Conversational Information Seeking: Theory and Application

#TheCISTutorial

Part 3: Large Language Models and Retrieval Augmentation

DISTRIBUTED REPRESENTATIONS [1]

Geoffrey E. Hinton
Computer Science Department
Carnegie-Mellon University
Pittsburgh PA 15213

Distributed representation of words is also called **word embedding**.

October 1984

Abstract

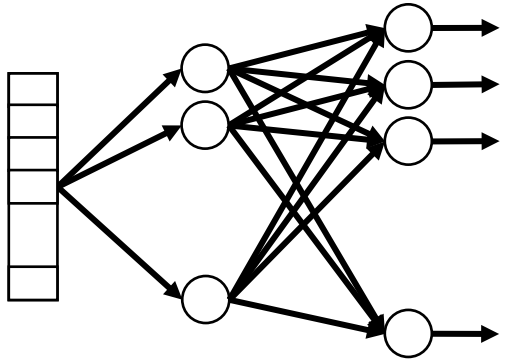
Given a network of simple computing elements and some entities to be represented, the most straightforward scheme is to use one computing element for each entity. This is called a *local* representation. It is easy to understand and easy to implement because the structure of the physical network mirrors the structure of the knowledge it contains. This report describes a different type of representation that is less familiar and harder to think about than local representations. Each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities. The strength of this more complicated kind of representation does not lie in its notational convenience or its ease of implementation in a conventional computer, but rather in the efficiency with which it makes use of the processing abilities of networks of simple, neuron-like computing elements.

Every representational scheme has its good and bad points. Distributed representations are no exception. Some desirable properties like content-addressable memory and automatic generalization arise very naturally

Geoffery E. Hinton. "Distributed Representations" Technical Report, CMU-CS-84-157, 1984.

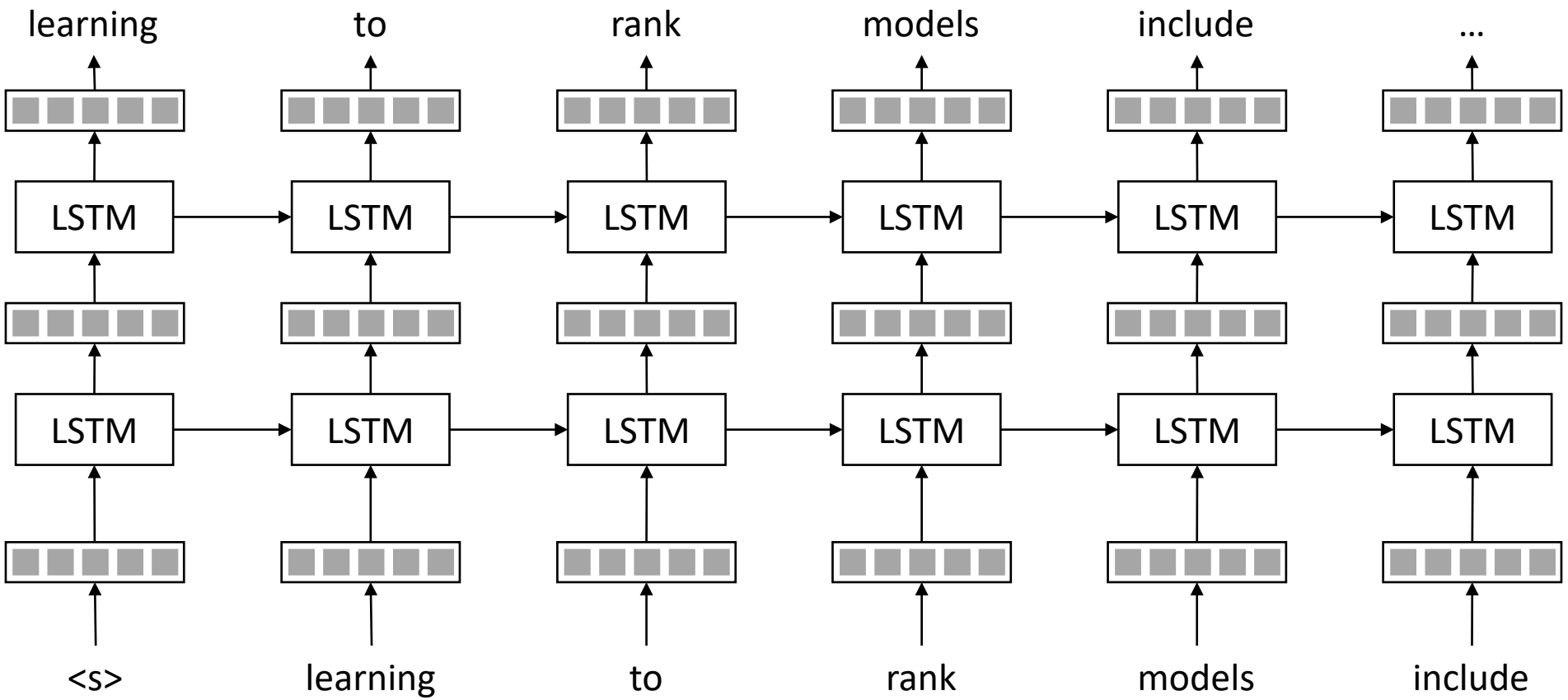
Contextual Embedding

- Word embedding methods, like word2vec and GloVe, are based on the **bag-of-words assumption**.



- They learn a representation for each word. But what would be the representation for ambiguous words? Words have different meanings in different contexts.
- Contextual embeddings address this problem by representing each word conditioned on its context (previous words or adjacent words).

Recap: Language Models using RNNs

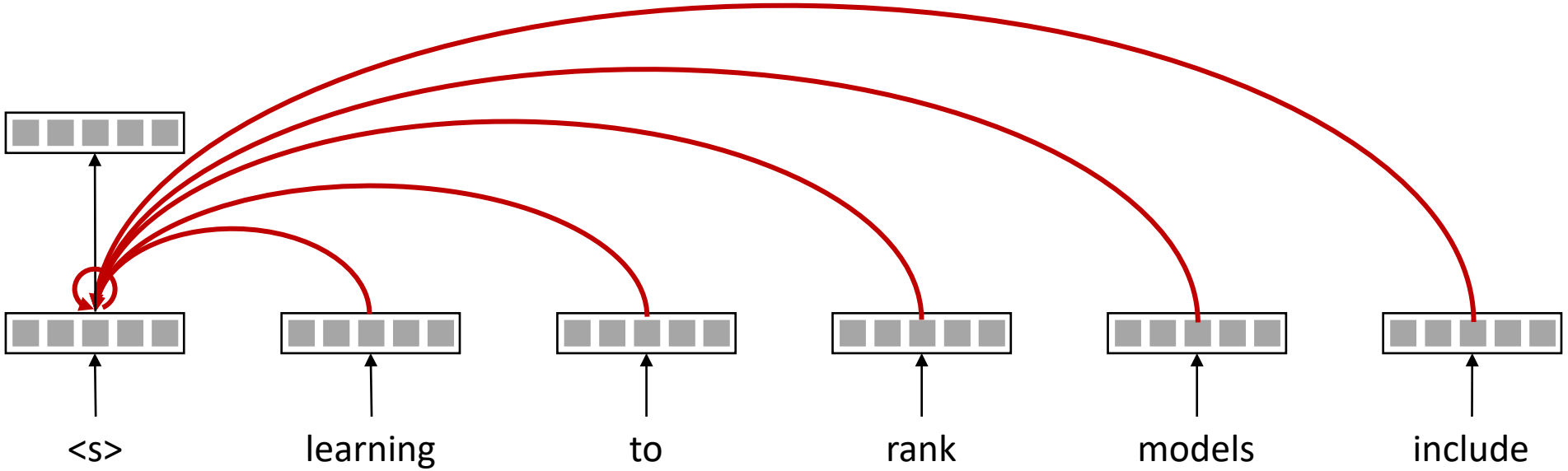


Problem: difficult to parallelize…

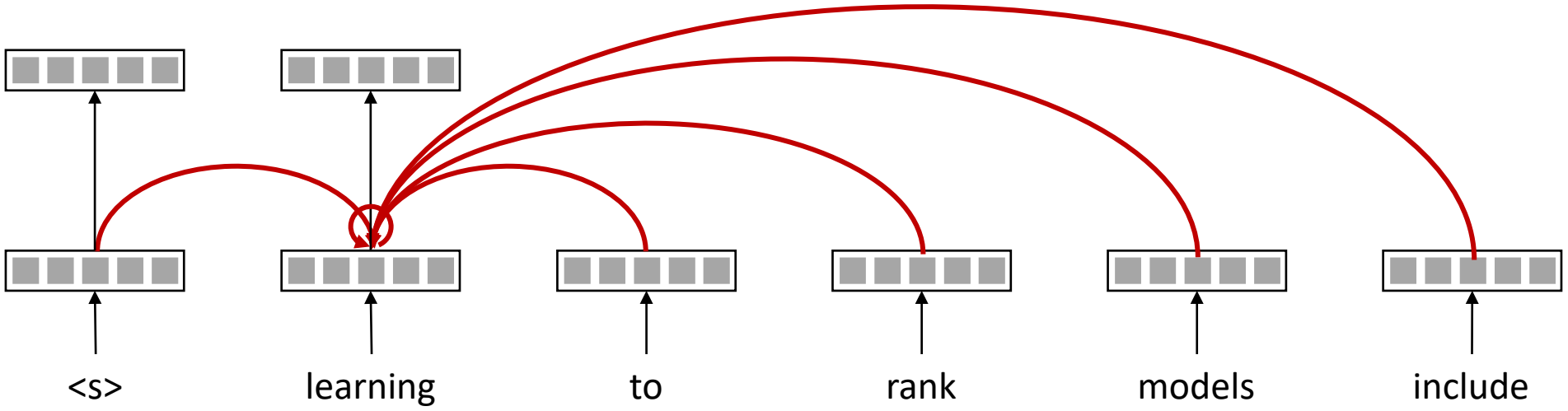
Transformer

- Transformer networks have been introduced in 2017 as an alternative to RNNs for text representation and generation tasks.
- Transformer uses an easy-to-parallelize operation called self-attention.
- Transformer networks are the current state-of-the-art in NLP and IR tasks.

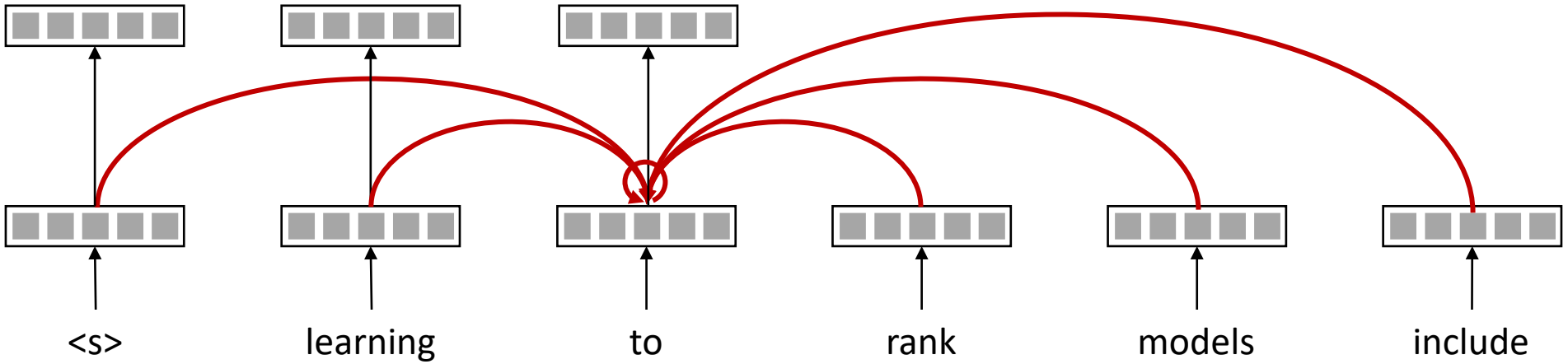
Self-Attention



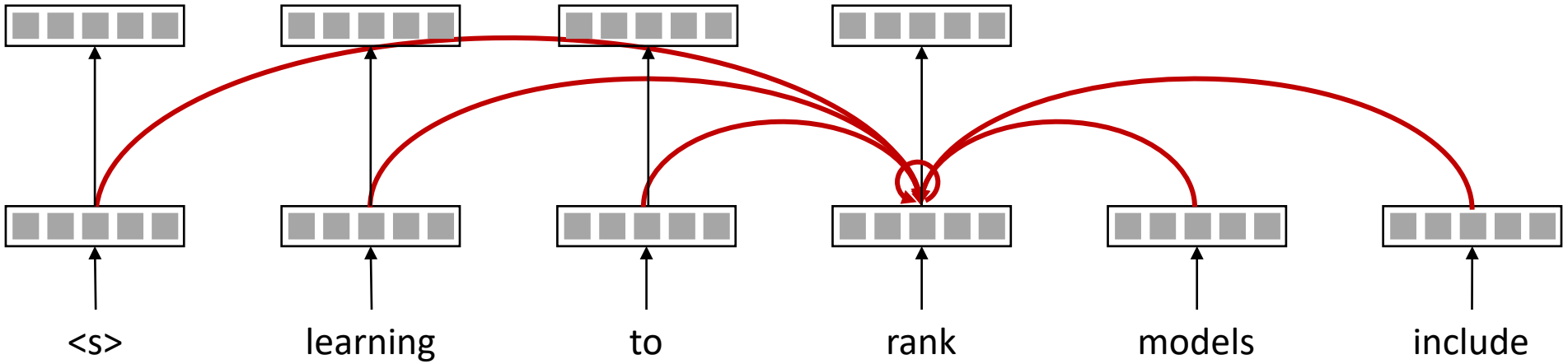
Self-Attention



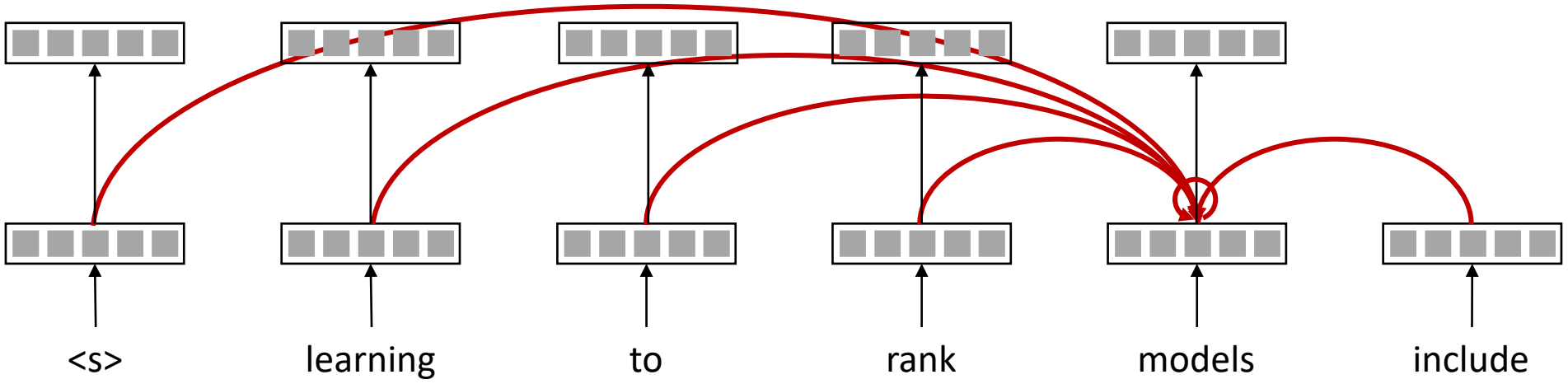
Self-Attention



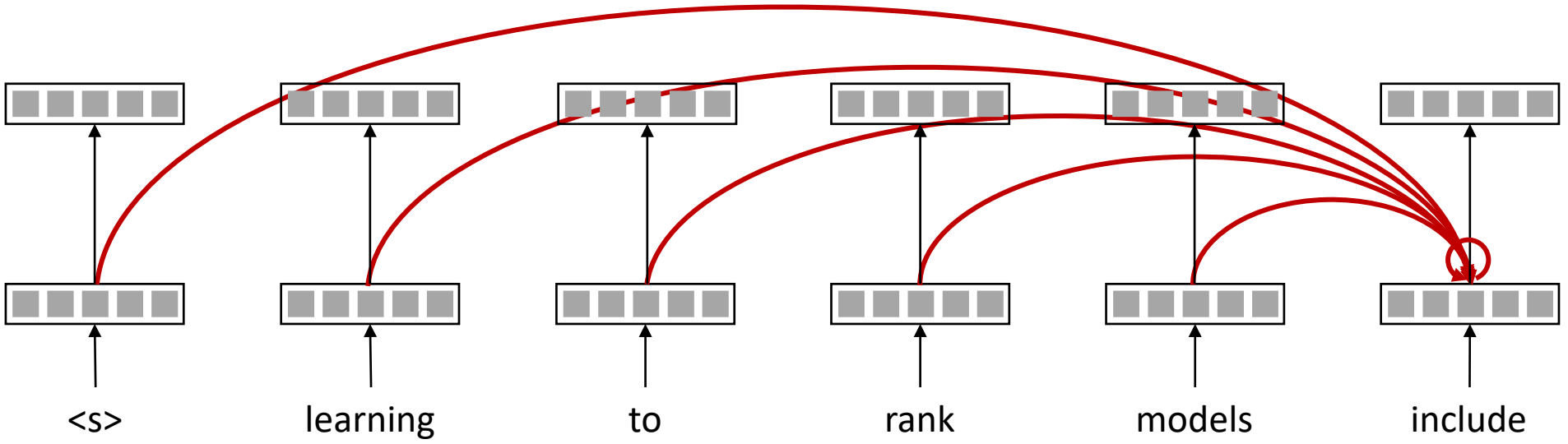
Self-Attention



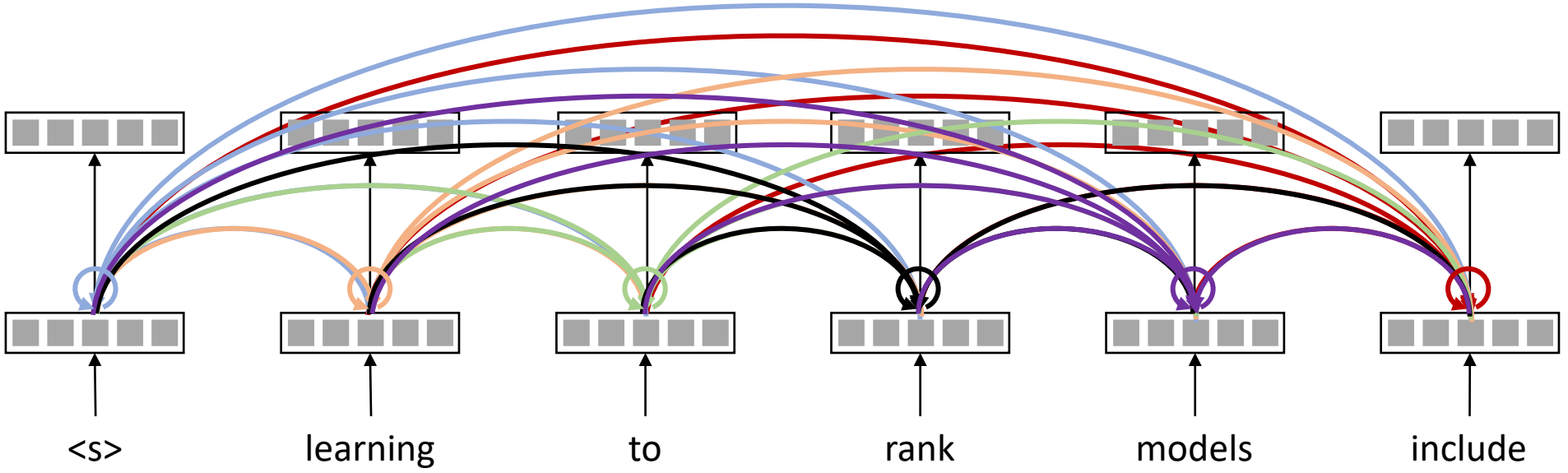
Self-Attention



Self-Attention



Self-Attention

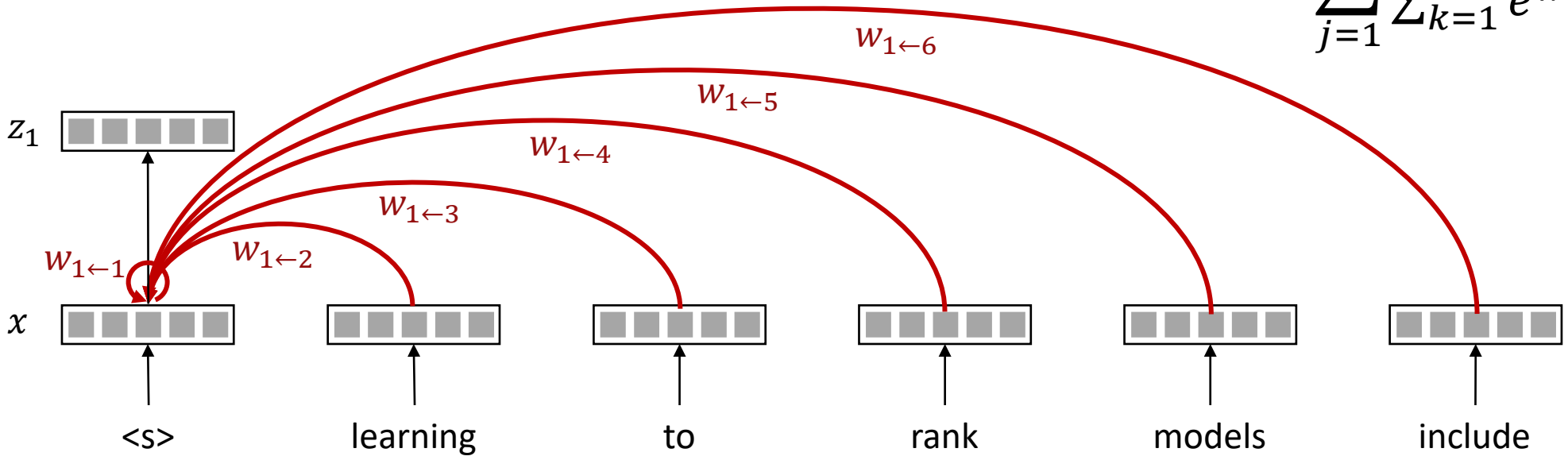


All self-attentions happen in parallel…

Updating Representations Through Self-Attention

$$z_i = \sum_{j=1}^{n} p_{i \leftarrow j} v_j$$

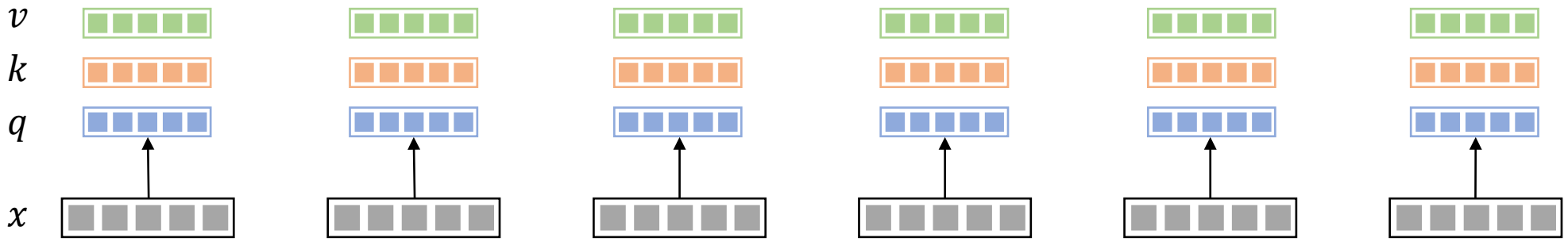
$$= \sum_{j=1}^{n} \frac{e^{w_{i \leftarrow j}}}{\sum_{k=1}^{n} e^{w_{i \leftarrow k}}} v_j$$



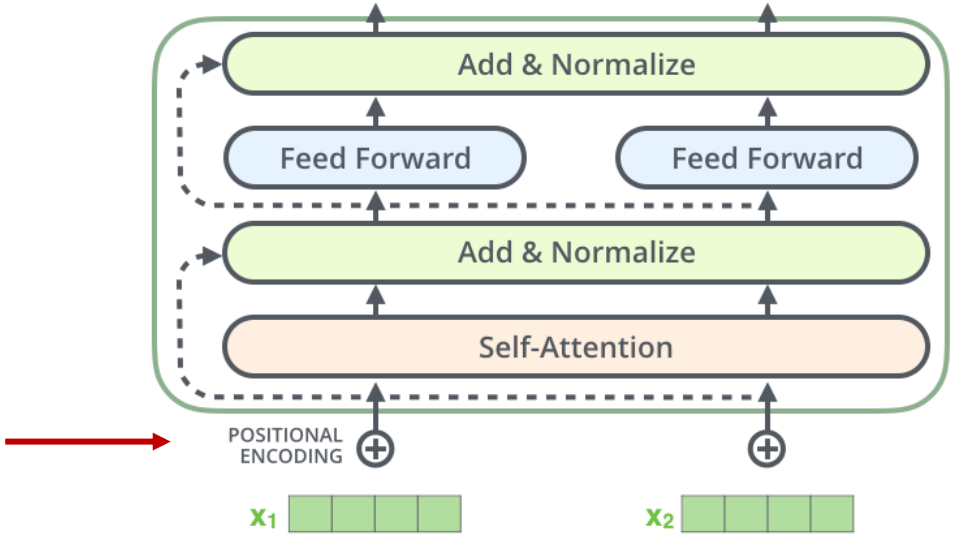
How to compute w and v parameters?

Self-Attention Computation in Transformer

- Step 1: Compute key, query, and value vectors for every token.
  - $q = x.\theta_Q$, $k = x.\theta_k$, $v = x.\theta_v$
- Step 2: Compute attention weights and probabilities
  - $w = \frac{qk^T}{\sqrt{d_k}} \Rightarrow p = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)$
- Step 3: Update representations: $z = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right).v$



A Transformer Encoder Layer

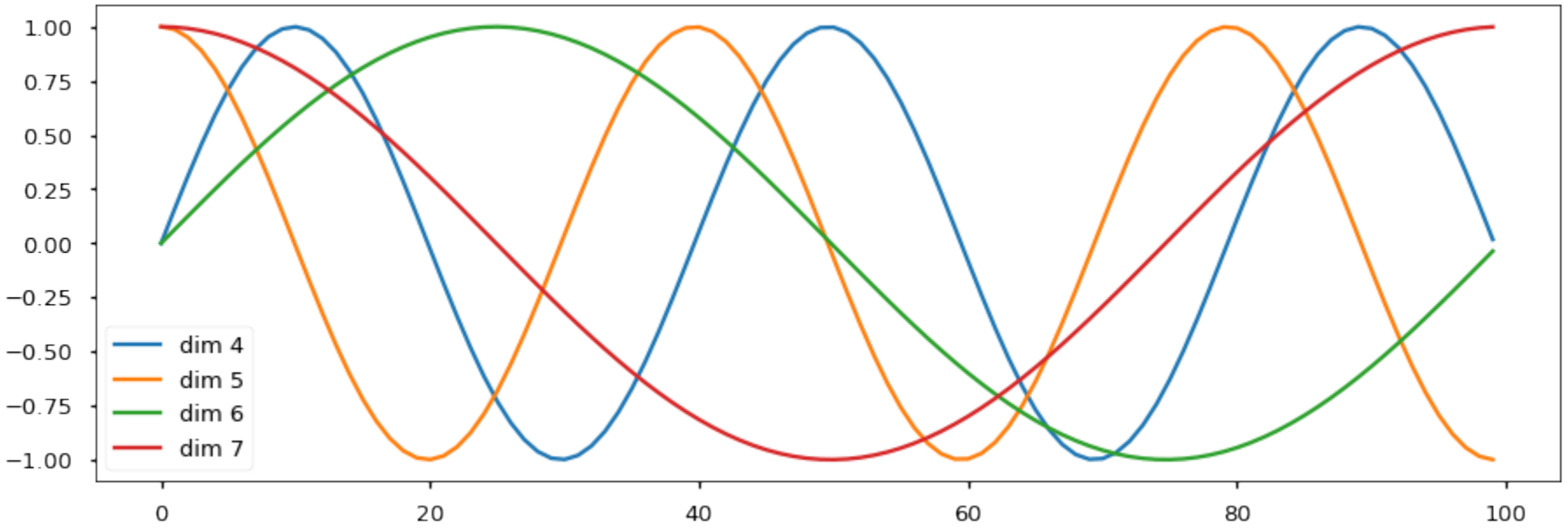


Vaswani et al. “Attention is all you need.” In NeurIPS 2017.

Positional Encoding in Transformer

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

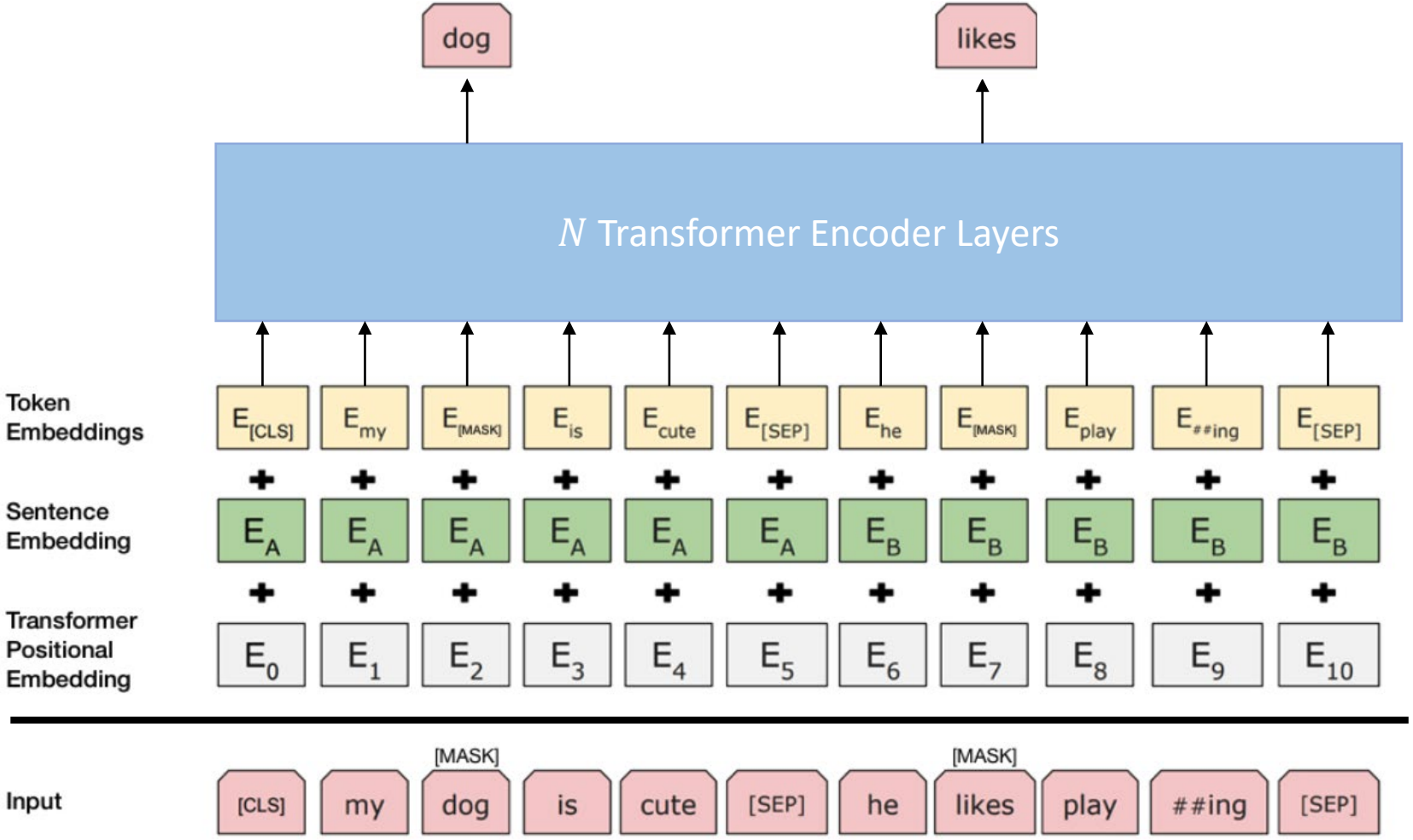


BERT: Bidirectional Encoder Representations from Transformers

- A Transformer-based neural network.
- Pre-trained based on a masked language modeling task.
- The pre-trained model can be fine-tuned for a given downstream task.
- BERT provides contextual word embeddings.

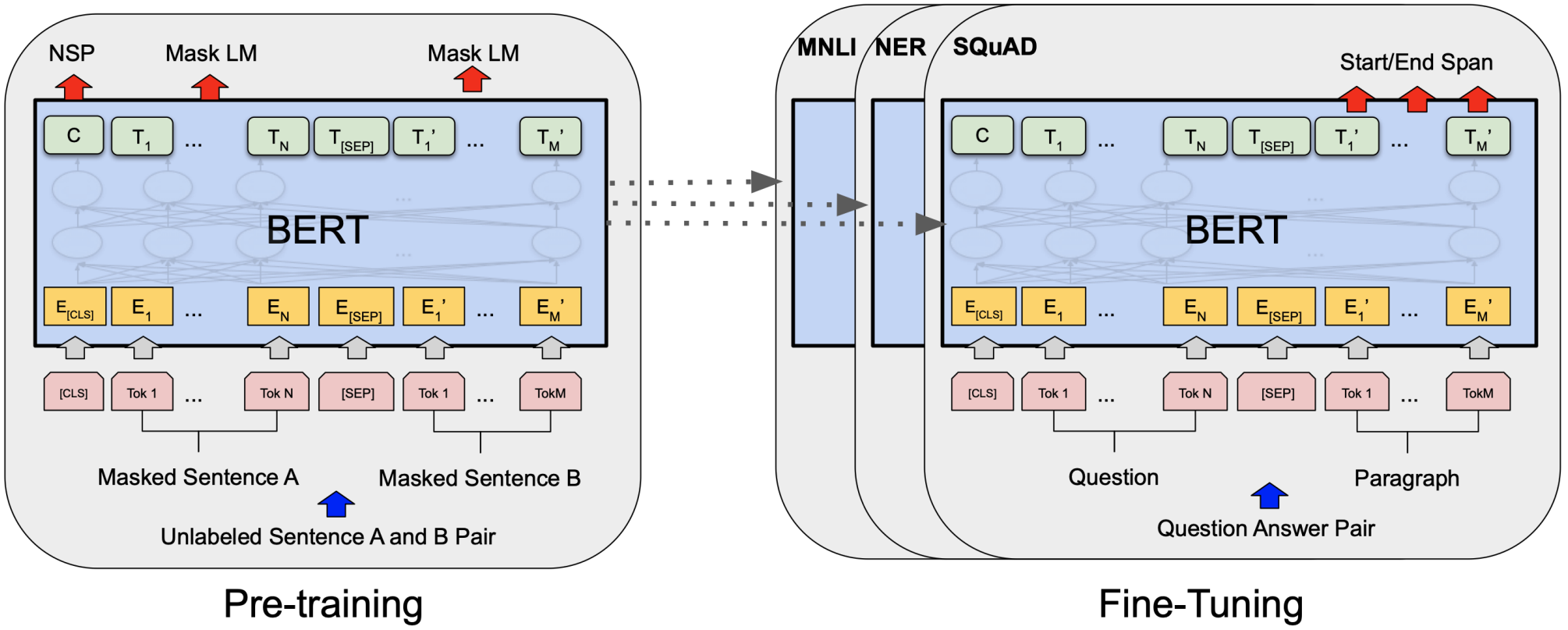
Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” In NAACL 2019.

BERT Pre-Training



Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” In NAACL 2019.

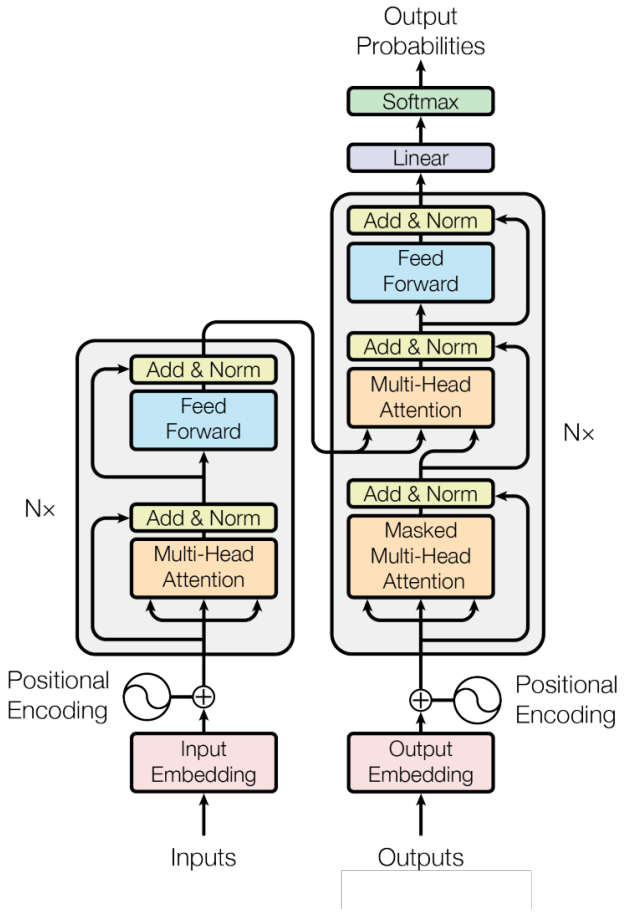
BERT Fine-Tuning



From BERT
to Generative Pre-trained Transformer (GPT)

BERT

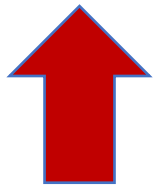
Encoder

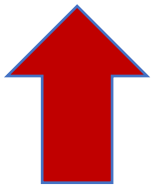


GPT

Decoder

GPT Variants

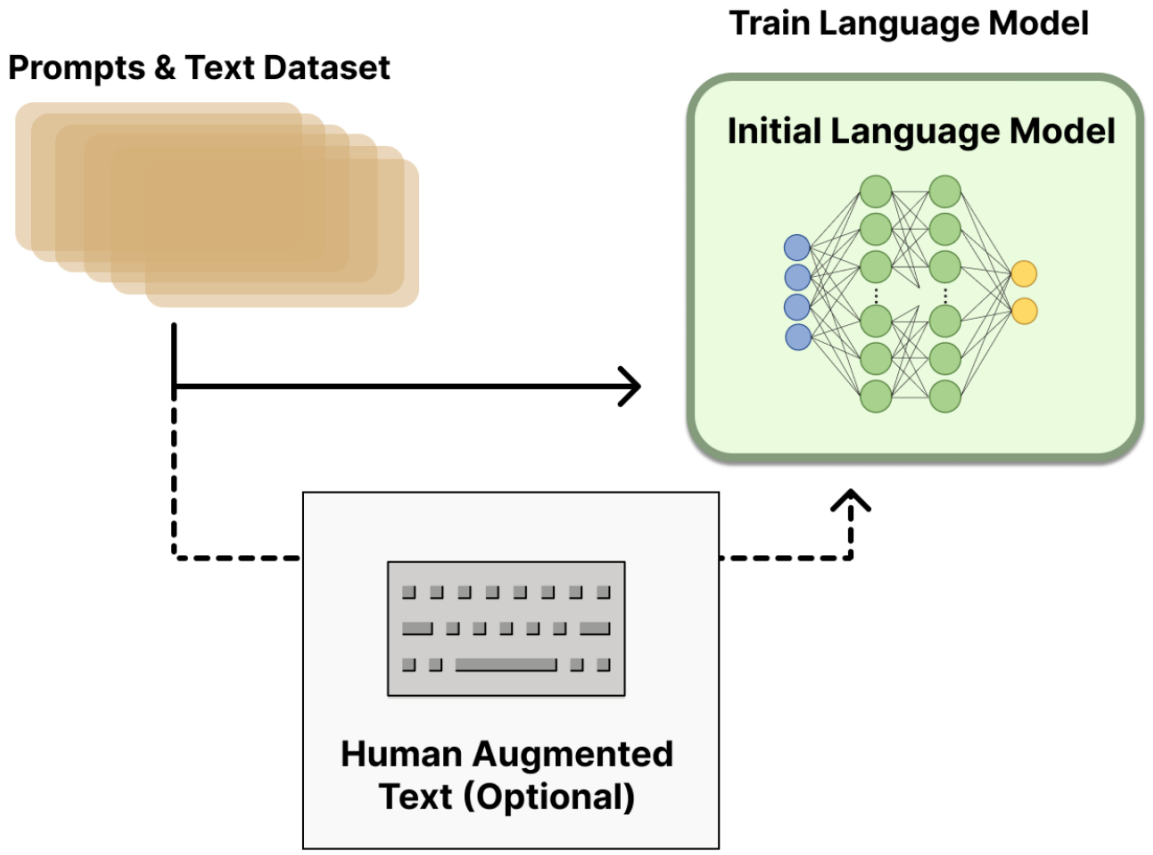




Model	Architecture	Parameter count	Training data	Release date
Original GPT (GPT-1)[22]	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus:[23] 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, 2018[5]
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)[24]
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	570 GB plaintext, 0.4 trillion tokens. Mostly CommonCrawl, WebText, English Wikipedia, and two books corpora (Books1 and Books2).	June 11, 2020[25] (then March 15, 2022, for a revision ultimately termed GPT-3.5)
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.[9]	Undisclosed	Undisclosed	March 14, 2023

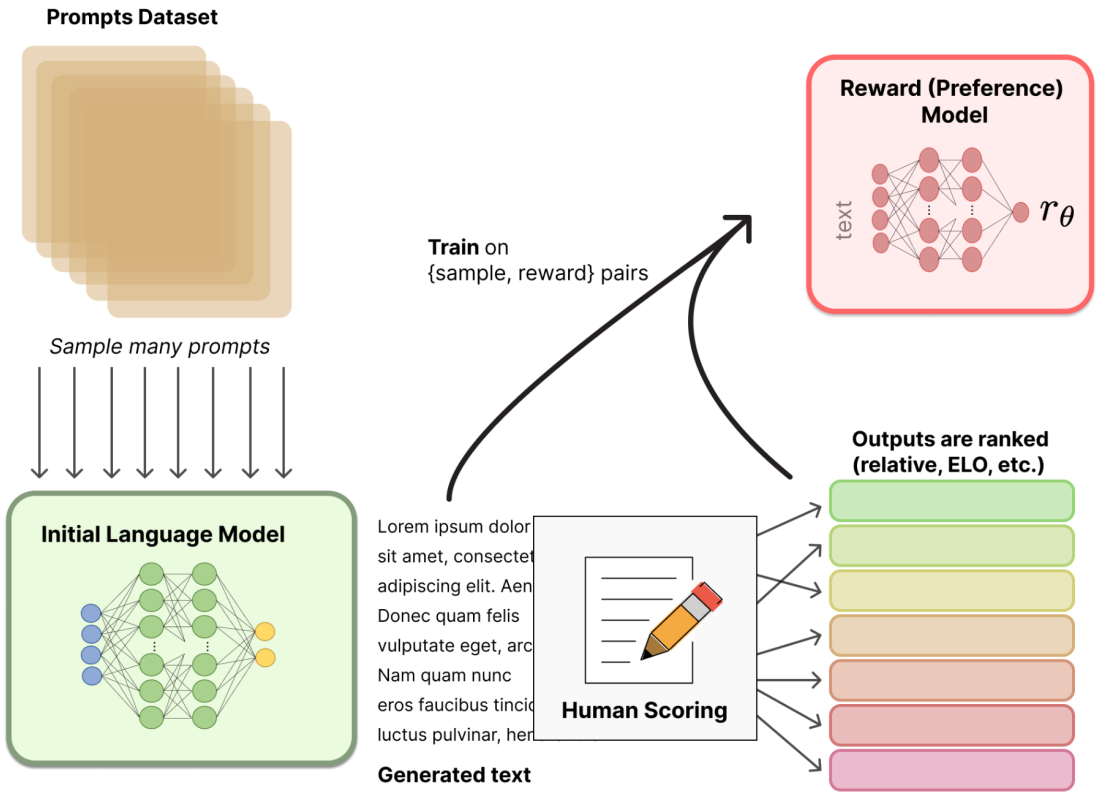
RLHF:
Reinforcement Learning with **Human Feedback**

- Step 1: Pre-training the LLM



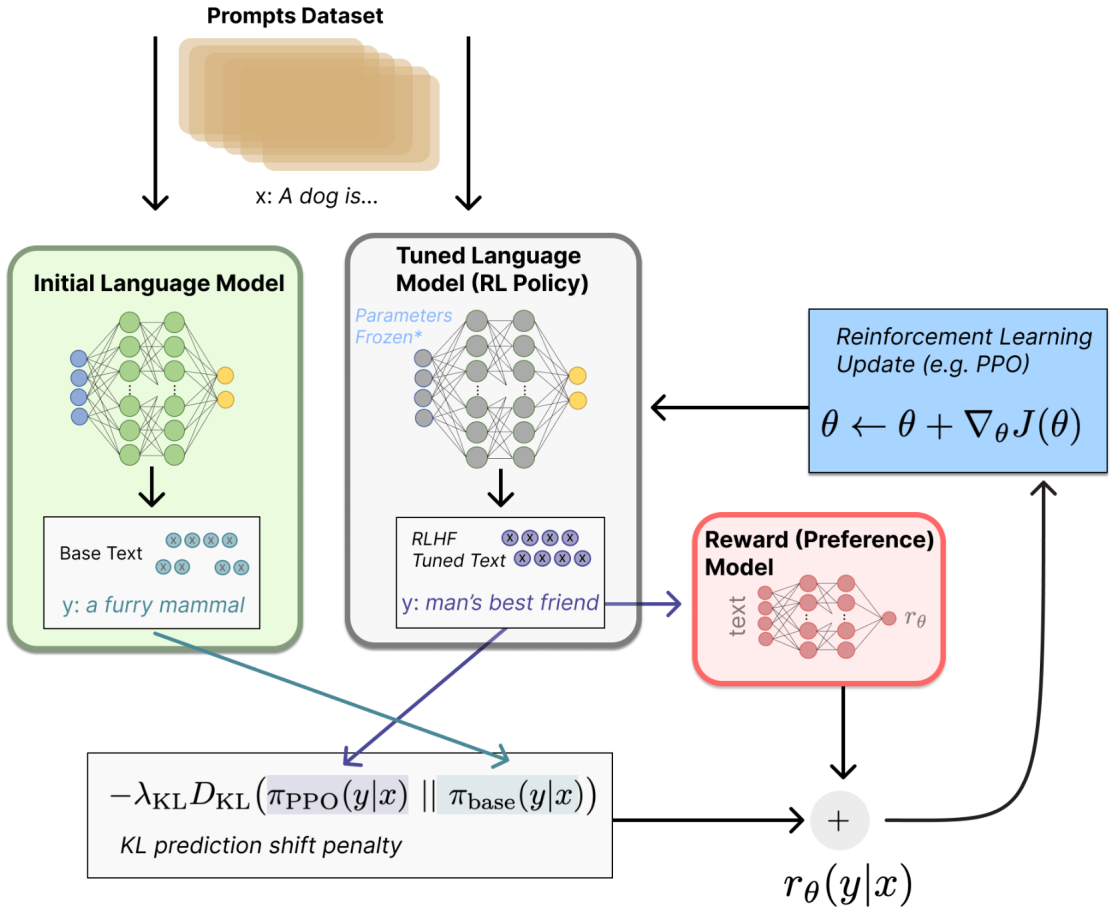
RLHF:
Reinforcement Learning with **Human Feedback**

- Step 2: Reward Model Training



RLHF:
Reinforcement Learning with **Human Feedback**

- Step 3: LLM Fine-tuning with RL
  - Using PPO: Proximal Policy Optimization



GPT-4

What is the major conference focusing on the web research?

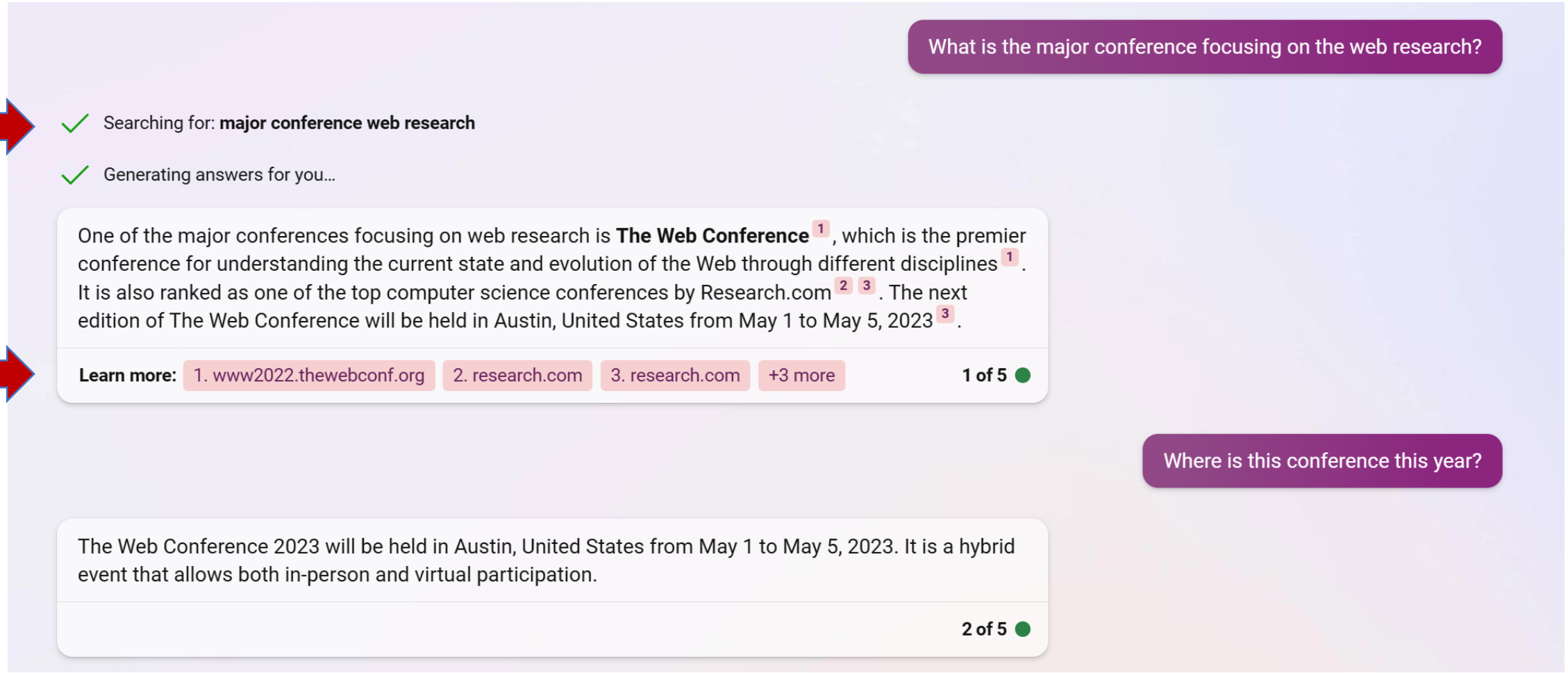
One of the major conferences focusing on web research is the World Wide Web Conference (WWW). The WWW conference is an annual international academic conference that focuses on current research and developments related to the World Wide Web. The conference brings together researchers, developers, and practitioners from different areas of computer science, such as web technologies, data mining, social networks, and machine learning, among others. The WWW conference is organized by the International World Wide Web Conference Committee (IW3C2), and it has been held annually since 1994 in different locations around the world.

Where is this conference this year?

As of my knowledge cutoff date of September 2021, the most recent edition of the World Wide Web Conference (WWW2021) was held virtually due to the COVID-19 pandemic. The conference was originally scheduled to take place in Ljubljana, Slovenia, but was moved to a fully virtual format.

As for the current year, 2023, I do not have information on the location of the upcoming WWW conference as the conference organizers have not yet announced the details. However, you can check the official website of the WWW conference (https://www2022.thewebconf.org/) for updates on the conference schedule and location as they become available.

Bing Chat


What is the major conference focusing on the web research?
Searching for: major conference web research
Generating answers for you...
One of the major conferences focusing on web research is The Web Conference [1], which is the premier conference for understanding the current state and evolution of the Web through different disciplines [1]. It is also ranked as one of the top computer science conferences by Research.com [2] [3]. The next edition of The Web Conference will be held in Austin, United States from May 1 to May 5, 2023 [3].
Learn more: 1. www2022.thewebconf.org 2. research.com 3. research.com +3 more
1 of 5
Where is this conference this year?
The Web Conference 2023 will be held in Austin, United States from May 1 to May 5, 2023. It is a hybrid event that allows both in-person and virtual participation.
2 of 5

Bard

What is the major conference focusing on the web research?

There are many major conferences focusing on web research. Some of the most well-known include:

- The International World Wide Web Conference (WWW)
- The Conference on Information and Knowledge Management (CIKM)
- The ACM SIGIR Conference on Research and Development in Information Retrieval
- The ACM SIGWEB Conference on the World Wide Web
- The Web Science Conference

These conferences bring together researchers from academia, industry, and government to share their latest work on the web, including topics such as:

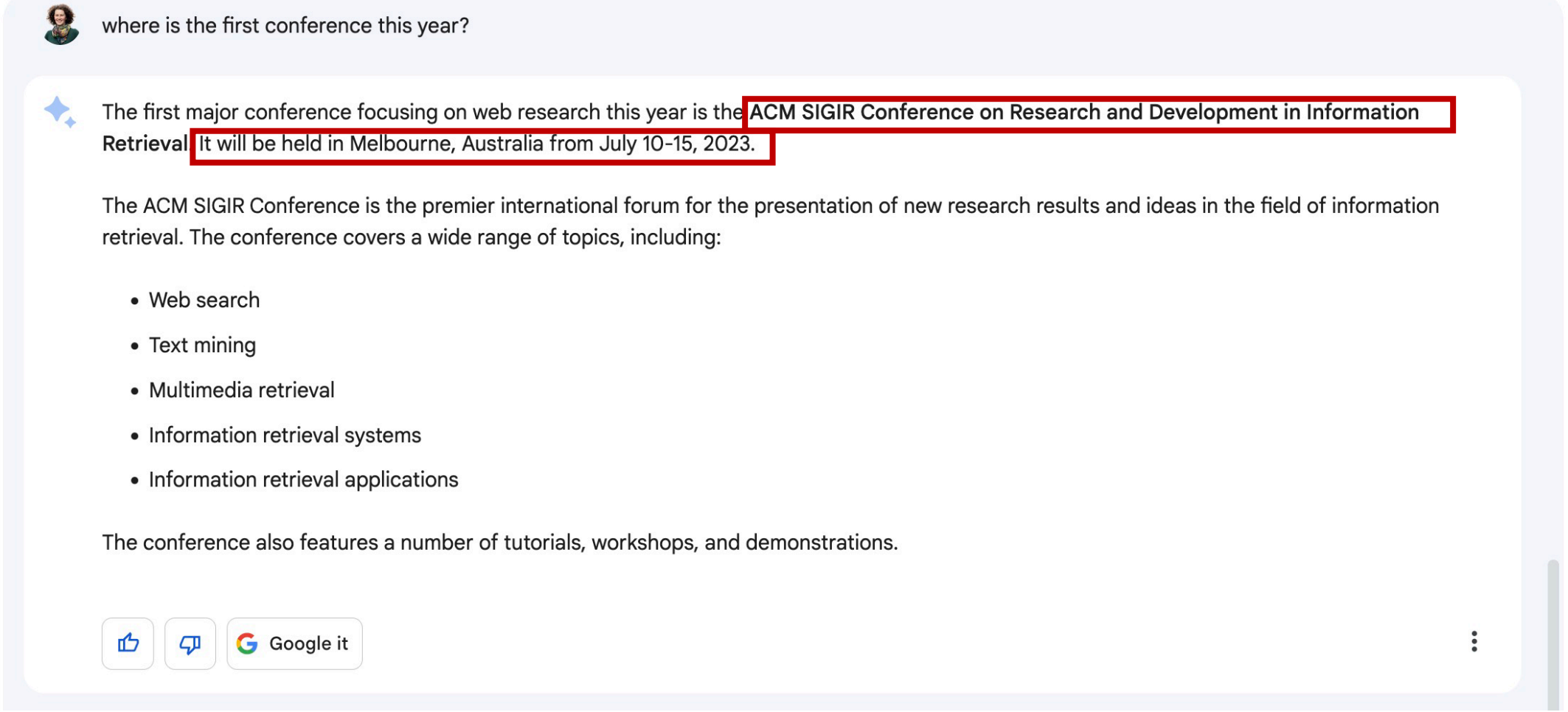
- Web search
- Web mining
- Web semantics
- Web architecture
- Web security
- Web accessibility

The conferences also provide a forum for researchers to network and collaborate on future projects.

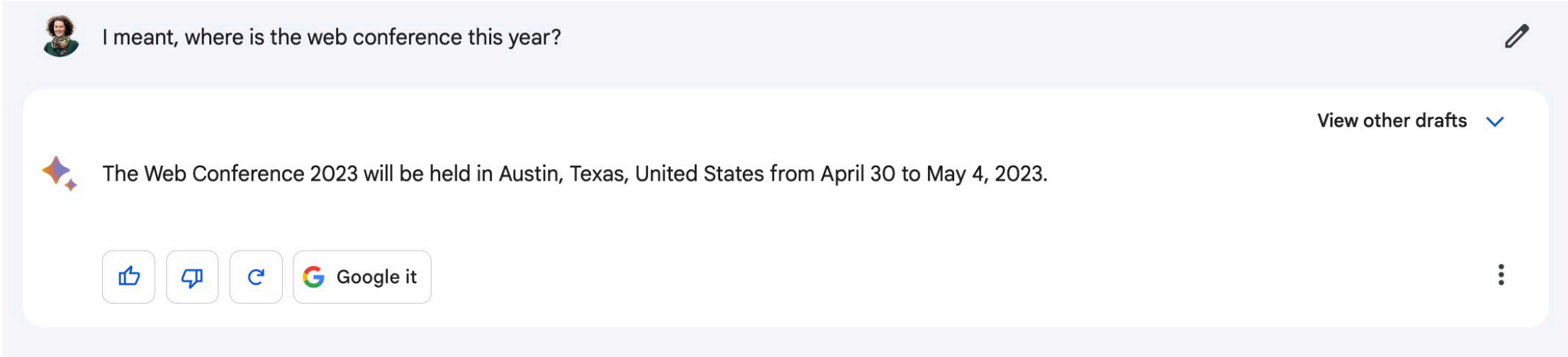
Enter a prompt here

Bard may display inaccurate or offensive information that doesn't represent Google's views.

Bard


where is the first conference this year?
The first major conference focusing on web research this year is the ACM SIGIR Conference on Research and Development in Information Retrieval. It will be held in Melbourne, Australia from July 10-15, 2023.
The ACM SIGIR Conference is the premier international forum for the presentation of new research results and ideas in the field of information retrieval. The conference covers a wide range of topics, including:
Web search
Text mining
Multimedia retrieval
Information retrieval systems
Information retrieval applications
The conference also features a number of tutorials, workshops, and demonstrations.
Google it

Bard


I meant, where is the web conference this year?
View other drafts
The Web Conference 2023 will be held in Austin, Texas, United States from April 30 to May 4, 2023.
Google it

Retrieval Augmentation

- Knowledge-Intensive Tasks
  - Open-domain question answering
  - Task-oriented dialogues
  - Fact verification
  - …
- Personalizing language models

- Notable Benchmarks:
  - KILT Benchmark
  - OK-VQA
  - LaMP

KILT: a Benchmark for Knowledge Intensive Language Tasks

Fabio Petroni[1] **Aleksandra Piktus**[1] **Angela Fan**[1,3] **Patrick Lewis**[1,2]
Majid Yazdani[1] **Nicola De Cao**[6] **James Thorne**[4] **Yacine Jernite**[5] **Vladimir Karpukhin**[1]
Jean Maillard[1] **Vassilis Plachouras**[1] **Tim Rocktäschel**[1,2] **Sebastian Riedel**[1,2]
[1]Facebook AI Research [2]University College London [3]LORIA
[4]University of Cambridge [5]HuggingFace [6]University of Amsterdam

OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge

Kenneth Marino*[1], Mohammad Rastegari[2], Ali Farhadi[2,3] and Roozbeh Mottaghi[2]

[1]Carnegie Mellon University
[2]PRIOR @ Allen Institute for AI
[3]University of Washington

LaMP: When Large Language Models Meet Personalization

Alireza Salemi[1], Sheshera Mysore[1], Michael Bendersky[2], Hamed Zamani[1]
[1]University of Massachusetts Amherst
[2]Google Research

Retrieval Augmentation

- Retrieval augmentation in the prompt
  - OpenAI Retrieval Plugin
- Fusion in Decoder (FiD), RAG, RetGen

Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

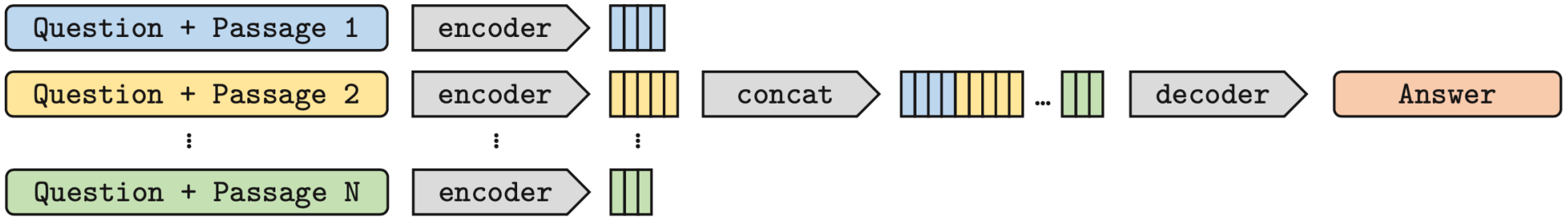
Gautier Izacard[1,2,3] Edouard Grave[1]
[1] Facebook AI Research, Paris
[2] ENS, PSL University, Paris
[3] Inria, Paris

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[⋆],
Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]
[†]Facebook AI Research; [‡]University College London; [⋆]New York University;

RetGen: A Joint framework for Retrieval and Grounded Text Generation Modeling

Yizhe Zhang [*] Siqi Sun Xiang Gao Yuwei Fang
Chris Brockett Michel Galley Jianfeng Gao Bill Dolan
Microsoft Corporation, Redmond, WA, USA



Current SOTA on Retrieval Augmentation

- On KILT:
  - FiD-Light (six datasets)
  - GENRE (two datasets)
- On OK-VQA
  - Prophet
  - PromptCat

FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation

Sebastian Hofstätter TU Wien (Google Internship)
Jiecao Chen Google
Karthik Raman Google
Hamed Zamani University of Massachusetts Amherst

Autoregressive Entity Retrieval

Nicola De Cao[1,2]*, Gautier Izacard[2,3,4], Sebastian Riedel[2,5], Fabio Petroni[2]
[1]University of Amsterdam, [2]Facebook AI Research
[3]ENS, PSL University, [4]Inria, [5]University College London

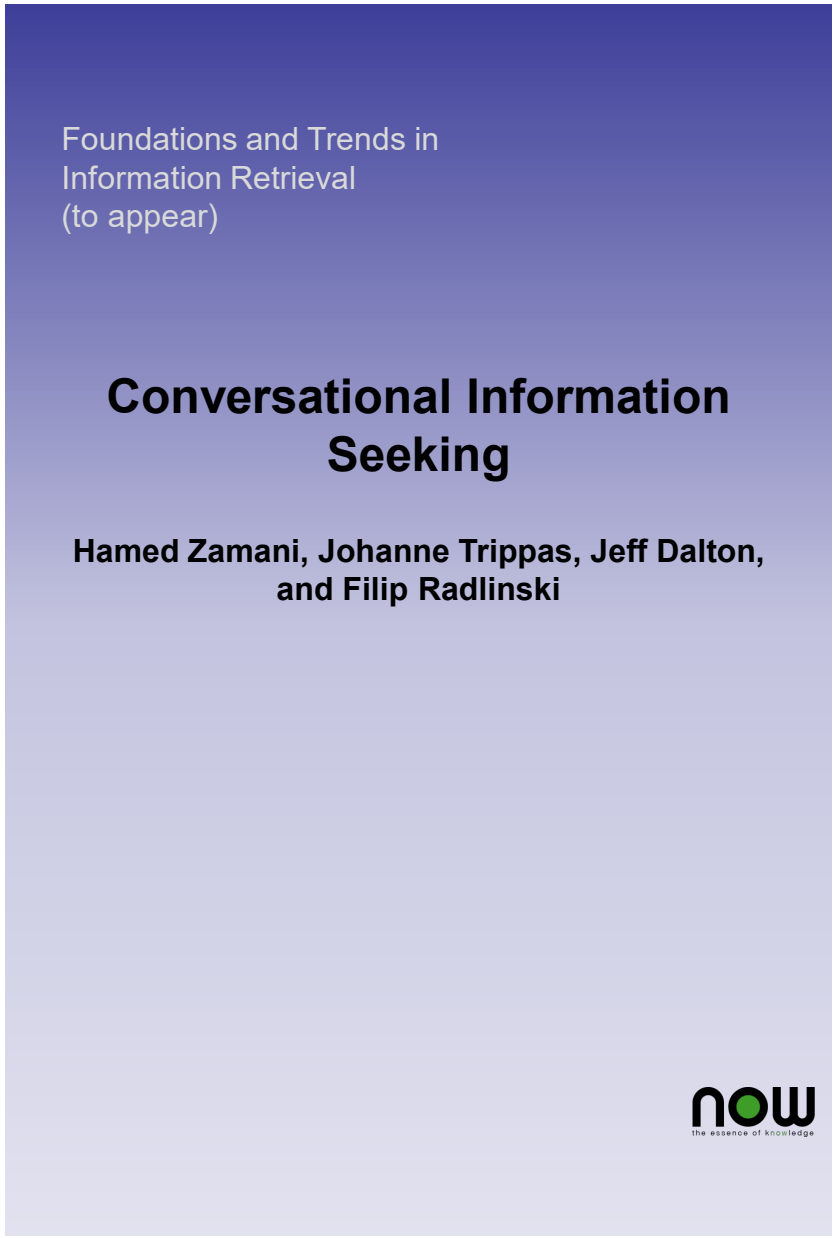
Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering

Zhenwei Shao[1] Zhou Yu[1*] Meng Wang[2] Jun Yu[1]
[1]Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, China.
[2]School of Computer Science and Information Engineering, Hefei University of Technology, China

PromptCap: Prompt-Guided Task-Aware Image Captioning

Yushi Hu[1*] Hang Hua[2*] Zhengyuan Yang[3]
Weijia Shi[1] Noah A. Smith[1,4] Jiebo Luo[2]
[1]University of Washington [2]University of Rochester
[3]Microsoft [4]Allen Institute for AI

Questions



https://arxiv.org/pdf/2201.08808.pdf