Natural Language Processing: RAG, Tools

HSE Faculty of Computer Science

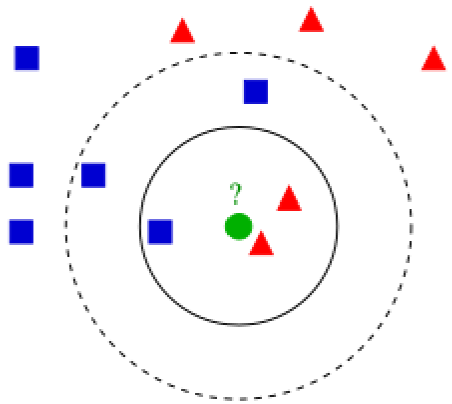
Machine Learning and Data-Intensive Systems

Murat Khazhgeriev

Table of Content

- **Approximate kNN**
- Retrieval-Augmented Generation (RAG)
- Introducing graphs to the system
- Agents

Vanilla kNN



Source: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Hierarchical Small Navigable World (HNSW)

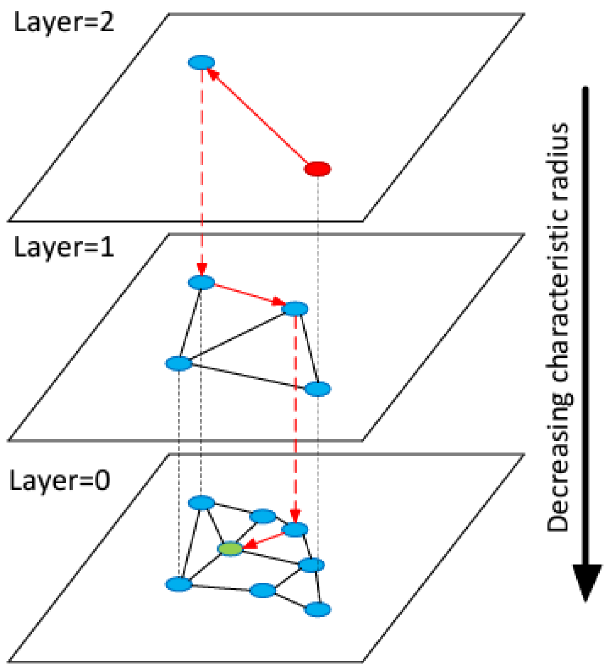
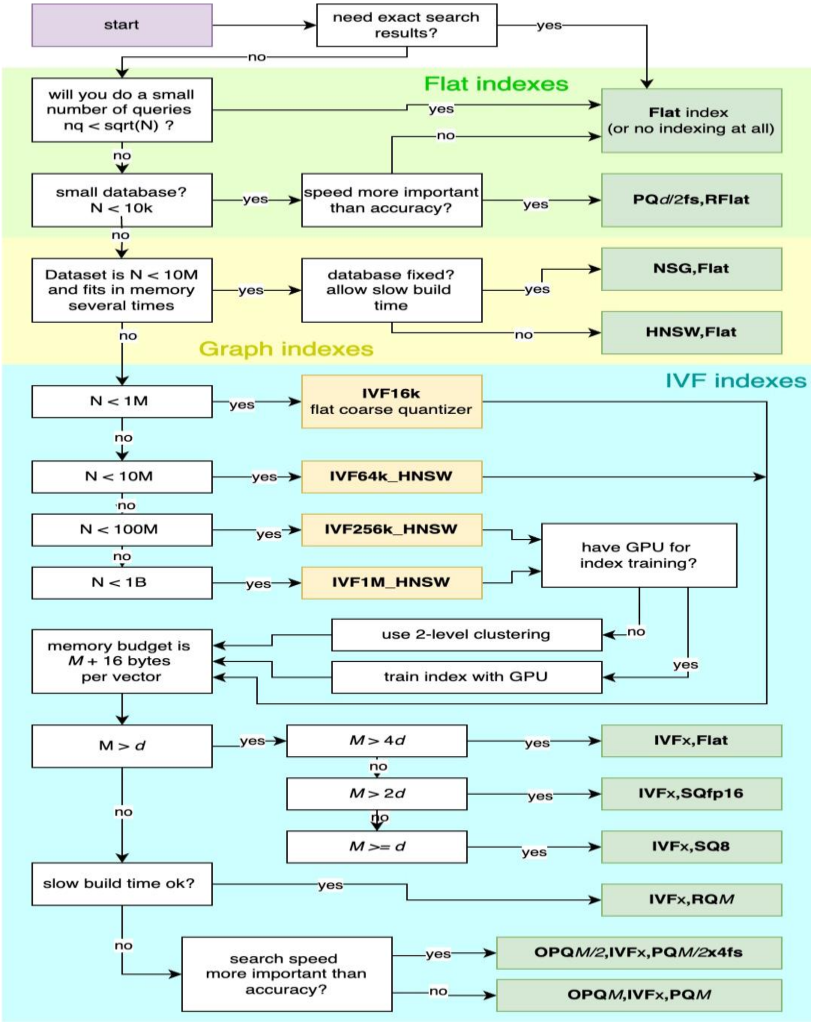


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

Source: https://arxiv.org/abs/1603.09320

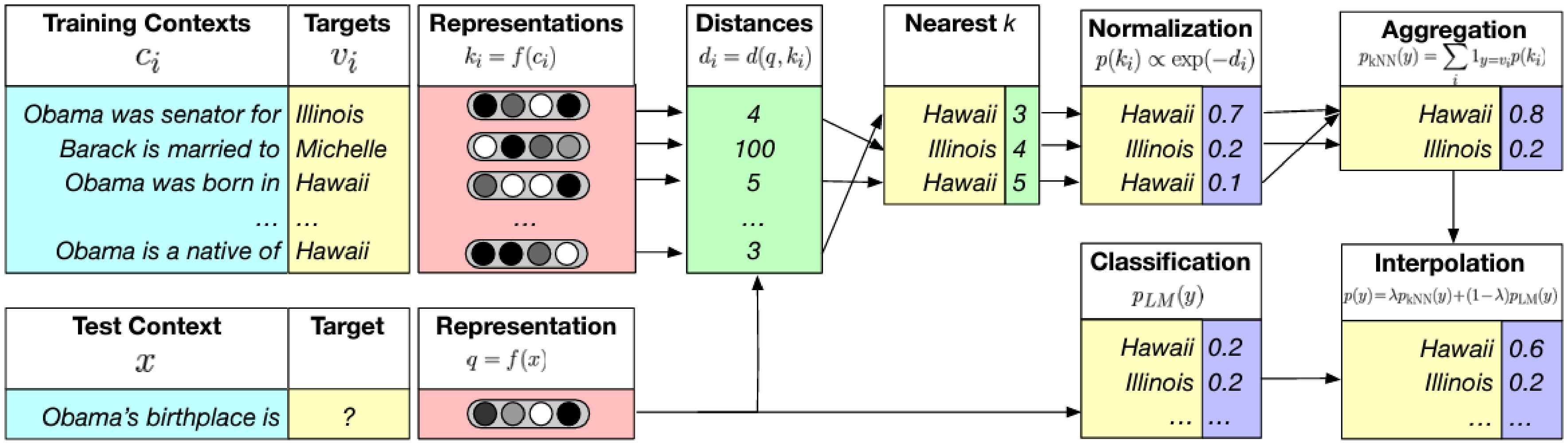
FAISS



Source: https://github.com/facebookresearch/faiss/wiki/Guidelines-to-choose-an-index

Table of Content

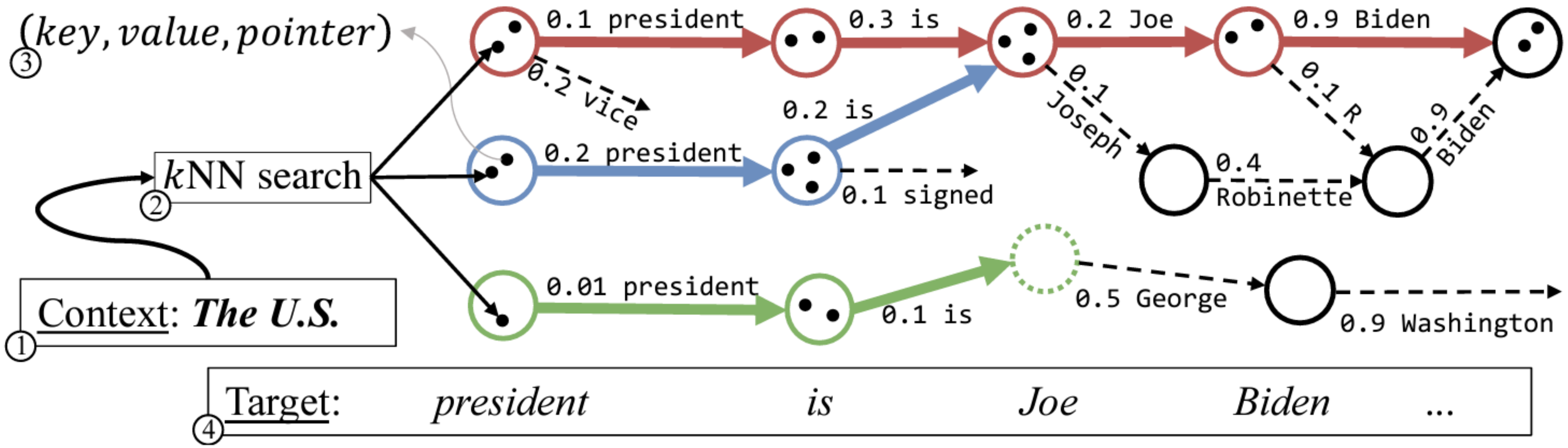
Generalization through Memorization: Nearest Neighbor Language Models



$$p(y|x) = \lambda\, p_{\text{kNN}}(y|x) + (1-\lambda)\, p_{\text{LM}}(y|x)$$

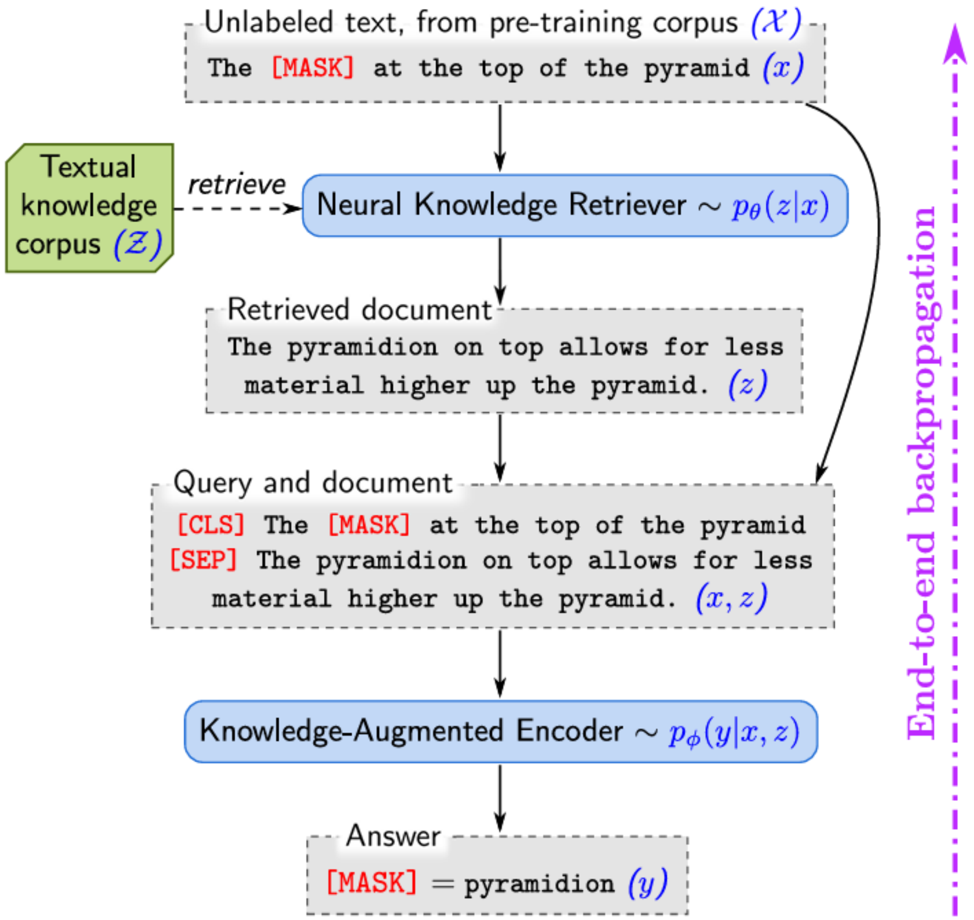
Source: https://arxiv.org/abs/1911.00172

Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval



Source: https://arxiv.org/abs/2201.12431

REALM: Retrieval-Augmented Language Model Pre-Training



Weighted probability for each neighbor is the answer:

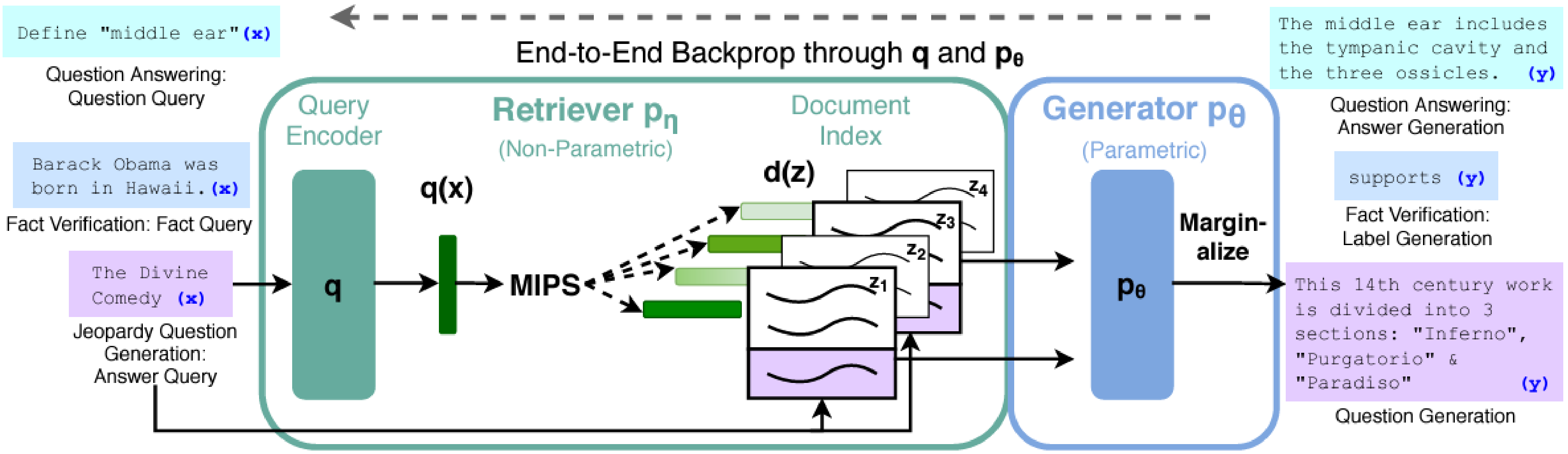
$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x).$$

The closer retrieved data, the more weight it has:

$$p(z \mid x) = \frac{\exp f(x, z)}{\sum_{z'} \exp f(x, z')},$$

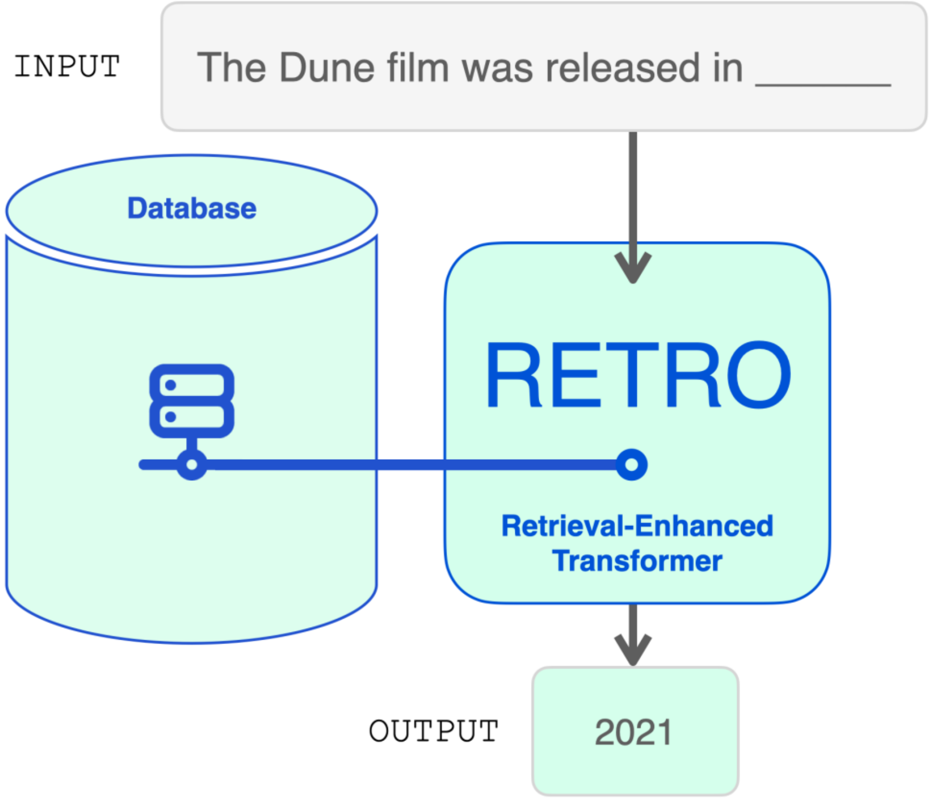
$$f(x, z) = \texttt{Embed}_{\texttt{input}}(x)^\top \texttt{Embed}_{\texttt{doc}}(z),$$

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks



Source: https://arxiv.org/abs/2005.11401

Improving language models by retrieving from trillions of tokens



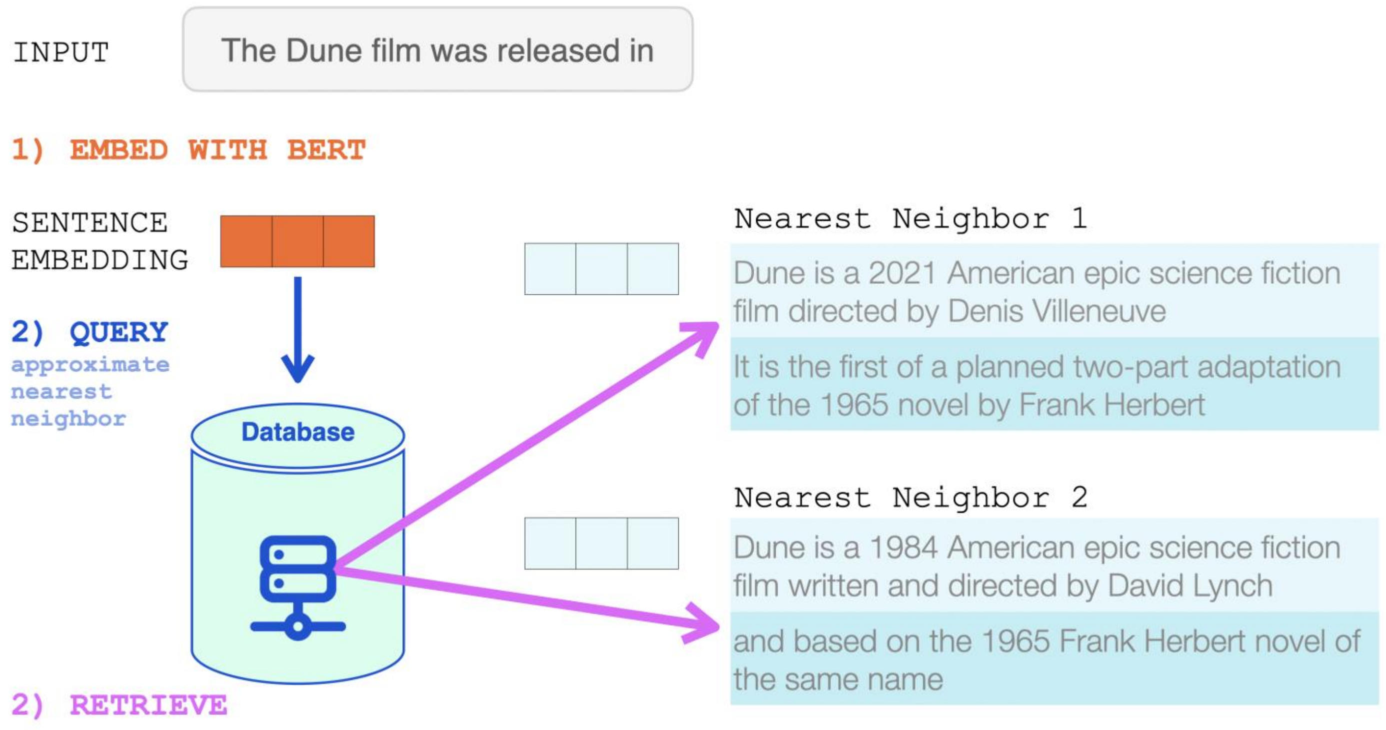
Source: https://arxiv.org/abs/2005.11401, image source: https://jalammar.github.io/illustrated-retrieval-transformer/

Improving language models by retrieving from trillions of tokens

Key (BERT sentence embedding)	Value (text. neighbor and completion chunks. Each up to 64 tokens in length)	
	Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	NEIGHBOR
	It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	COMPLETION
	Dune is a 1965 science fiction novel by American author Frank Herbert	NEIGHBOR
	originally published as two separate serials in Analog magazine	COMPLETION
...	...	

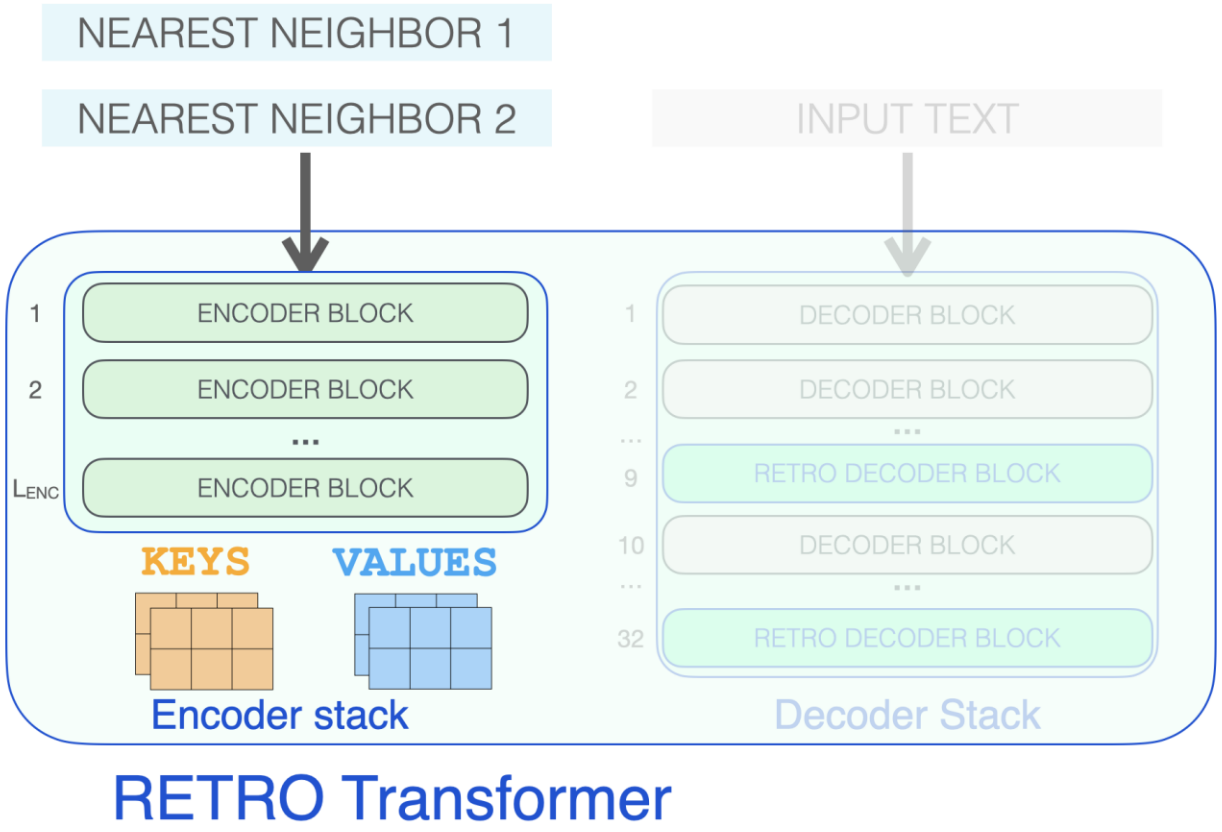
Source: https://arxiv.org/abs/2005.11401, image source: https://jalammar.github.io/illustrated-retrieval-transformer/

Improving language models by retrieving from trillions of tokens



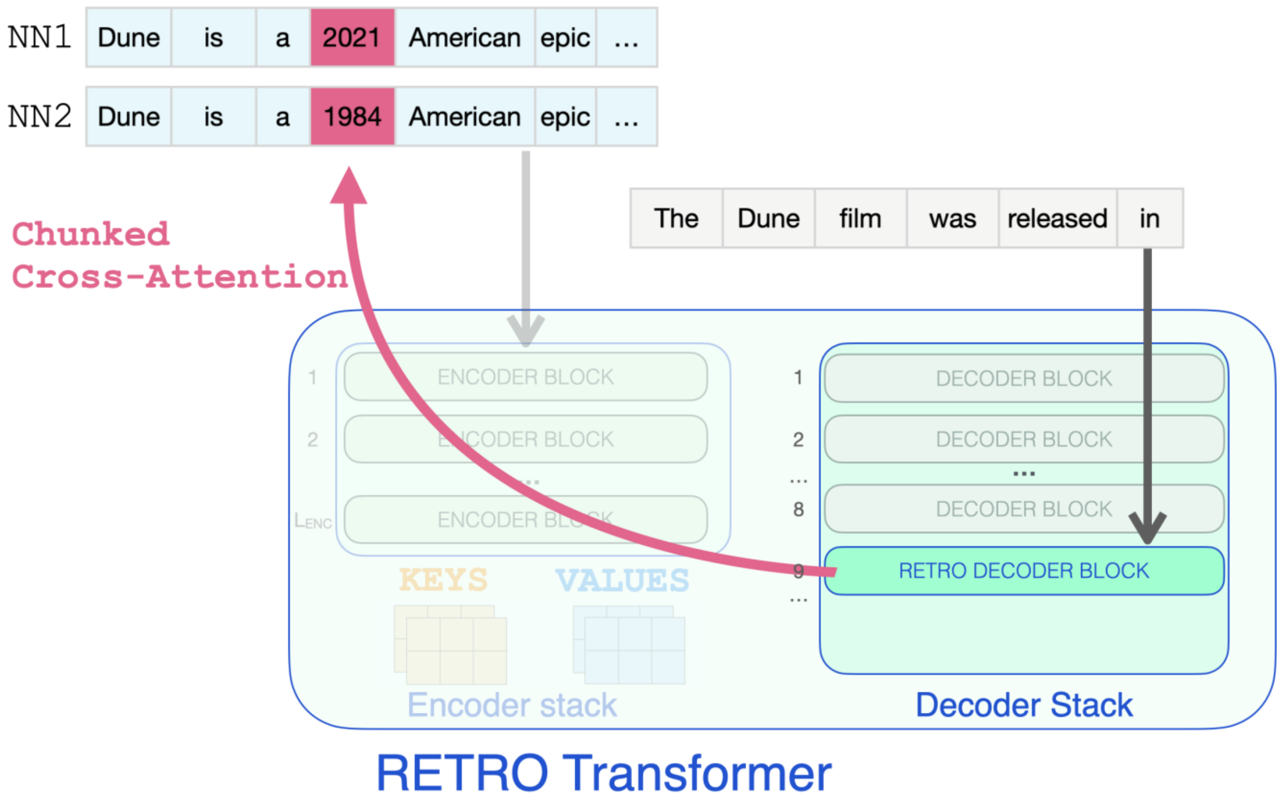
Source: https://arxiv.org/abs/2005.11401, image source: https://jalammar.github.io/illustrated-retrieval-transformer/

Improving language models by retrieving from trillions of tokens



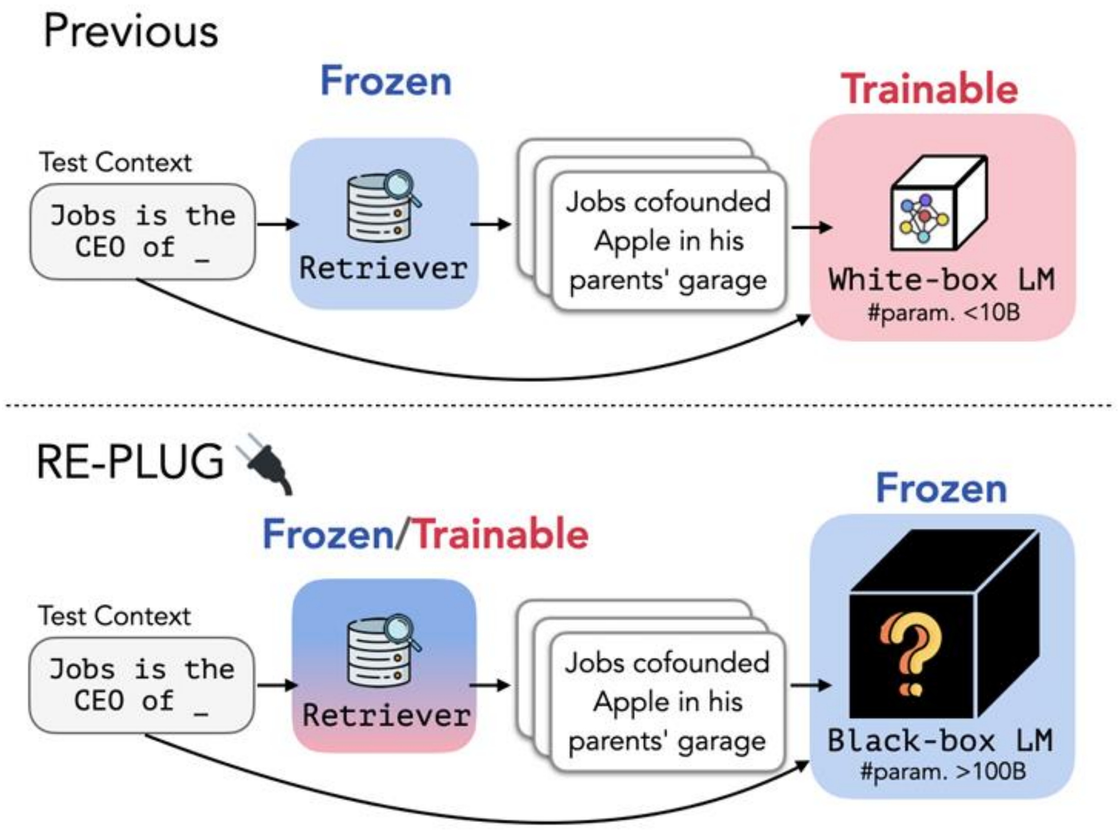
Source: https://arxiv.org/abs/2005.11401, image source: https://jalammar.github.io/illustrated-retrieval-transformer/

Improving language models by retrieving from trillions of tokens



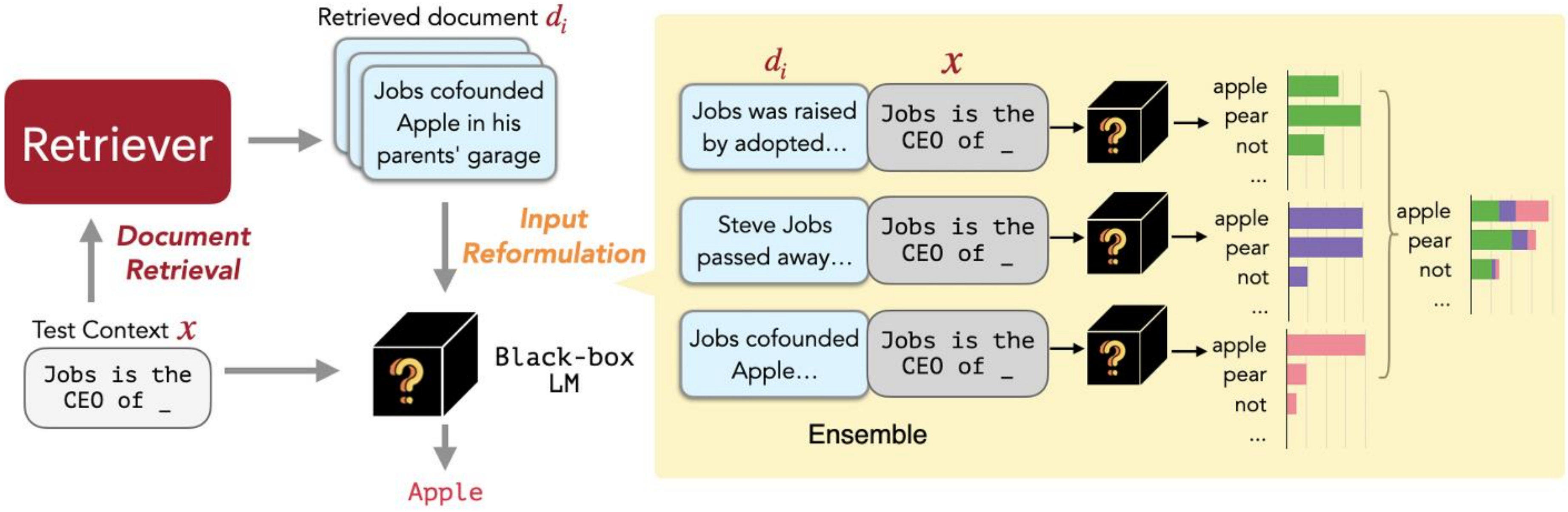
Source: https://arxiv.org/abs/2005.11401, image source: https://jalammar.github.io/illustrated-retrieval-transformer/

REPLUG: Retrieval-Augmented Black-Box Language Models



Source: https://arxiv.org/abs/2301.12652

REPLUG: Retrieval-Augmented Black-Box Language Models



Source: https://arxiv.org/abs/2301.12652

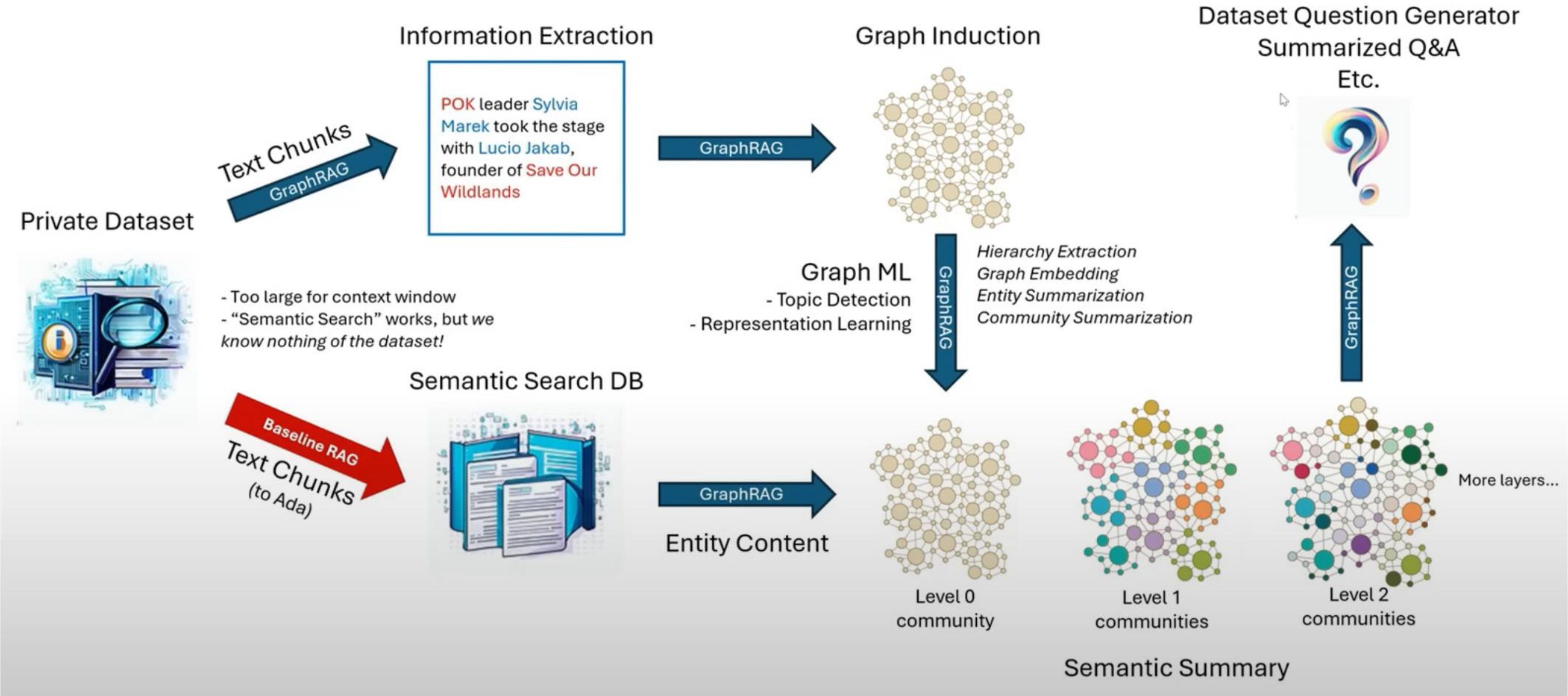
Retrieval meets Long Context Large Language Models

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

Source: https://arxiv.org/abs/2310.03025

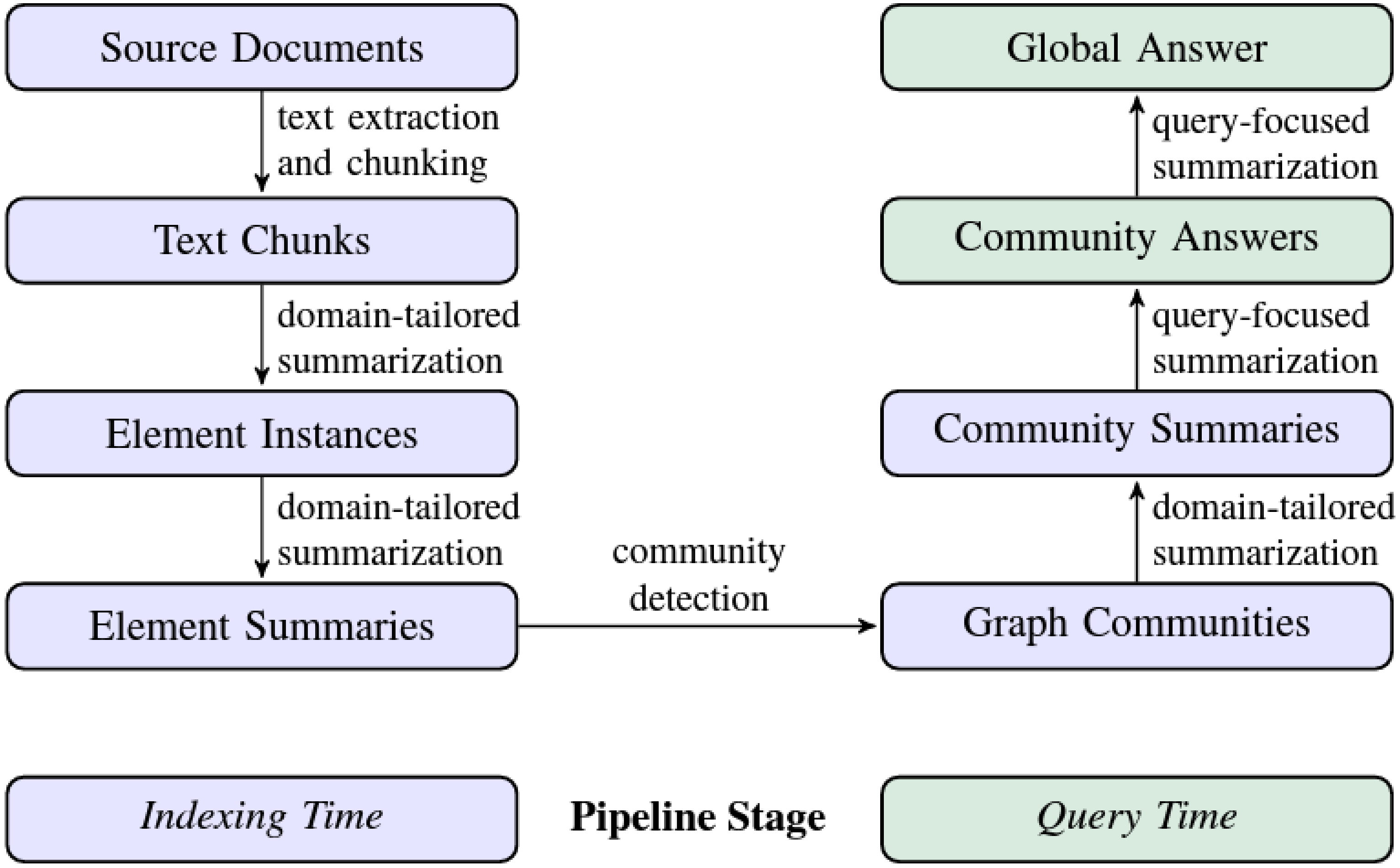
Table of Content

From Local to Global: A Graph RAG Approach to Query-Focused Summarization



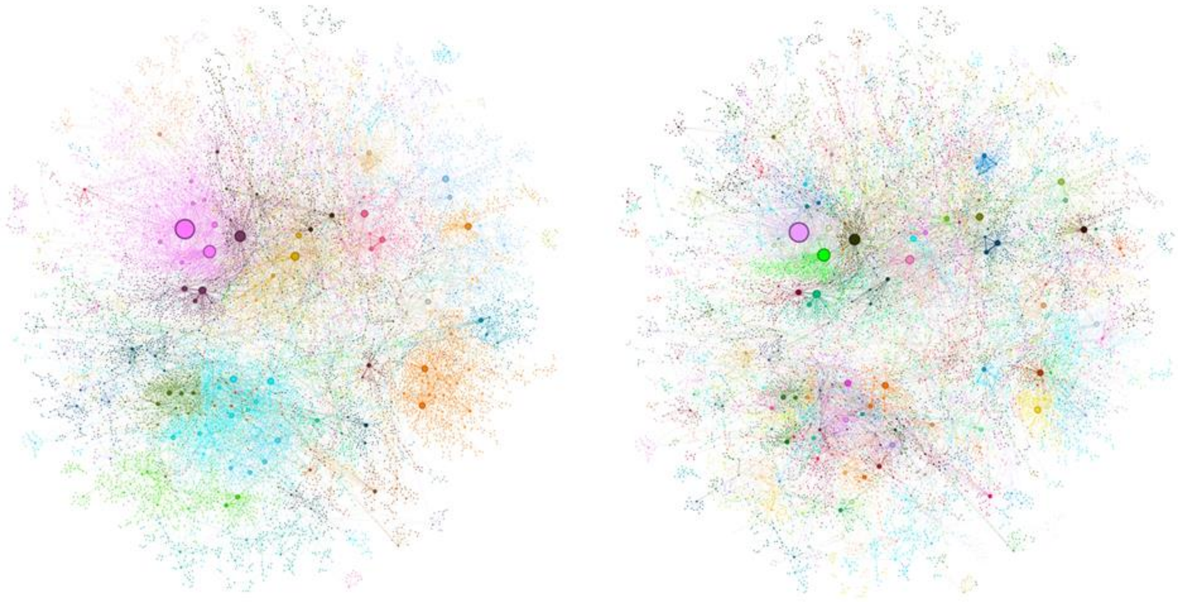
Source: https://www.youtube.com/watch?v=r09tJfON6kE&t=778s&ab_channel=AlexChao

From Local to Global: A Graph RAG Approach to Query-Focused Summarization



Source: https://microsoft.github.io/graphrag/

From Local to Global: A Graph RAG Approach to Query-Focused Summarization



(a) Root communities at level 0

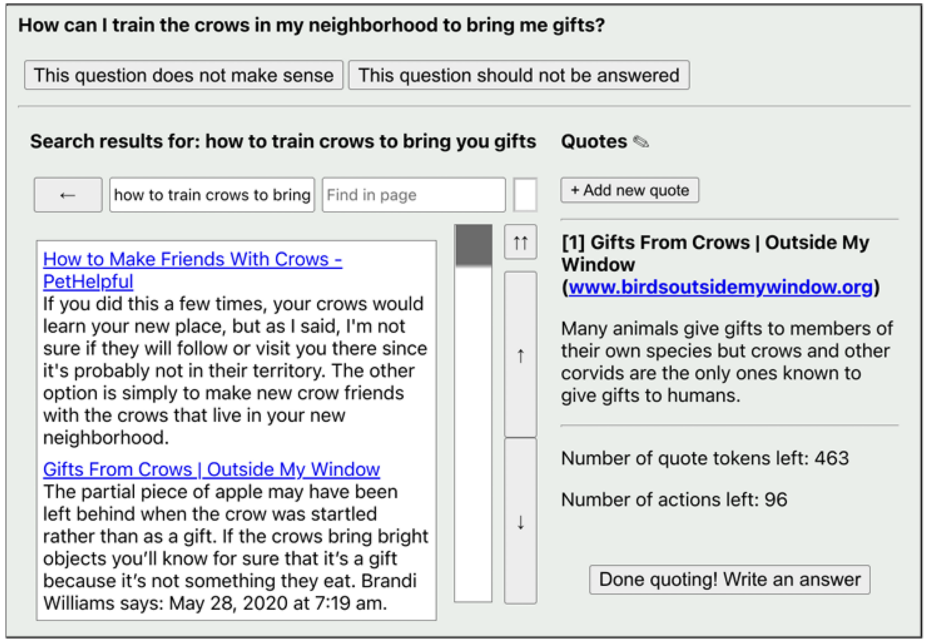
(b) Sub-communities at level 1

Figure 3: Graph communities detected using the Leiden algorithm (Traag et al., 2019) over the MultiHop-RAG (Tang and Yang, 2024) dataset as indexed. Circles represent entity nodes with size proportional to their degree. Node layout was performed via OpenORD (Martin et al., 2011) and Force Atlas 2 (Jacomy et al., 2014). Node colors represent entity communities, shown at two levels of hierarchical clustering: (a) Level 0, corresponding to the hierarchical partition with maximum modularity, and (b) Level 1, which reveals internal structure within these root-level communities.

Source: https://microsoft.github.io/graphrag/

Table of Content

WebGPT



(a) Screenshot from the demonstration interface.

```
♦Question
How can I train the crows in my neighborhood to bring me gifts?

♦Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemywindow.org)
> Many animals give gifts to members of their own species but crows and
other corvids are the only ones known to give gifts to humans.

♦Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemywindow.org
Quote
Back

♦Title
Search results for: how to train crows to bring you gifts

♦Scrollbar: 0 - 11
♦Text
【0†How to Make Friends With Crows - PetHelpful†pethelpful.com】
If you did this a few times, your crows would learn your new place, but
as I said, I'm not sure if they will follow or visit you there since it's
probably not in their territory. The other option is simply to make new
crow friends with the crows that live in your new neighborhood.

【1†Gifts From Crows | Outside My Window†www.birdsoutsidemywindow.org】
The partial piece of apple may have been left behind when the crow was
startled rather than as a gift. If the crows bring bright objects you'll
know for sure that it's a gift because it's not something they eat.
Brandi Williams says: May 28, 2020 at 7:19 am.

♦Actions left: 96
♦Next action
```

(b) Corresponding text given to the model.

Figure 1: An observation from our text-based web-browsing environment, as shown to human demonstrators (left) and models (right). The web page text has been abridged for illustrative purposes.

Source: https://openai.com/index/webgpt/

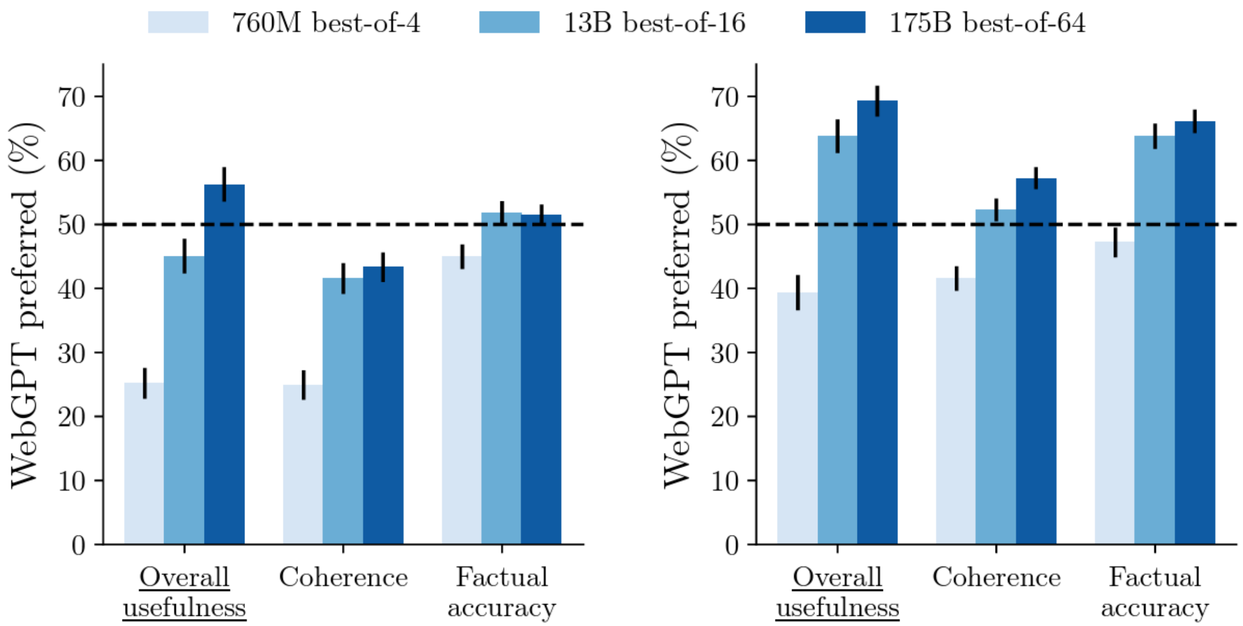
WebGPT

Table 1: Actions the model can take. If a model generates any other text, it is considered to be an invalid action. Invalid actions still count towards the maximum, but are otherwise ignored.

Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

Source: https://openai.com/index/webgpt/

WebGPT



(a) WebGPT vs. human demonstrations.

(b) WebGPT vs. ELI5 reference answers.

Figure 2: Human evaluations on ELI5 comparing against (a) demonstrations collected using our web browser, (b) the highest-voted answer for each question. The amount of rejection sampling (the n in best-of-n) was chosen to be compute-efficient (see Figure 8). Error bars represent ± 1 standard error.

Source: https://openai.com/index/webgpt/

WebGPT

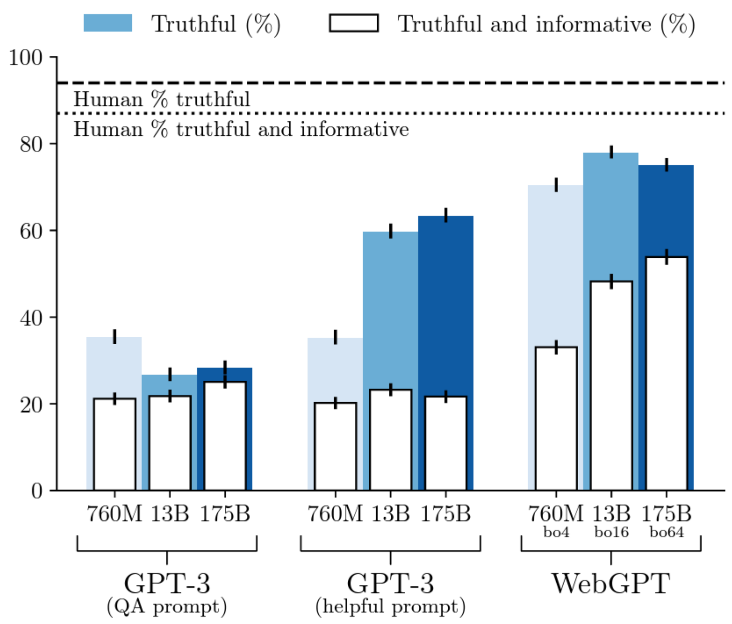


Figure 3: TruthfulQA results. The amount of rejection sampling (the n in best-of-n) was chosen to be compute-efficient (see Figure 8). Error bars represent ± 1 standard error.

Source: https://openai.com/index/webgpt/

Toolformer: Language Models Can Teach Themselves to Use Tools

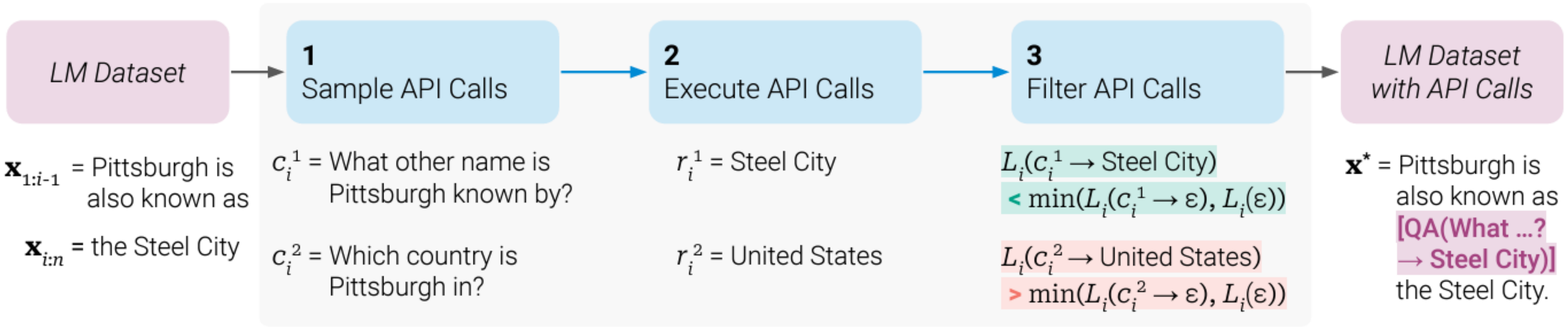


Figure 2: Key steps in our approach, illustrated for a *question answering* tool: Given an input text $\mathbf{x}$, we first sample a position i and corresponding API call candidates $c_i^1, c_i^2, \ldots, c_i^k$. We then execute these API calls and filter out all calls which do not reduce the loss L_i over the next tokens. All remaining API calls are interleaved with the original text, resulting in a new text $\mathbf{x}^*$.

Source: https://arxiv.org/abs/2302.04761

Toolformer: Language Models Can Teach Themselves to Use Tools

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: $\mathbf{x}$

Output:

Figure 3: An exemplary prompt $P(\mathbf{x})$ used to generate API calls for the question answering tool.

Source: https://arxiv.org/abs/2302.04761

Toolformer: Language Models Can Teach Themselves to Use Tools

The New England Journal of Medicine is a registered trademark of **[QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society]** the MMS.

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

The name derives from "la tortuga", the Spanish word for **[MT("tortuga") → turtle]** turtle.

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.

Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text.

Source: https://arxiv.org/abs/2302.04761

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face

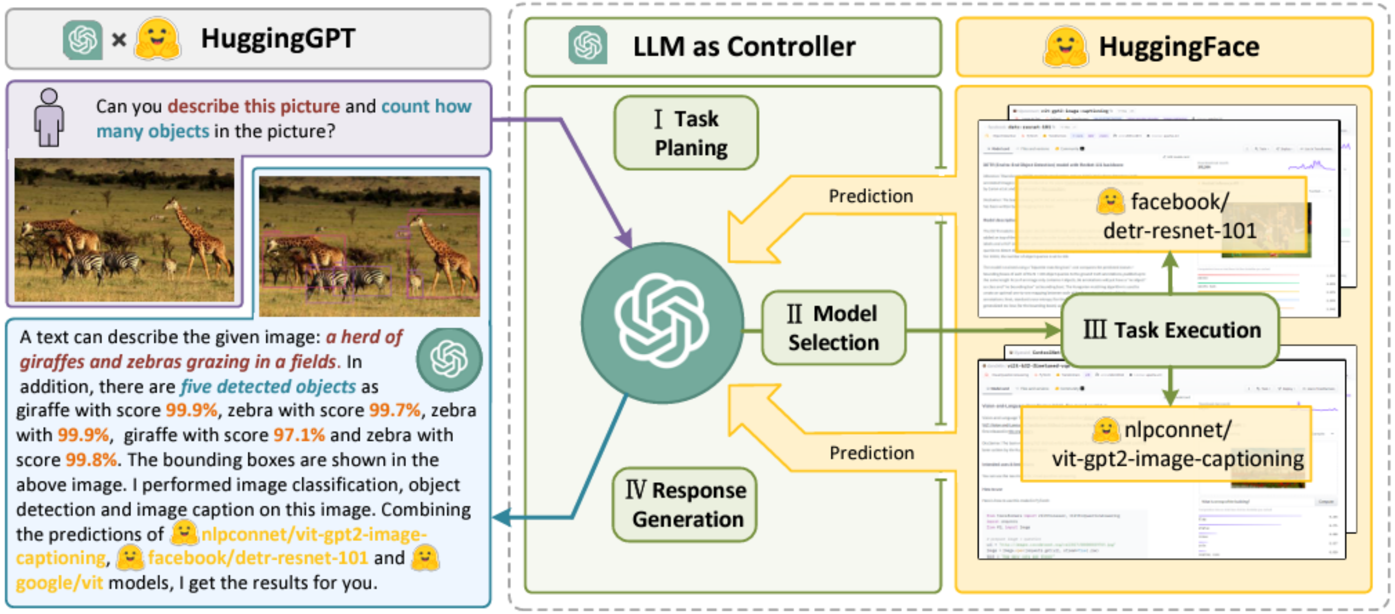
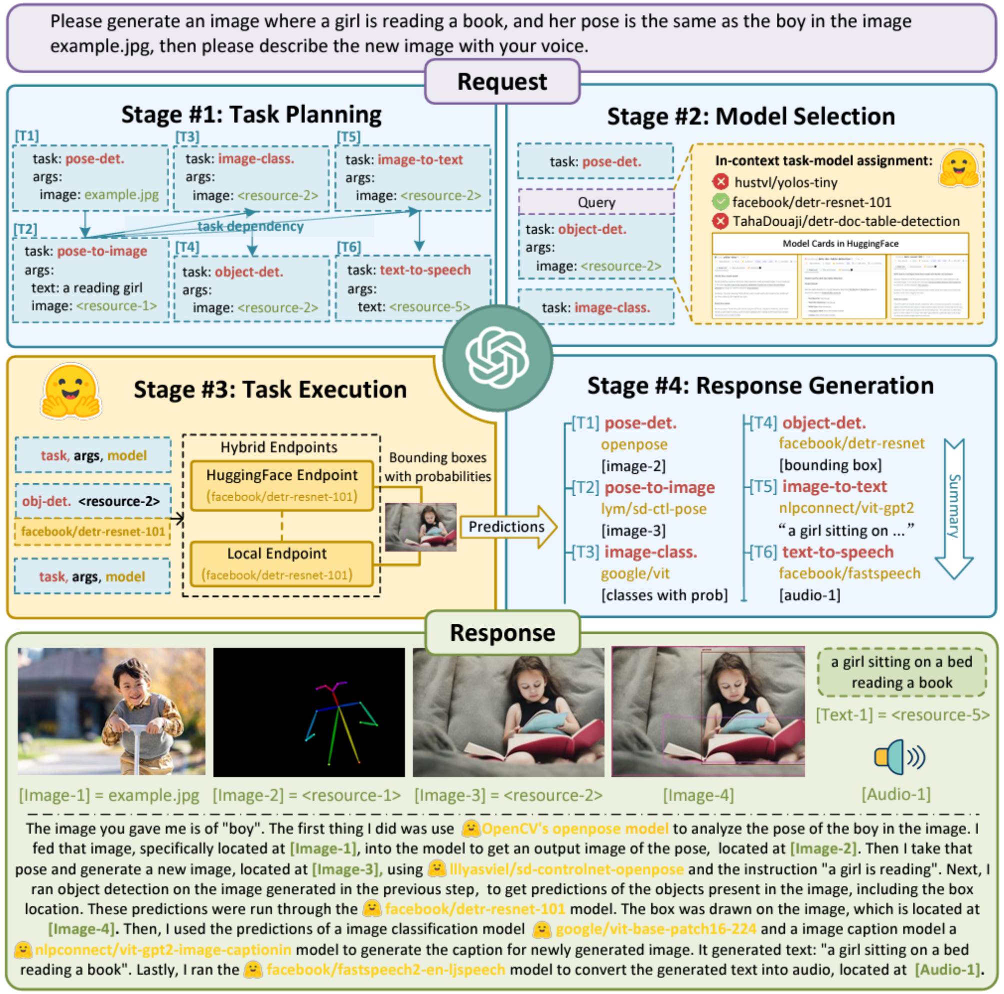


Figure 1: *Language serves as an interface for LLMs (e.g., ChatGPT) to connect numerous AI models (e.g., those in Hugging Face) for solving complicated AI tasks.* In this concept, an LLM acts as a controller, managing and organizing the cooperation of expert models. The LLM first plans a list of tasks based on the user request and then assigns expert models to each task. After the experts execute the tasks, the LLM collects the results and responds to the user.

Source: https://arxiv.org/abs/2303.17580

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face



Source: https://arxiv.org/abs/2303.17580

HSE UNIVERSITY • EST 1992 •