search Engine Architecture](#)]

The Infinite Index



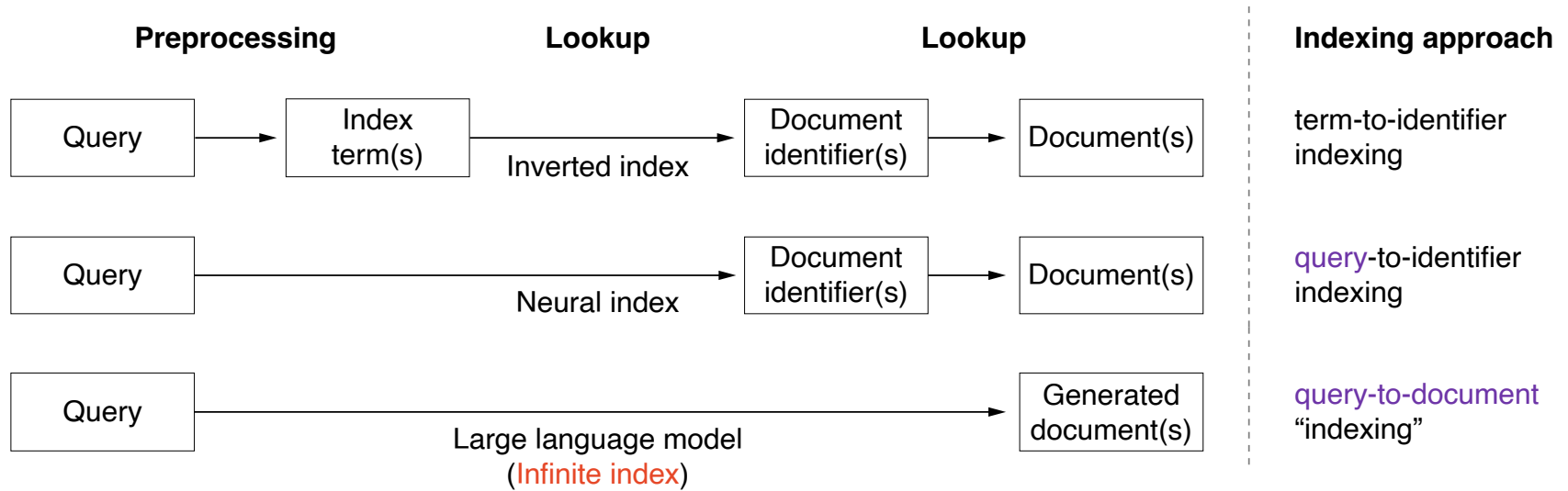
[[Search Engine Architecture](#)]

The Infinite Index



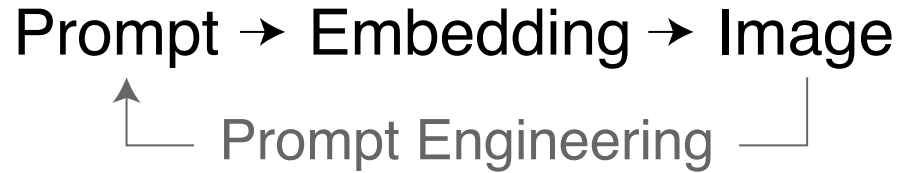
[[Search Engine Architecture](#)]

The Infinite Index

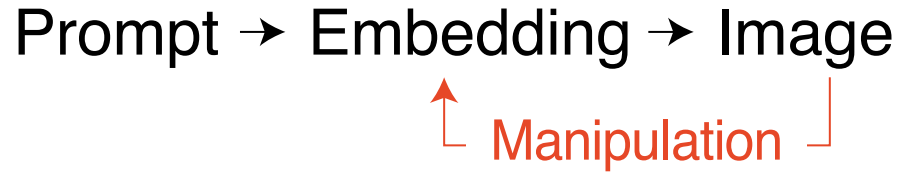


[[Search Engine Architecture](#)]

What retrieval model works on an infinite index?



- ❑ Text-to-image models require users to declare their desired output
- ❑ For a given seed, a prompt designates precisely one image
- ❑ The first image generated is usually not satisfactory
- ❑ Prompt engineering (trial and error) is required to reach the desired result



- ❑ Text-to-image models require users to declare their desired output
- ❑ For a given seed, a prompt designates precisely one image
- ❑ The first image generated is usually not satisfactory
- ❑ Prompt engineering (trial and error) is required to reach the desired result
- ❑ We develop a retrieval model that allows users to “surf” embedding space
- ❑ Prompt embedding manipulation allows targeted retrieval from image space

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

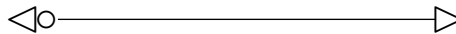
Big treehouse in rain forest with two floors, green roof, and spiral staircase.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



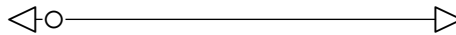
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

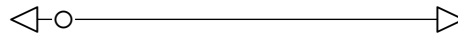
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



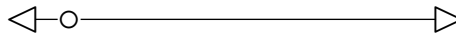
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

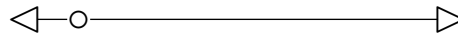
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

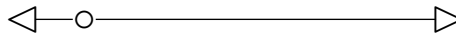
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

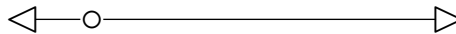
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

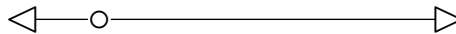
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

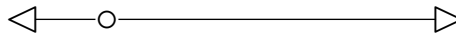
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

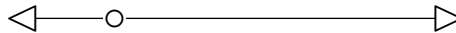
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

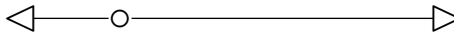
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

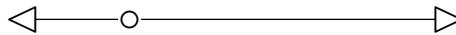
Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

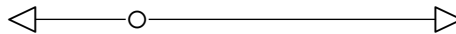
Big treehouse in rain forest with two floors, green roof, and spiral staircase.



The Infinite Index

Query 1:

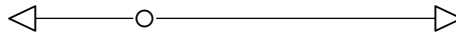
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

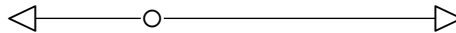
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



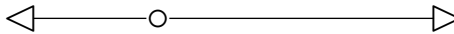
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

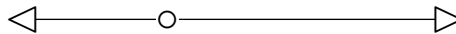
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

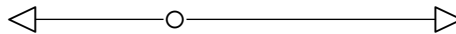
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

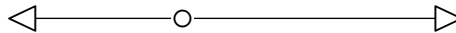
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

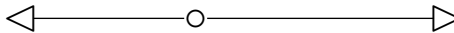
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

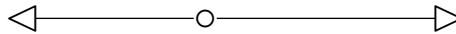
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

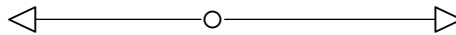
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

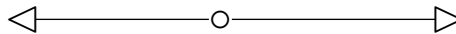
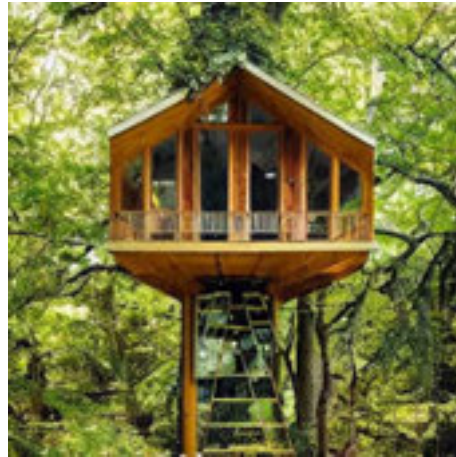
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

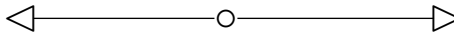
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

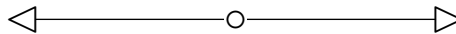
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



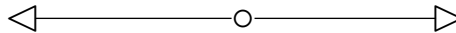
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

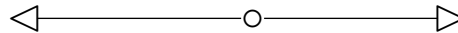
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

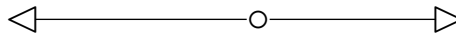
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

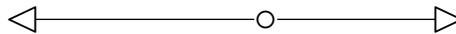
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

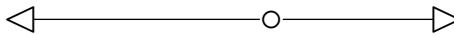
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

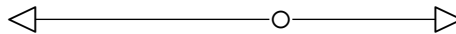
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



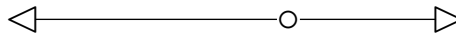
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



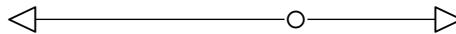
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



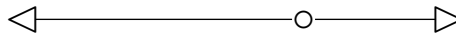
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



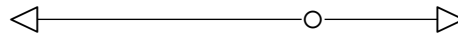
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



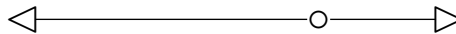
Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

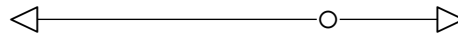
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

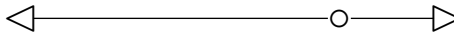
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

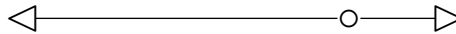
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

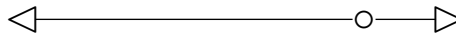
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

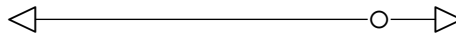
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

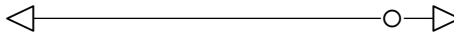
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

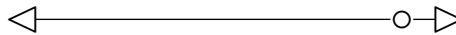
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

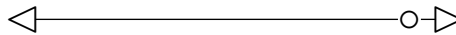
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

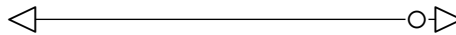
Golden treehouse in lush forest with big glass window and intricate woodwork.



The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.



Query 2:

Big treehouse in rain forest with two floors, green roof, and spiral staircase.

The Infinite Index

Query 1:

Golden treehouse in lush forest with big glass window and intricate woodwork.

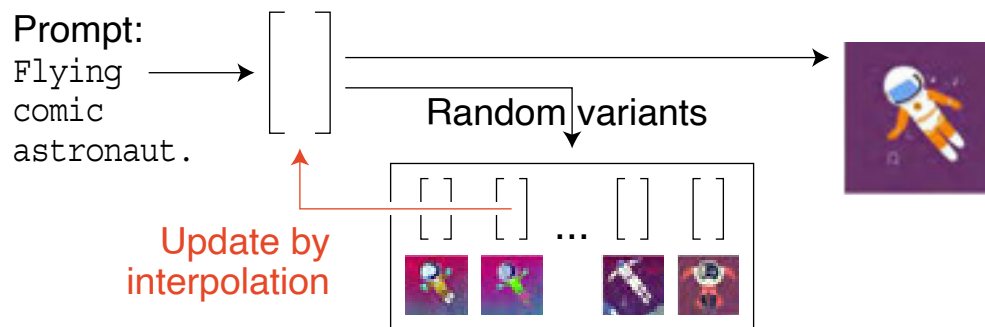


The Infinite Index [IJCAI 2024]



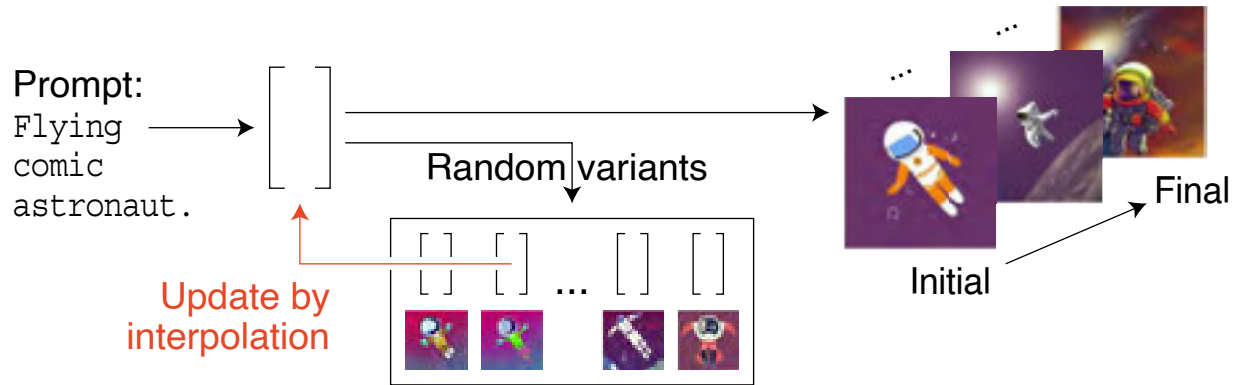
- A user enters a vague prompt / query and is dissatisfied with the first result.

The Infinite Index [IJCAI 2024]



- ❑ A user enters a vague prompt / query and is dissatisfied with the first result.
- ❑ Our approach offers a selection of alternatives
- ❑ Alternatives are sampled in the “vicinity” of the prompt embedding
- ❑ A user selects “interesting” alternative(s)
- ❑ The system generates new results by interpolation

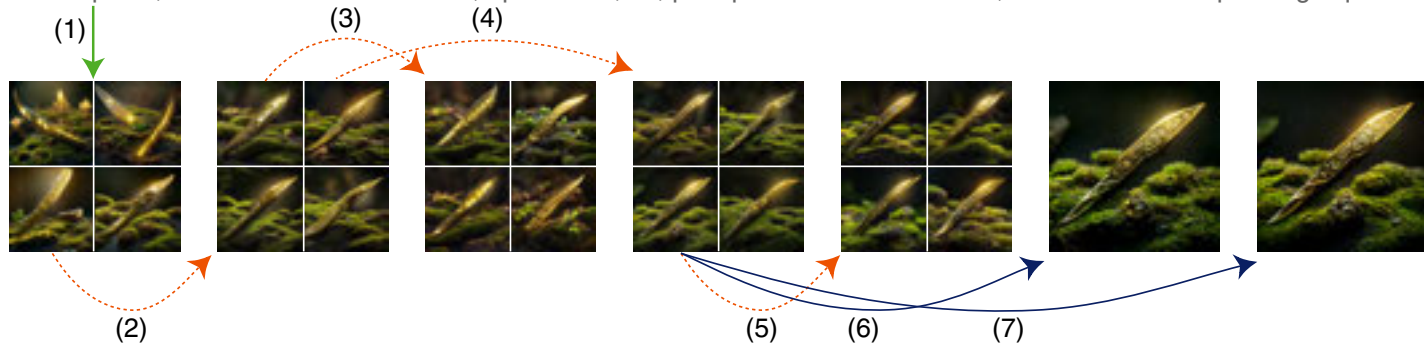
The Infinite Index [IJCAI 2024]



- ❑ A user enters a vague prompt / query and is dissatisfied with the first result.
- ❑ Our approach offers a selection of alternatives
- ❑ Alternatives are sampled in the “vicinity” of the prompt embedding
- ❑ A user selects “interesting” alternative(s)
- ❑ The system generates new results by interpolation
- ❑ In each iteration, the user “surfs” the embedding space
- ❑ Selection of liked alternatives gives a sense of direction

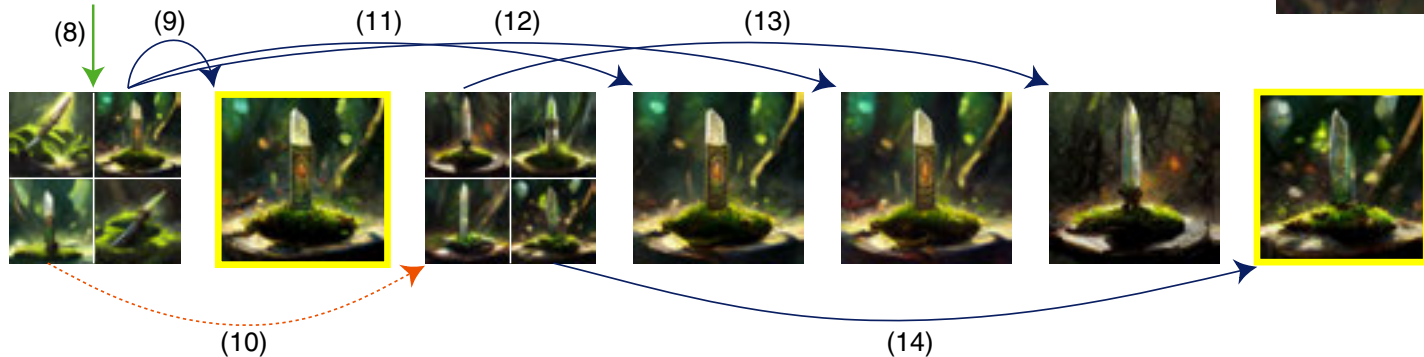
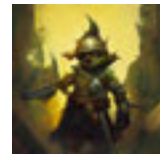
The Infinite Index [IJCAI 2024]

Initial prompt: an ancient golden dagger lying on moss, illuminated by godrays, close up, digital painting, matte painting, midjourney, concept art, detailed art, sci-fi cinematic painting, magic the Gathering, volumetric light, masterpiece, volumetric realistic render, epic scene, 8k, post-production detailed art, sci-fi cinematic painting --q 2



Reformulated prompt: a medieval dagger lying on moss, lit by god rays, art by Adrian Smith + Paul bonner, magic gathering style, warcraft, blizzard style, hearthstone, fantasy concept art, medieval, masterpiece, mystical, witchcraft

+



The Infinite Index [IJCAI 2024] [demo]

1. Initialization

2. Image Selection

3. History



Interpolation Value

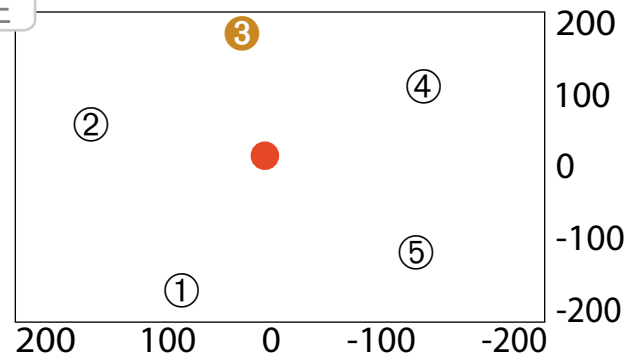
0.3

Generate

Current



TSNE



Take-away messages

- ❑ Retrieval models are made useful based on human feedback
- ❑ Language models are made useful based on human feedback, too
- ❑ Users wish to make vague requests
- ❑ Prompt engineering resembles declarative retrieval
- ❑ Retrieval-augmented generation addresses hallucination
- ❑ New retrieval technologies are required to support image generation
- ❑ Prompt embedding manipulation may relief

Evaluation is remains a challenge

Netspeak — Mozilla Firefox

Netspeak x +

https://netspeak.org/#?see+...works

Netspeak

One word leads to another.

English German

see ... works i x ↺

how to ? this
 see ... works
 it's [great well]
 and knows #much
 { more show me }
 m...d ? g?p

The ? finds one word.
 The ... finds many words.
 The [] compare options.
 The # finds similar words.
 The { } check the order.
 The space is important.

see how it works	150,000	20%
see if it works	100,000	14%
see works	57,000	7.5%
see how this works	55,000	7.3%
see what works	51,000	6.7%
see the works	51,000	6.7%
see if that works	28,000	3.7%
see your good works	28,000	3.7%
see how that works	25,000	3.3%
see how technorati works	23,000	3.0%
see if this works	17,000	2.3%
see more works	17,000	2.2%
see if it really works	15,000	2.1%
see his works	12,000	1.7%
see how well it works	11,000	1.5%
see other works	8,900	1.2%

netspeak - Mozilla Firefox

netspeak

x

+

←

→

🔒

https://netspeak.org/Bq=i+lovemy?

133K

...

☆

🔄

⬇️

📄

☰

Netspeak One word leads to another.

English

German

I love my ?

i x ↺

how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }
m...d ? g?p

The ? finds one word.
The ... finds many words.
The [] compare options.
The # finds similar words.
The { } check the order.
The space is important.

i love my job	72,000	10%
i love my country	44,000	6.2%
i love my family	41,000	5.9%
i love my wife	38,000	5.4%
i love my new	34,000	4.9%
i love my friends	33,000	4.7%
i love my pet	27,000	3.8%
i love my dog	26,000	3.7%
i love my husband	26,000	3.7%
i love my life	24,000	3.4%
i love my baby	24,000	3.4%
i love my soldier	22,000	3.1%
i love my cat	21,000	2.9%
i love my computer	18,000	2.6%
i love my work	16,000	2.4%
i love my mom	16,000	2.3%

2022 WHAT'S IN MY AI? – ALT VIEW



Google Patents.....	0.48%
The New York Times.....	0.06%
Los Angeles Times.....	0.06%
The Guardian.....	0.06%
Public Library of Science.....	0.06%
Forbes.....	0.05%
Huffington Post.....	0.05%
Patents.com.....	0.05%
Scribd.....	0.04%
Other.....	99.09%

Common Crawl

Google.....	3.4%
Archive.....	1.3%
Blogspot.....	1.0%
GitHub.....	0.9%
The New York Times.....	0.7%
Wordpress.....	0.7%
Washington Post.....	0.7%
Wikia.....	0.7%
BBC.....	0.7%
Other.....	89.9%

Reddit links

Biography.....	27.8%
Geography.....	17.7%
Culture and Arts.....	15.8%
History.....	9.9%
Biology, Health, Medicine.....	7.8%
Sports.....	6.5%
Business.....	4.8%
Other society.....	4.4%
Science & Math.....	3.5%
Education.....	1.8%

English Wikipedia

Romance.....	26.1%
Fantasy.....	13.6%
Science Fiction.....	7.5%
New Adult.....	6.9%
Young Adult.....	6.8%
Thriller.....	5.9%
Mystery.....	5.6%
Vampires.....	5.4%
Horror.....	4.1%
Other.....	18.0%

BookCorpus (GPT-1 only)

