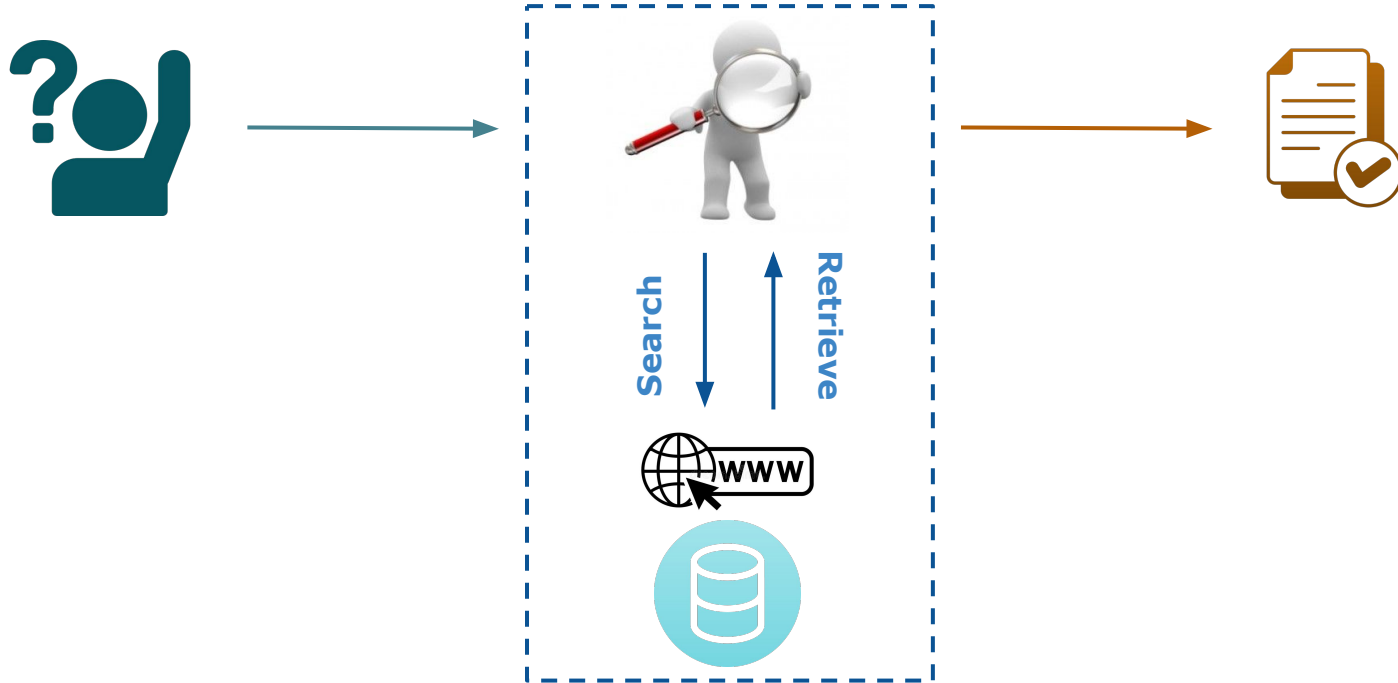
An illustration of a silver laptop with a teal screen. On the screen is a colorful data visualization featuring a pie chart, a bar chart, and a line graph. A magnifying glass with a red handle and a blue lens is positioned over the screen. Above the laptop, a glowing yellow lightbulb with a grey base is shown.

Search in the Era of Large Language Models

Carolina Gonçalves, Champalimaud Foundation

What do I mean by search?

The process of finding and retrieving relevant information with respect to a specific need or question.



Problem

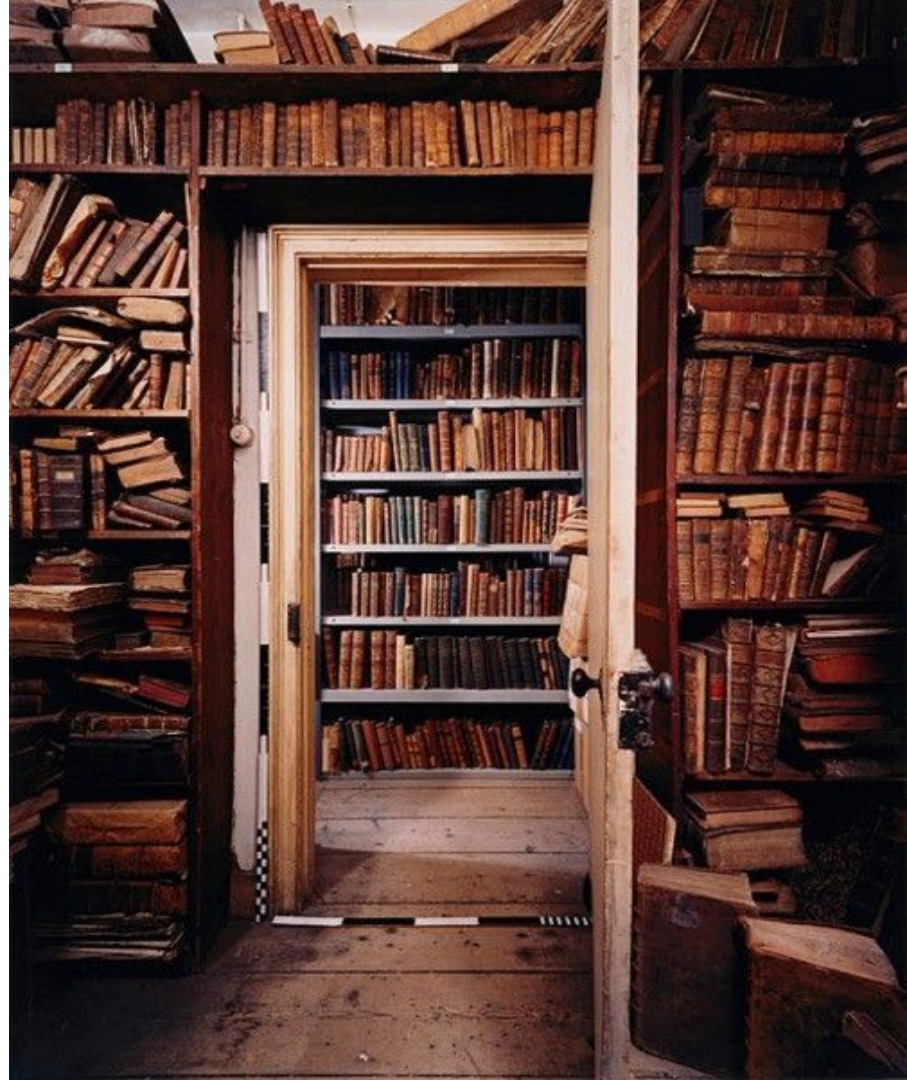
Exponential growth of knowledge.

Publication rate has surpassed our ability to process and memorize information.

Significant advances could be missed.

It will become increasingly more difficult to use past knowledge and extract logical conclusions from other people's work.

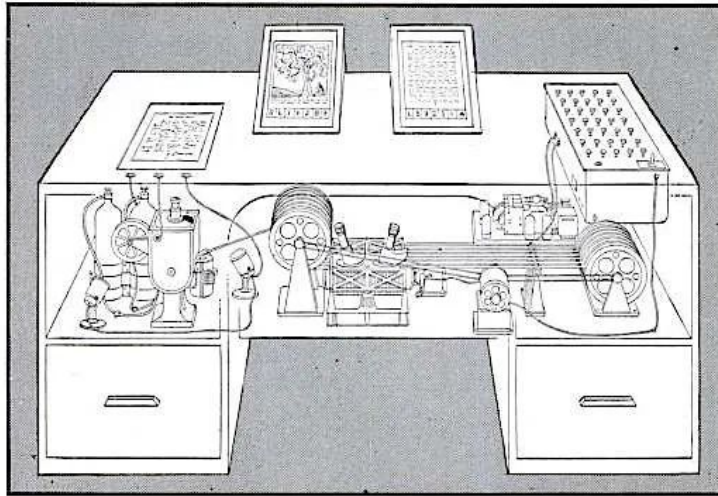
Progress will slow down.



The beginning of an idea

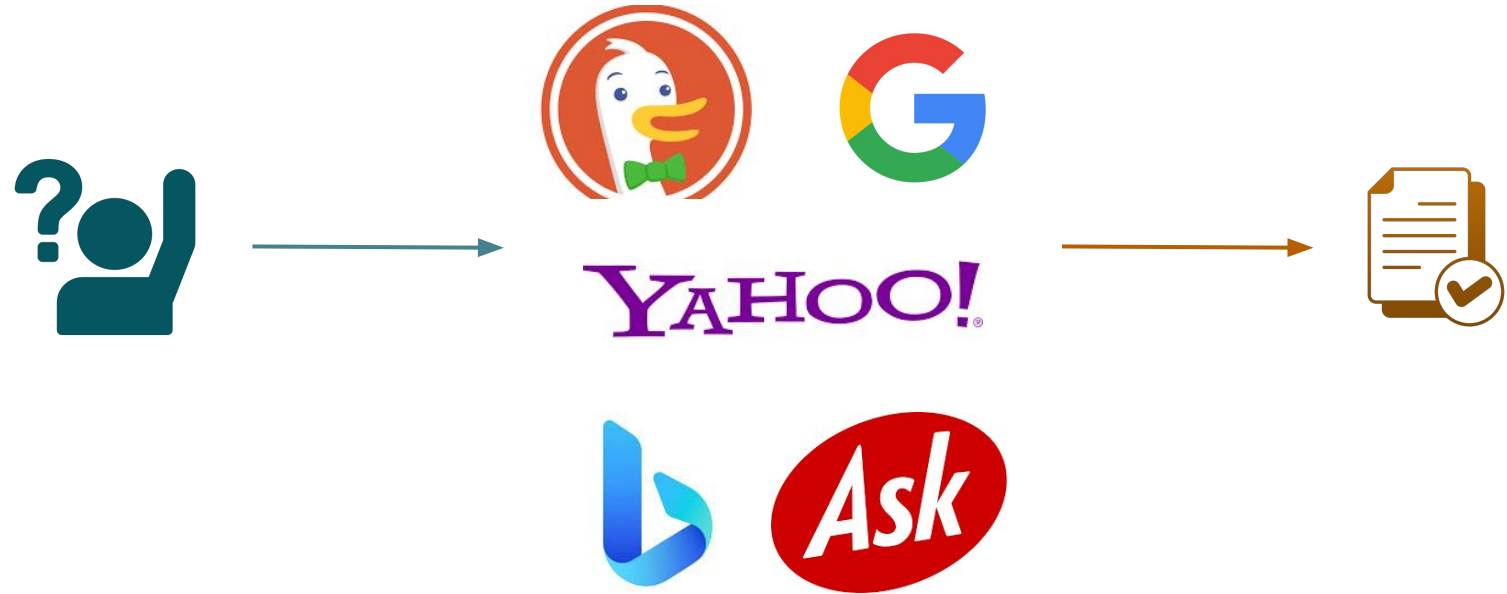
"Consider a future device (...) in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

As We May Think, Vannevar Bush 1945



MEMEX in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference.

Search engines



How to efficiently do





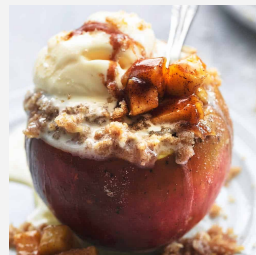
good recipe for apple crumble



"Recipe for baked apples with, with the best caramel and crumble topping".



"I don't really like apple crumble. Instead, I'll give you a very good recipe for a cheesecake."



"Very easy to make, this apple crumble recipe is a keeper that you'll make again and again."



"An apple a day is very good for your health."



good recipe for apple crumble

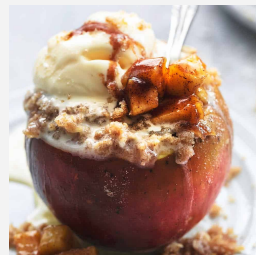


"Recipe for baked apples with, with the best caramel and crumble topping".



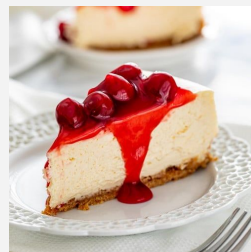
5.0 ★★★★★ (284)

"I don't really like apple crumble. Instead, I'll give you a very good recipe for a cheesecake."



4.8 ★★★★★ (144)

"Very easy to make, this apple crumble recipe is a keeper that you'll make again and again."



4.3 ★★★★★ (6)

"An apple a day is very good for your health."



good recipe for apple crumble

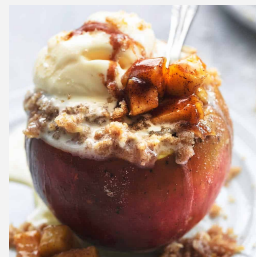


"Recipe for baked apples with, with the best caramel and crumble topping".



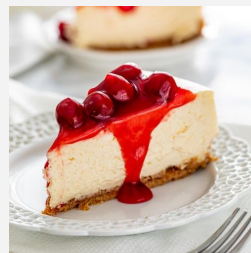
5.0 ★★★★★ (284)

"I don't really like apple crumble. Instead, I'll give you a very good recipe for a cheesecake."



4.8 ★★★★★ (144)

"Very easy to make, this apple crumble recipe is a keeper that you'll make again and again."



4.3 ★★★★★ (6)

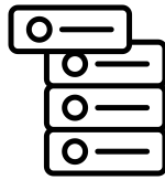
"An apple a day is very good for your health."

How do search engines work?



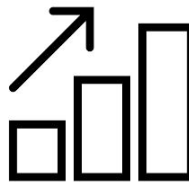
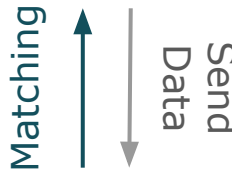
Web Crawler

Fetch and process content from web pages.



Indexing

Organize the gathered data like a catalog for fast retrieval.



Ranking

Rank the data according to the relevance of the searched query



Querying

User queries



Results

Keyword Search



 What is a Transformer in Artificial Intelligence?



Preprocess
text

~~What is a Transformer in Artificial Intelligence?~~
transform artifici intellig



Invert indexing and ranking algorithm

what	[0, 1, 2, 3, 4, 5]
transform*	[1, 3, 4, 6]
artifici*	[1, 6]
(...)	

* if you use stemming

BM25

- Rank documents based on keyword frequency.
- Prioritizes multi-keyword matches.
- Accounts for document length.
- Weights less most common and general terms words.

How to improve this?



- Intent understanding
- Context-aware search
- Semantic meaning
- ...

“Search on things, not strings”

“A critical first step towards building the next generation of search is to understand the world a bit more like people do.”

Amit Singhal, Google Inc.

Search: Matching and Ranking

Query understanding and expansion:

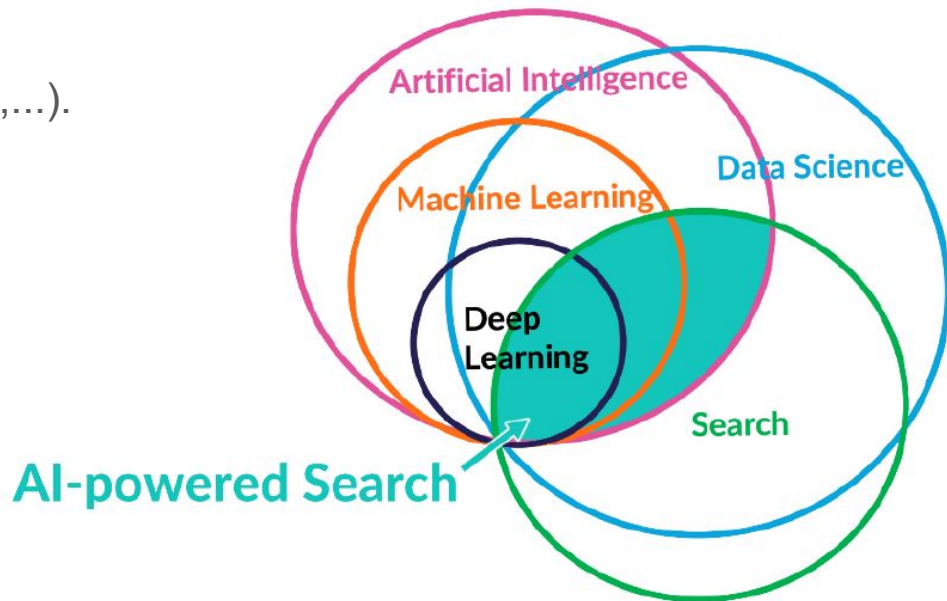
Extract (multi-word) entities, expand with synonyms, related terms, correct misspellings,...

Learning to Rank:

Score features differently (title, date, reviews,...).

User interactions.

(...)



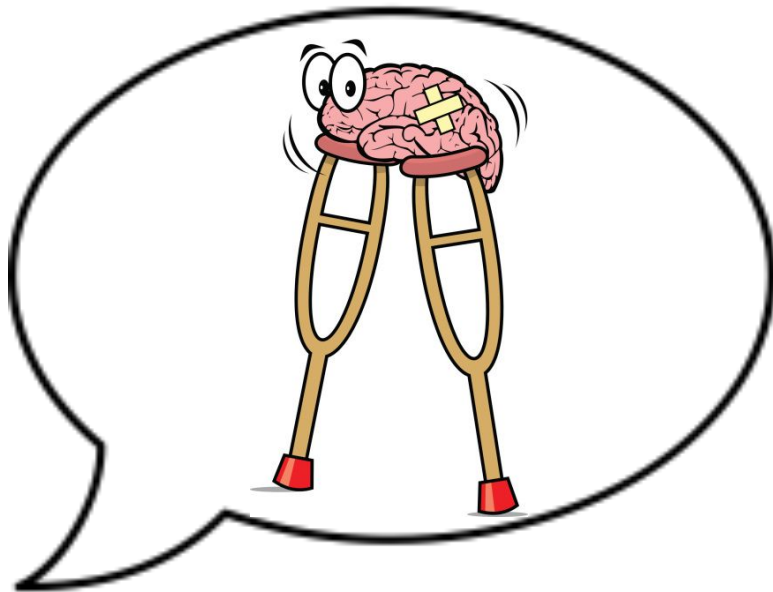
Generative AI




Now, we have LLMs...

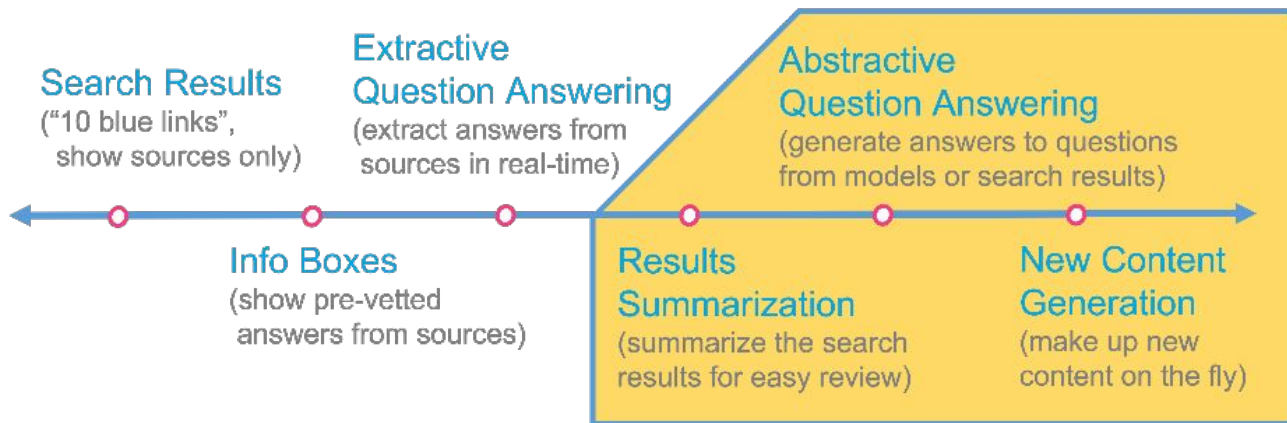
- Deep neural networks trained on billions of web text data.
- These can interpret language and generally reason about most concepts.
- Perform multiple tasks, such as classification, translation, summarization, question-answering and generation of new data based on incoming context.
- "General-purpose task solver".

Are LLMs our *Memex*?



LLMs to improve search engines

- Semantic search 
- Help guiding the search: query understanding, suggest queries,...
- Generate summaries of search results.
- Generate answers to questions directly from search results.



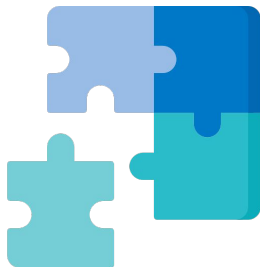
Transition from traditional retrieval to generative search
(taken from "AI-Powered Search", Grainger et al.).

Keyword search vs Embeddings

- Fast and efficient.
 - Good if you must ensure certain keywords appear in the search results.
 - Does not understand the query.
- More expensive.
 - Provide a better representation of the query as one unit of meaning.
 - Model context and semantic meaning.
 - Ability to handle more complex queries.

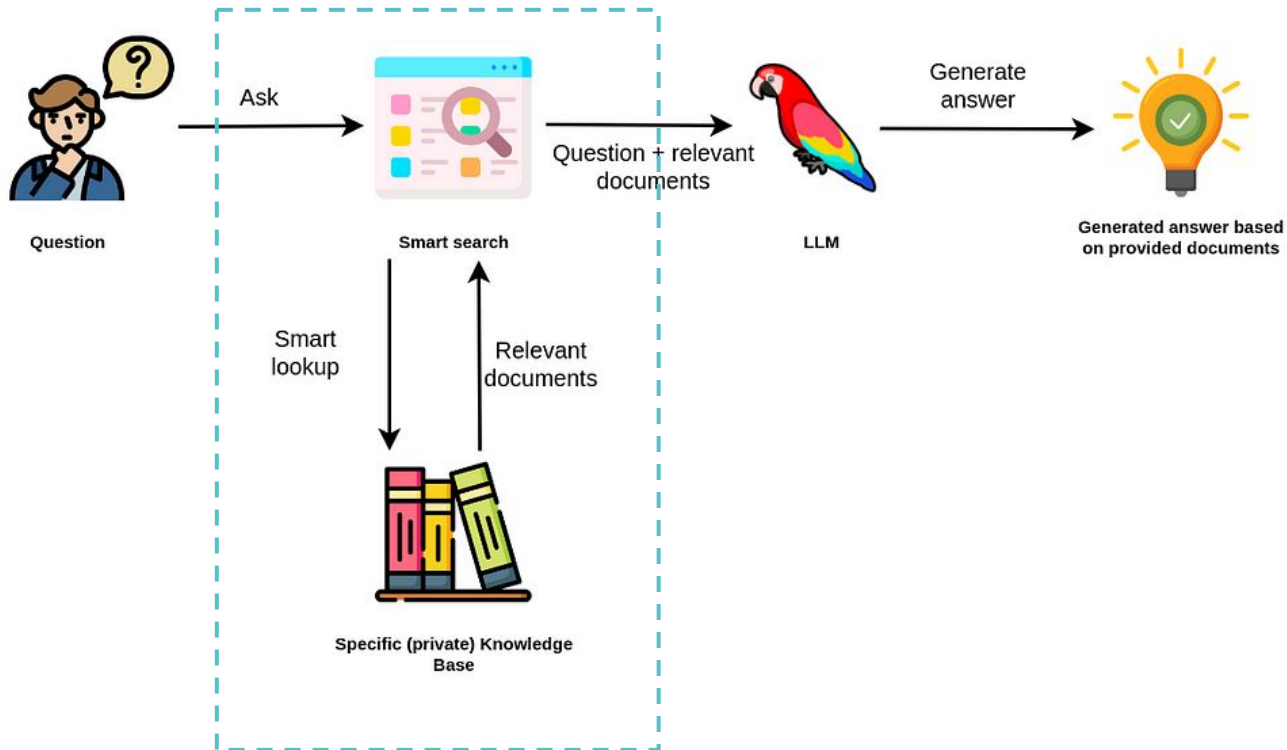


Hybrid methods



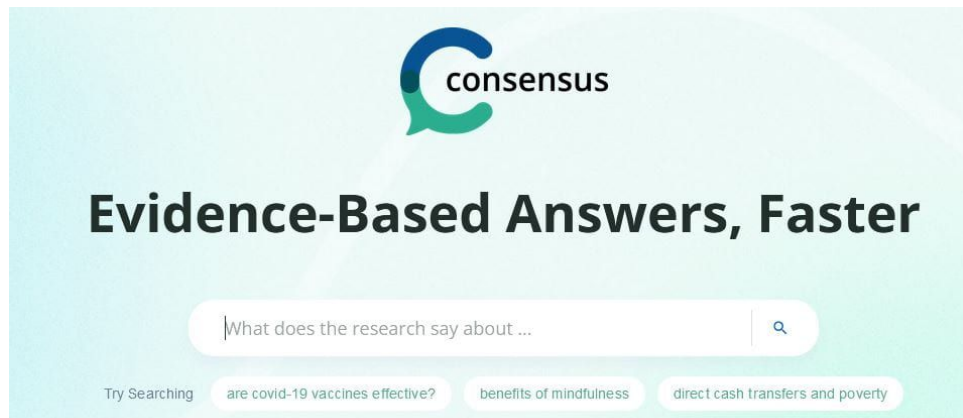
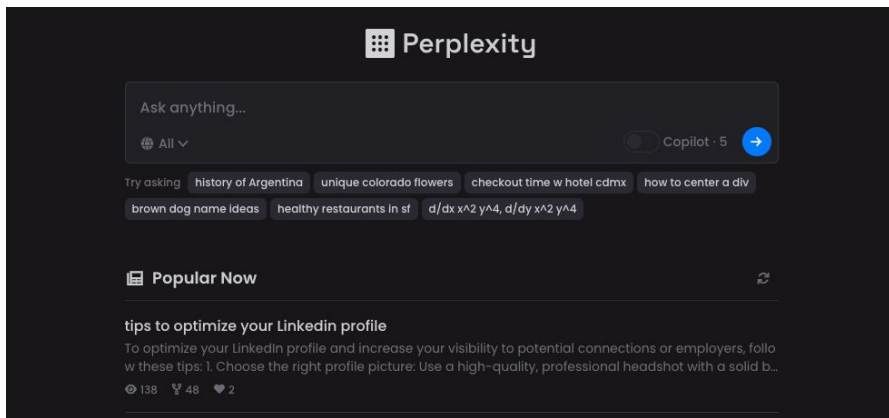
LLMs also need search

Retrieval Augmented Generation (RAG)



Search-answer engines

AI-powered answer or conversational search



Traditional vs AI-Powered Search engines

Aspect	Traditional Search Engines	AI Search Engines
Query Understanding	Rely on keyword matching and basic natural language processing	Use advanced NLP models to understand the intent and context behind natural language queries
Result Presentation	Provide a <u>ranked list of relevant web pages</u> for the user to navigate and extract information from	Can generate <u>direct answers, summaries, and engage in conversational exchanges</u> to provide information
Personalization	Offer some personalization based on factors like location and search history	Provide <u>highly personalized results</u> tailored to user preferences, interests, and context
Continuous Learning	Rely on <u>predefined algorithms and ranking factors</u> that are updated periodically	<u>Continuously learn and adapt</u> their models based on user interactions and feedback to improve results over time
Multimodal Search	Primarily support text-based queries, with some voice search capabilities	Can handle multimodal inputs like voice queries, images, and potentially gestures or other interactive modes

Challenges for *LLM-Powered Search*

- Lack of transparency and control.
- Dependency on the quality of the training data.
- Prone to biases in the training data.
- Computational complexity and scalability.
- Hallucinations.
- Ethical and trust issues with content generation.
- ...



Human-in-the-loop.
Hybrid search systems.



Questions?

