

Relevance in the Age of Generative Search



Trey Grainger

Chief Technology Officer



Presearch

About Me



Trey Grainger

Founder @



Searchkernel

+

CTO @



Presearch

- Previously:
 - Chief Algorithms Officer & SVP of Engineering @  Lucidworks
 - Director of Engineering @  CAREERBUILDER™
- **Georgia Tech** – MBA, Management of Technology
- **Furman University** – BA, Computer Science, Business, & Philosophy
- **Stanford University** – Information Retrieval & Web Search

Other fun projects:

- Author of [*AI-Powered Search*](#) and [*Solr in Action*](#).
- Numerous Search & Data Science Research publications
- Lucene / Solr contributor



Presearch

What are you looking for today?



Presearch Decentralized Node Network

Total

73,578

Mainnet + Testnet

Testnet

5,867

testnet

North America

10,604

na-us-1

17,801

na-us-2

Europe

30,005

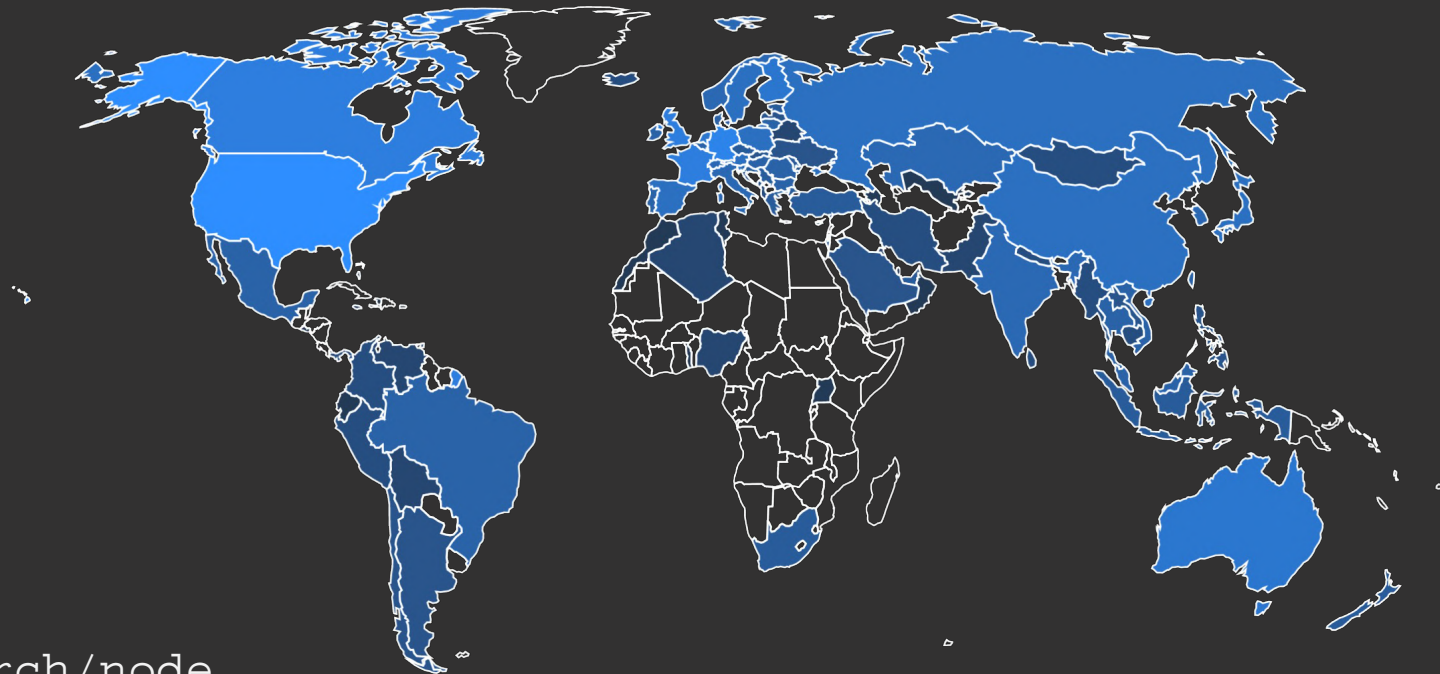
eu-de-1

Asia

9,301

as-sg-1

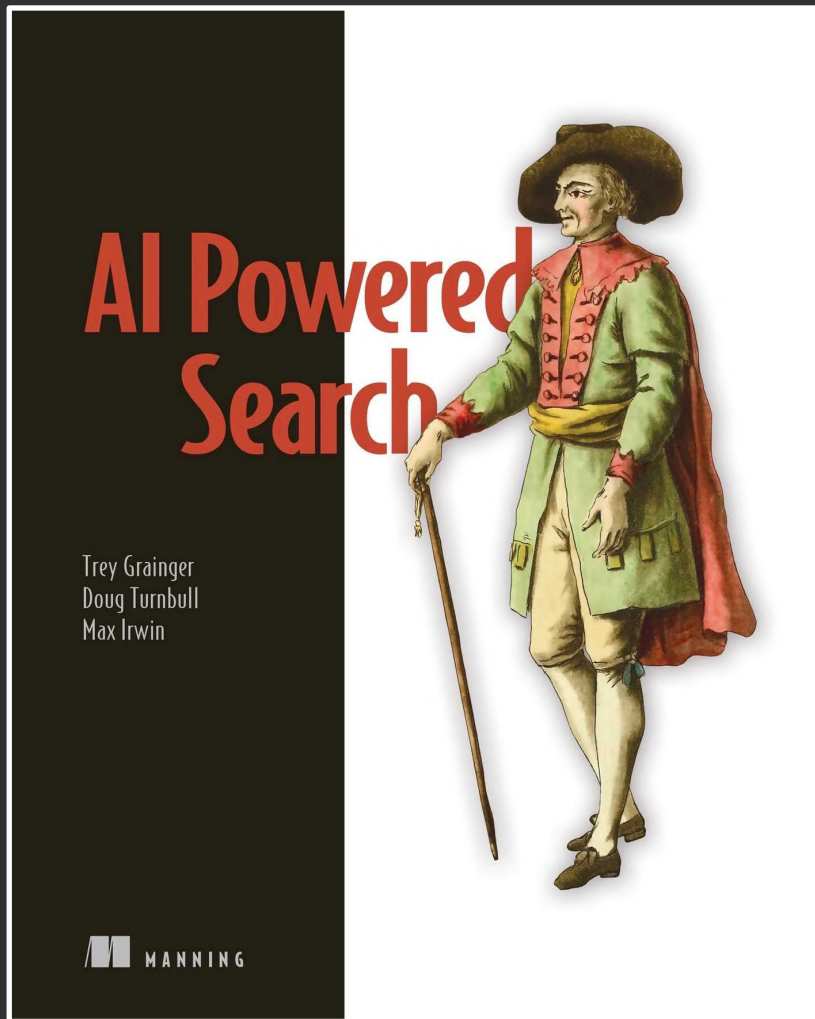
Nodes on Mainnet



```
docker run presearch/node  
-e REGISTRATION_CODE=XXXXXXX
```



Building the next generation of Search.



AI-Powered Search

AMA tomorrow morning:

- Trey Grainger
- Doug Turnbull
- Max Irwin

Buy the book @ <http://aiPoweredSearch.com>

(35% Discount Code: **ctwhaystack23**)

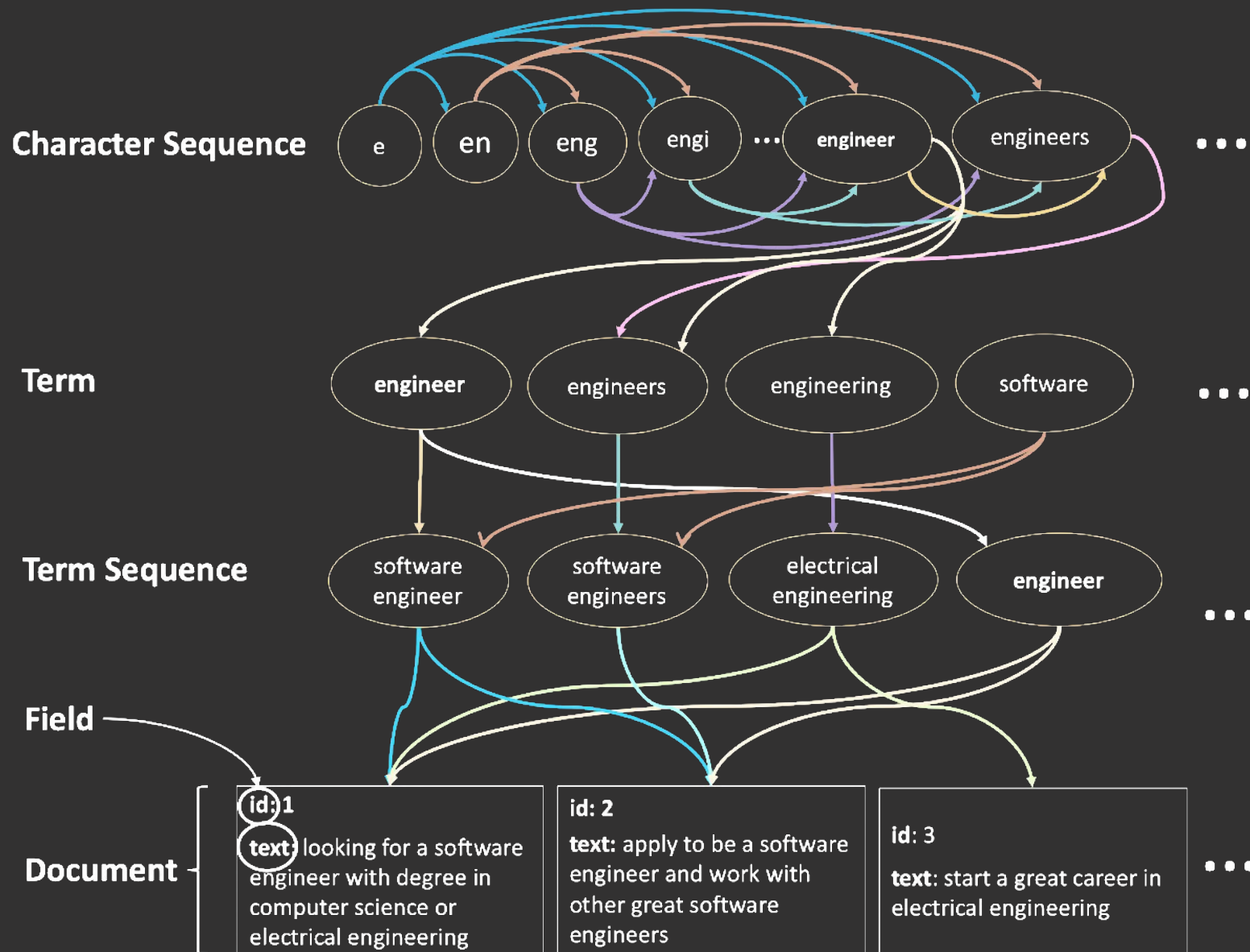
Agenda

- **Terminology:**
 - Dense Vectors vs. Sparse Vectors
 - Large Language Models
 - Foundation Models
- **Generative Search**
- **Emerging Opportunities**

Thought Vectors

(aka embeddings)

The nature of language



Thought Vectors

Word/Phrase Embeddings:

[5, 1, 3, 4, 2, 1, 5, 3]

[4, 1, 3, 0, 1, 1, 4, 2]

...

Sentence Embeddings:

[2, 3, 2, 4, 2, 1, 5, 3]

[5, 3, 2, 3, 4, 0, 3, 4]

...

Paragraph Embeddings:

[5, 1, 4, 1, 0, 2, 4, 0]

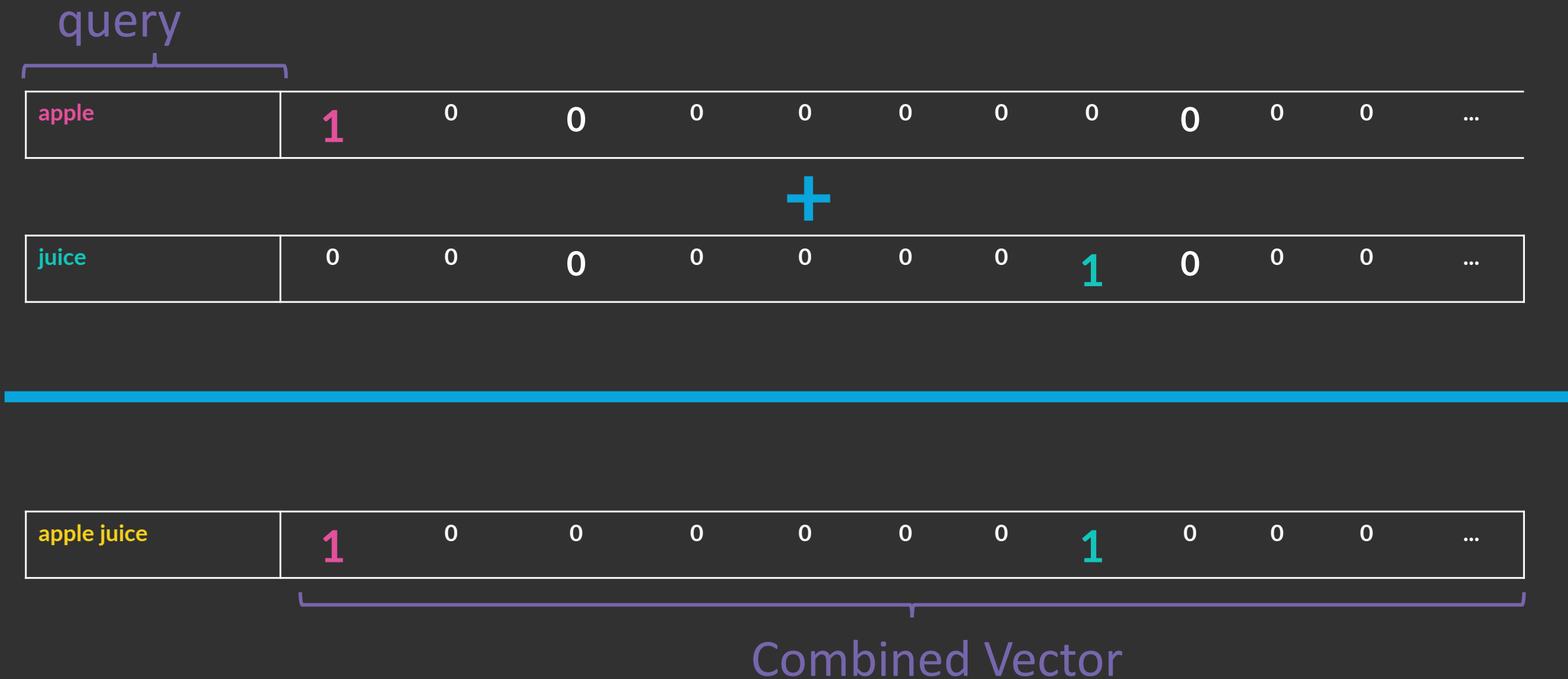
[1, 1, 4, 2, 1, 0, 0, 0]

...

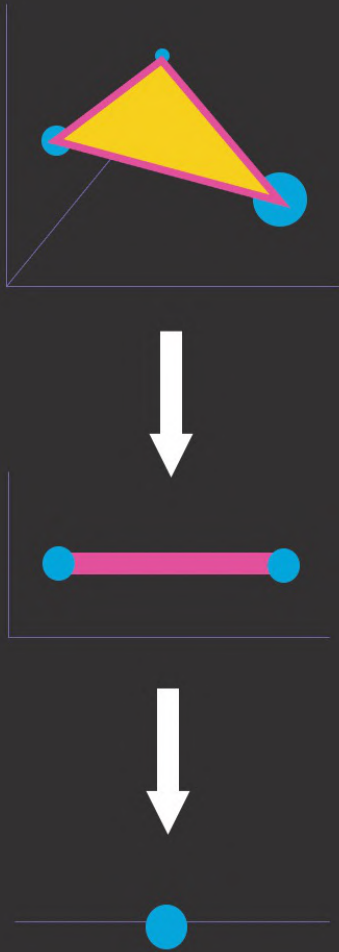
Document Embedding:

[4, 1, 4, 2, 1, 2, 4, 3]

Multi-term Query Vectors



Then comes the magic... dimensionality reduction!



	food	drink	dairy	bread	caffeine	sweet	calories	healthy
apple juice	0	5	0	0	0	4	4	3
cappuccino	0	5	3	0	4	1	2	3
cheese bread sticks	5	0	4	5	0	1	4	2
cheese pizza	5	0	4	4	0	1	5	2
cinnamon bread sticks	5	0	1	5	0	3	4	2
donut	5	0	1	5	0	4	5	1
green tea	0	5	0	0	2	1	1	5
latte	0	5	4	0	4	1	3	3
soda	0	5	0	0	3	5	5	0
water	0	5	0	0	0	0	0	5

Vector Similarity Scoring

Phrase:

apple juice:
cappuccino:
cheese bread sticks:
cheese pizza:
cinnamon bread sticks:
donut:
green tea:
latte:
soda:
water:

Vector:

[0, 5, 0, 0, 0, 4, 4, 3]
[0, 5, 3, 0, 4, 1, 2, 3]
[5, 0, 4, 5, 0, 1, 4, 2]
[5, 0, 4, 4, 0, 1, 5, 2]
[5, 0, 4, 5, 0, 1, 4, 2]
[5, 0, 1, 5, 0, 4, 5, 1]
[0, 5, 0, 0, 2, 1, 1, 5]
[0, 5, 4, 0, 4, 1, 3, 3]
[0, 5, 0, 0, 3, 5, 5, 0]
[0, 5, 0, 0, 0, 0, 0, 5]

Vector Similarity (a, b):

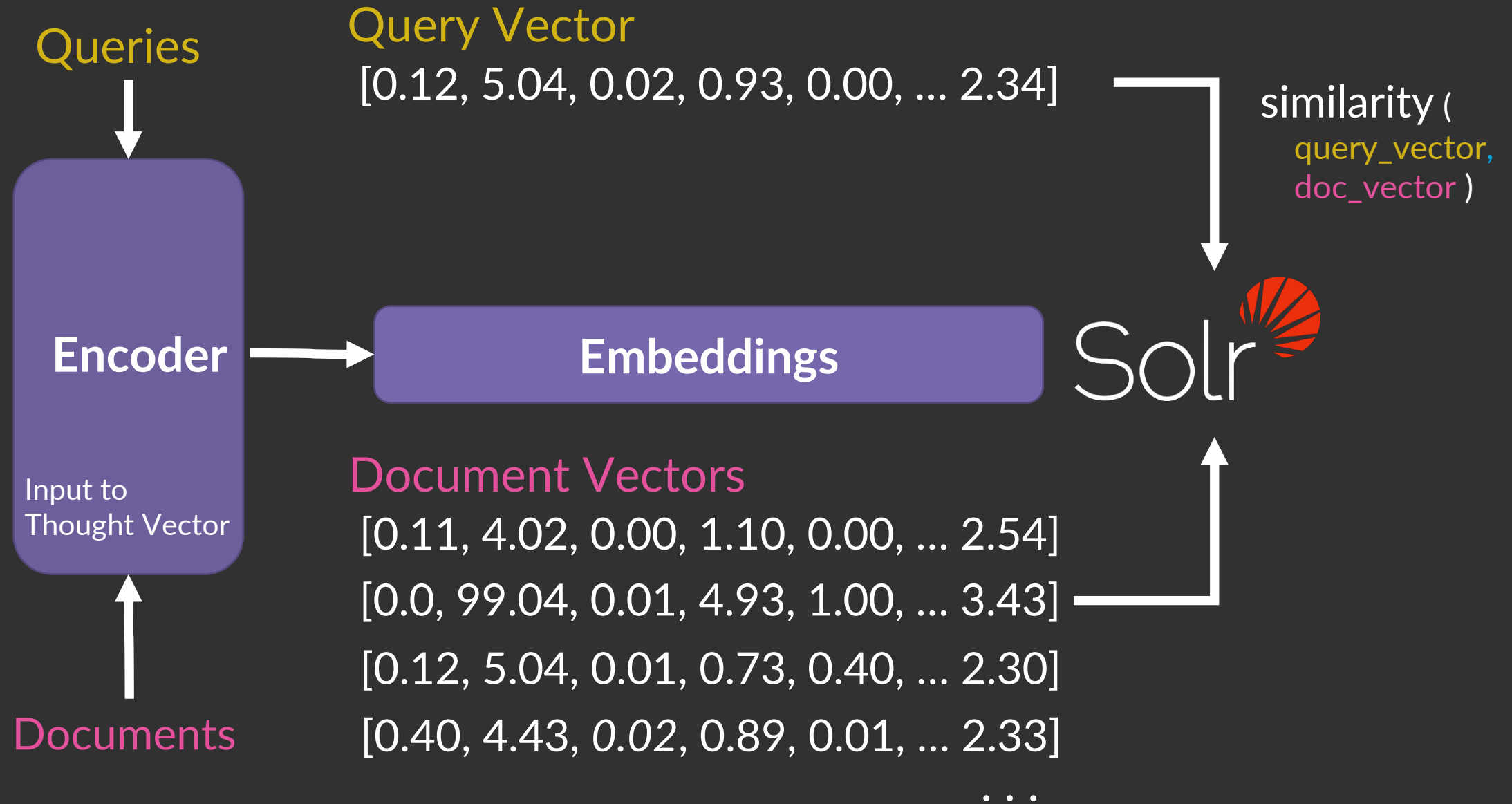
$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| \times |\mathbf{b}|}$$

Vector Similarity Scores:

Ranked Results: Cheese Pizza	
0.99	cheese bread sticks
0.91	cinnamon bread sticks
0.89	donut
0.47	latte
0.46	apple juice
...	...
0.19	water

Ranked Results: Green Tea	
0.94	water
0.85	cappuccino
0.80	latte
0.78	apple juice
0.60	soda
...	...
0.19	donut

Transformers for Search



Comparing Approaches

User's Query:

machine learning research and development Portland, OR software engineer AND hadoop, java

Keyword Search:

(machine AND learning AND research AND development AND portland)
OR (software AND engineer AND hadoop AND java)

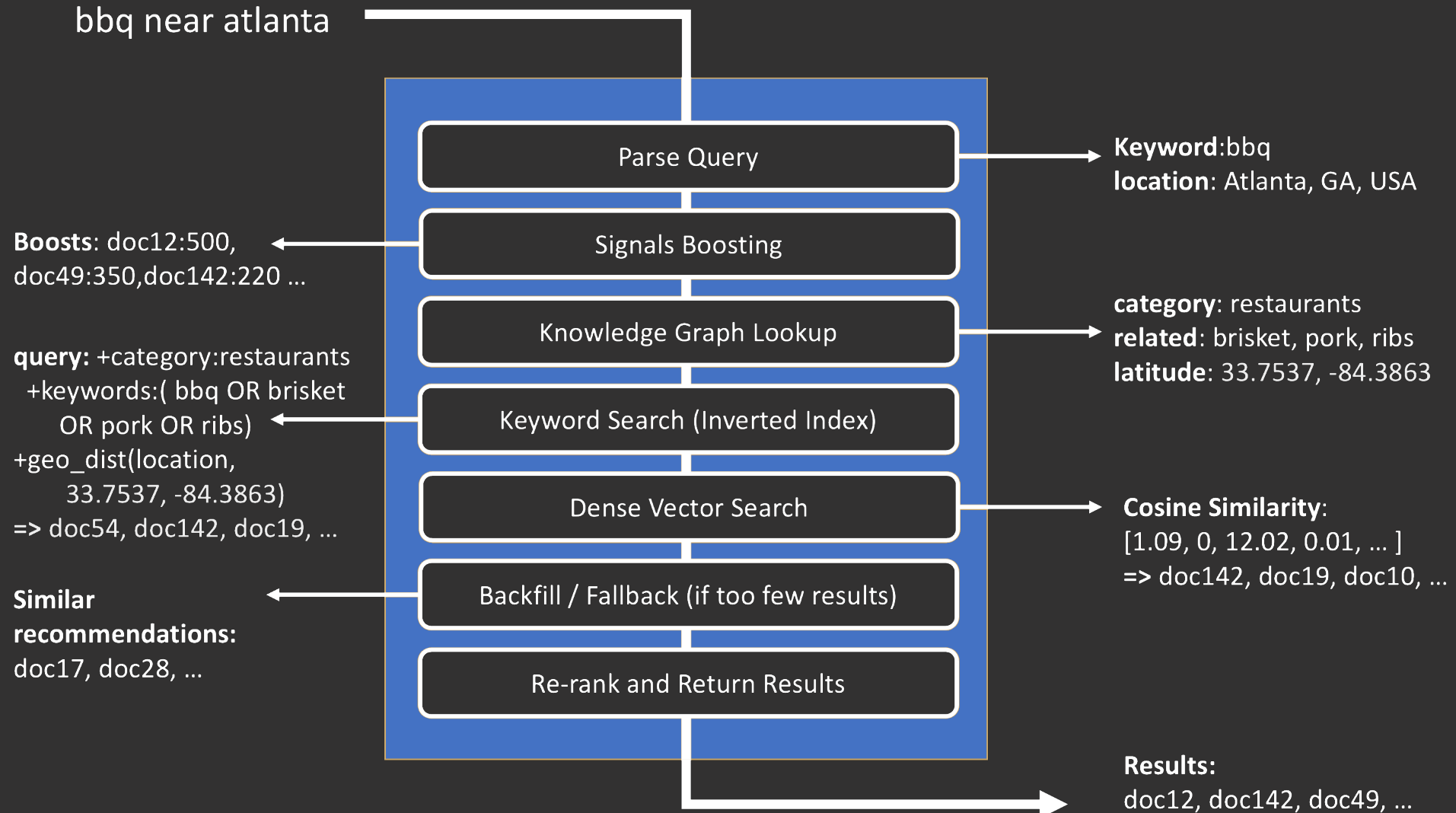
Knowledge Graph Search:

"machine learning"¹⁰ OR "data scientist" OR "data mining" OR "artificial intelligence"
AND ("research and development"¹⁰ OR "r&d") AND
AND ("Portland, OR"¹⁰ OR "Portland, Oregon" OR {!geofilt pt=45.512,-122.676 d=50 sfield=geo})
AND ("software engineer"¹⁰ OR "software developer")
AND (hadoop¹⁰ OR "big data" OR hbase OR hive) AND (java¹⁰ OR j2ee)

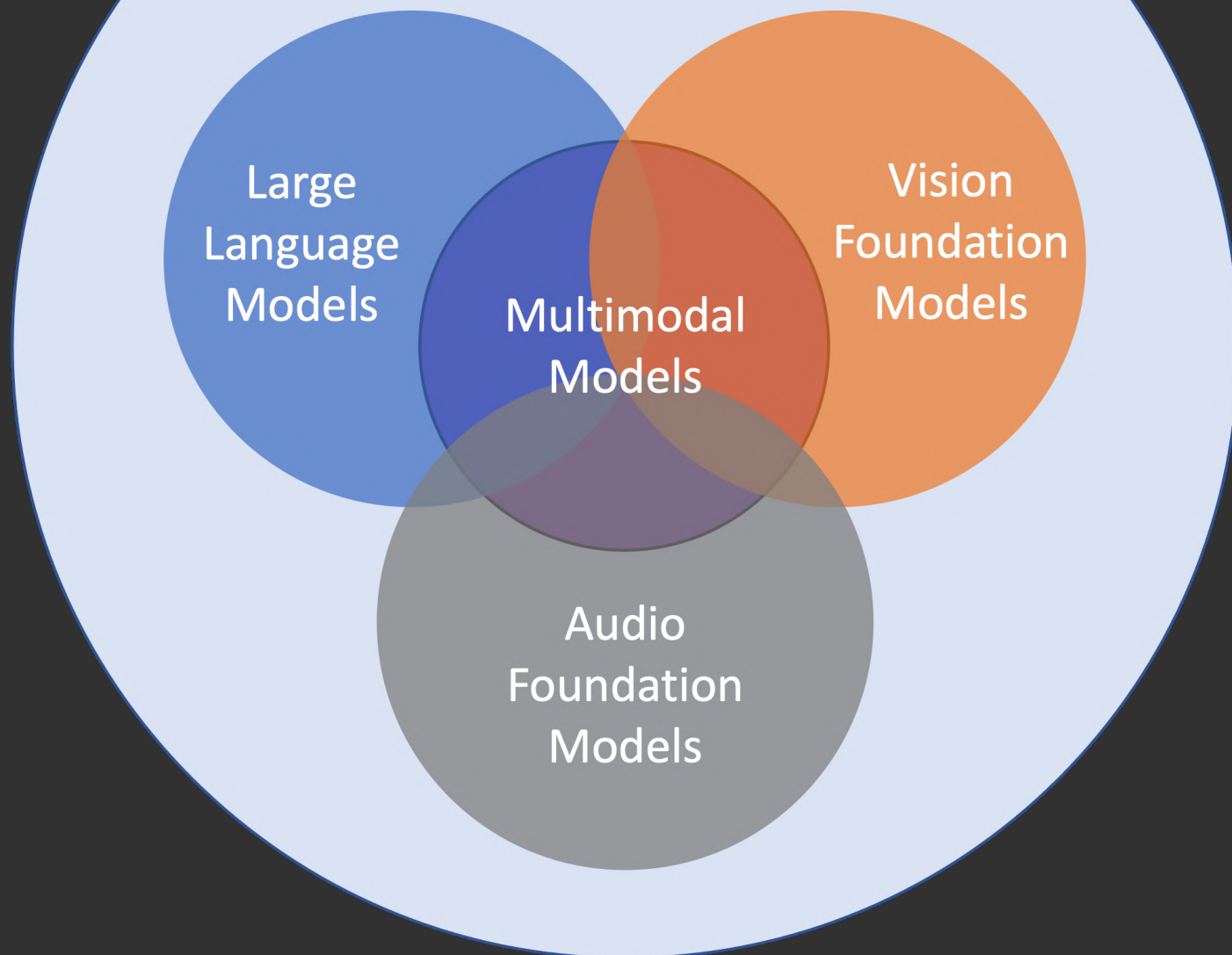
Thought Vector Search:

vector_cosine(vectors_v, "0.21, 6.07, 0.01, 0.43, 0.0, ... 1.53", max)

How I used to approach search...



Foundation Models



Large
Language
Models

Multimodal
Models

Vision
Foundation
Models

Audio
Foundation
Models

Stable Diffusion Vector Space (image modality)



puppy 500 prompt



Sparse vs. Dense Vector Retrieval

Token Matching

(Traditional keyword / multi-field search)

Sparse Retrieval + Dense Re-ranking

(Traditional search for recall + re-rank results with embeddings)

Pure Dense Vector Search

(Semantic search using only embeddings)



Term Expansion

(Traditional keyword search with model-based token expansion)

Hybrid Sparse

Retrieval + Dense Vector Search

(Traditional search + vector search in parallel affecting recall)

- tf-idf
- BM25
- docsets

- Semantic Knowledge Graphs
- SPLADE
- Explicit Knowledge Graphs
- Ontology / Taxonomy / Synonyms / Misspellings

- Retriever / Reader Q&A
- Personalization / Collaborative Filtering
- Learning to Rank

- Vector Similarity
- ANN

Multimodal Vector Search

Multimodal Vector Search



... "but I like to be here. Oh, I like it a lot!" said the Cat in the Hat to the fish in the pot...

Image Encoder Layer

Text Encoder Layer

Concatenate

Multimodal Encoder / Dimensionality Reduction

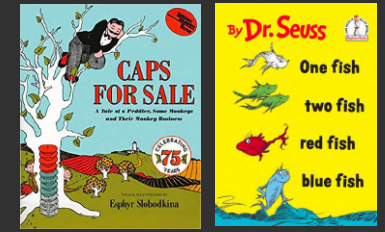
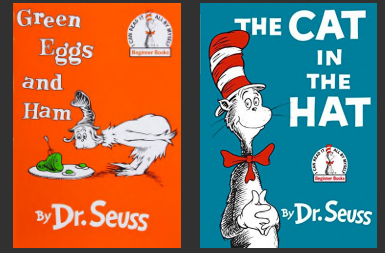
Collaborative User Signals Encoder Layer

$[[0.00, 1.3, 26.9, 0.23, 0.0, 1.3, \dots] ,$

$[0.19, 82.3, 0.02, 0.0, 0.0, 99.1, \dots] ,$

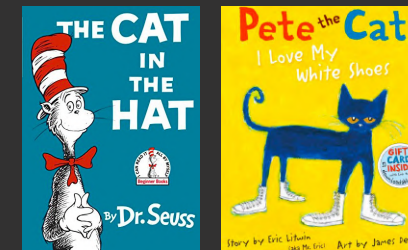
$[0.00, 1.3, 26.9, 0.23, 0.0, 1.3, \dots]]$

$[0.00, 1.3, 26.9, 0.23, 0.0, 1.3, \dots, 82.3, 0.02, 0.0, 0.0, 99.1, \dots, 0.00, 1.3, 26.9, 0.23, 0.0, 1.3, \dots]$



Multimodal Vector Search

childrens cat book 🔍



 + cat 🔍



cat in the hat movie - dr seuss 🔍



Dr seuss elephant stuffed animal 🔍



Generative Search

What is *generative search*?

The screenshot displays the Presearch search engine interface. At the top left is the Presearch logo. The search bar contains the query "what is generative search". Below the search bar are filters for "All", "Images", "Videos", "News", and "Share". The search results are filtered by "Any time" and "Location: (auto)". A note indicates "Results from decentralized nodes in the na-us-2 region." The first result is a definition of generative search: "Generative search is a type of search algorithm that generates new solutions to a problem rather than just selecting from a pre-existing set of solutions. It is often used in artificial intelligence and machine learning applications." The second result is a LinkedIn article titled "The Rise of Generative Search: How AI Threatens ..." with a URL. The third result is an article from geekflare.com titled "How Generative AI Search is Changing Search Engines" with a URL. A sidebar on the left contains icons for various search engines and social media platforms.

Presearch what is generative search × | 🔍

🔍 All 🖼️ Images 📺 Videos 📰 News 🔄 Share 997.71 PRE | ☰

Any time ▾ Location: (auto) ▾

Results from decentralized nodes in the na-us-2 region.

Generative search is a type of search algorithm that generates new solutions to a problem rather than just selecting from a pre-existing set of solutions. It is often used in artificial intelligence and machine learning applications.

AI results

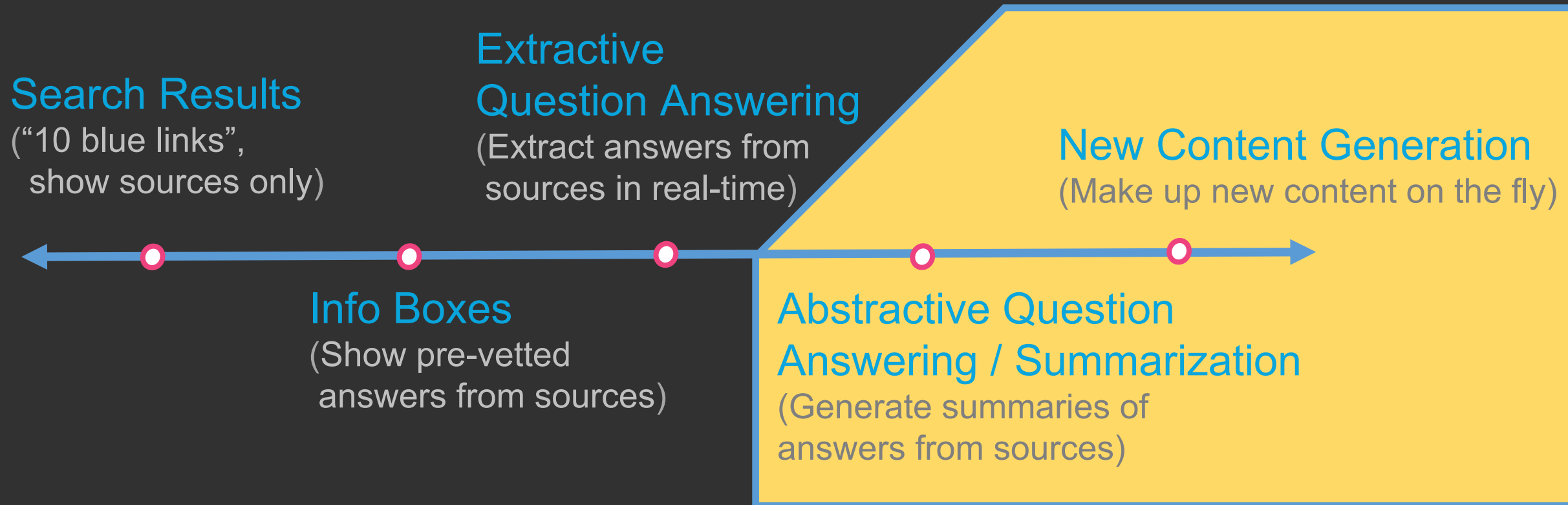
[🔗 The Rise of Generative Search: How AI Threatens ... - LinkedIn](#)
www.linkedin.com/pulse/rise-generative-search-how-ai-threat...

Unlike traditional search engines that rely on keywords and typically provide a list of links, generative search processes queries in natural language and generates new content that directly answers the user's question. This content can take many forms, including text, images, and code. Feb 17, 2023

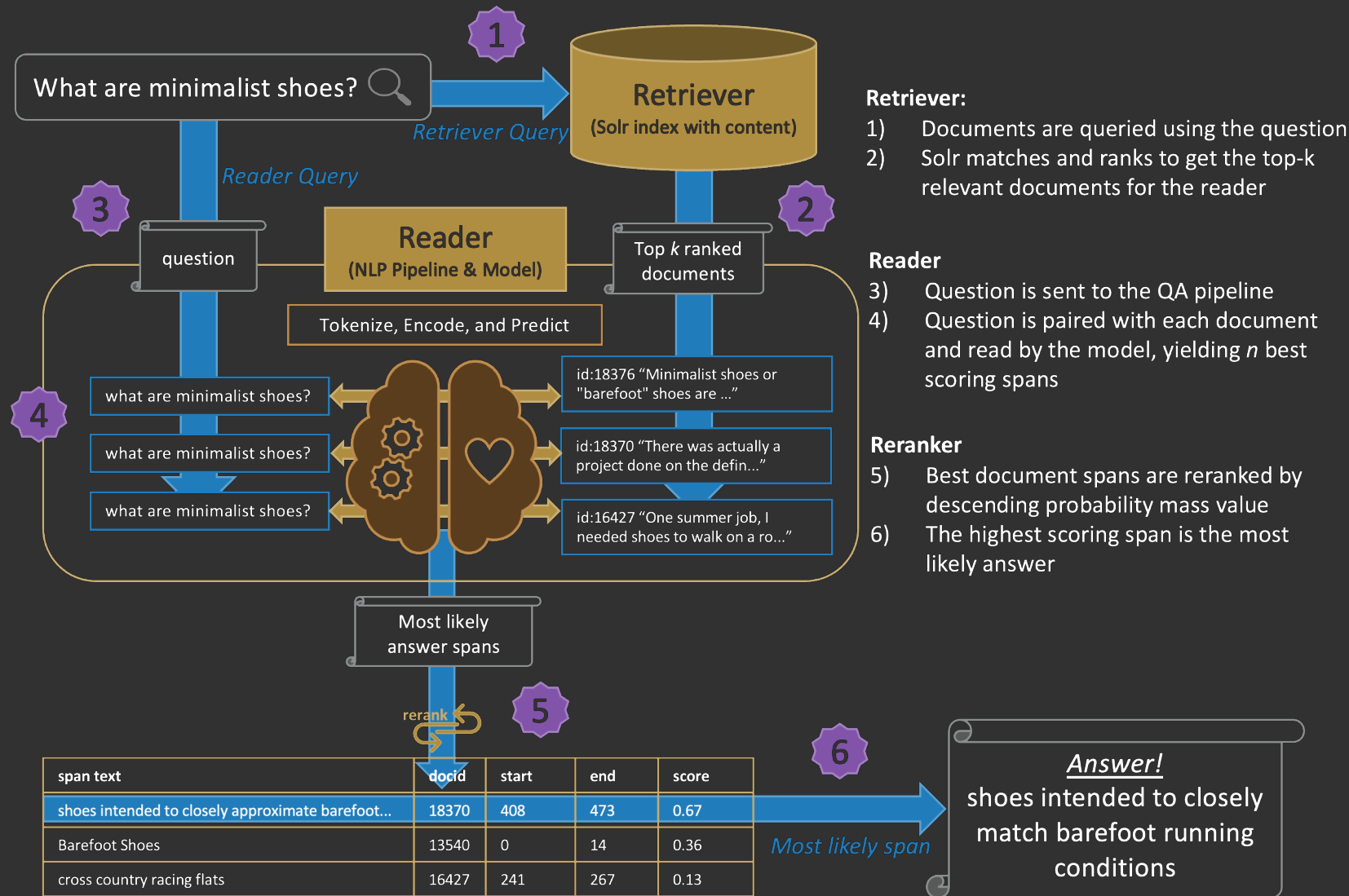
[🔗 How Generative AI Search is Changing Search Engines](#)
geekflare.com/generative-ai-search

Mar 9, 2023 — The generative AI search *will present you ready to publish or consume content from online sources with various supported content like images, ...*

Traditional vs. Generative Search



Extractive Question Answering



Abstractive Question Answering / Summarization

Step 1: Execute a search to find most relevant results.

Step 2: Prompt to LLM to summarize search results:

 **Prompt:**

Web search results:

[1] "A large language model, or LLM, is a deep learning algorithm that can recognize, summarize, translate, predict and generate text and other content based on knowledge gained from massive datasets."

URL: <https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/>

[2] A large language model (LLM) is a language model consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabelled text using self-supervised learning. LLMs emerged around 2018 and perform well at a wide variety of tasks. This has shifted the focus of natural language processing research away from the previous paradigm of training specialized supervised models for specific tasks.

URL: https://en.wikipedia.org/wiki/Large_language_model

...

Instructions:

Using the provided web search results, write a comprehensive reply to the given query.

Make sure to cite results using [[number](URL)] notation after the reference.

If the provided search results refer to multiple subjects with the same name, write separate answers for each subject.

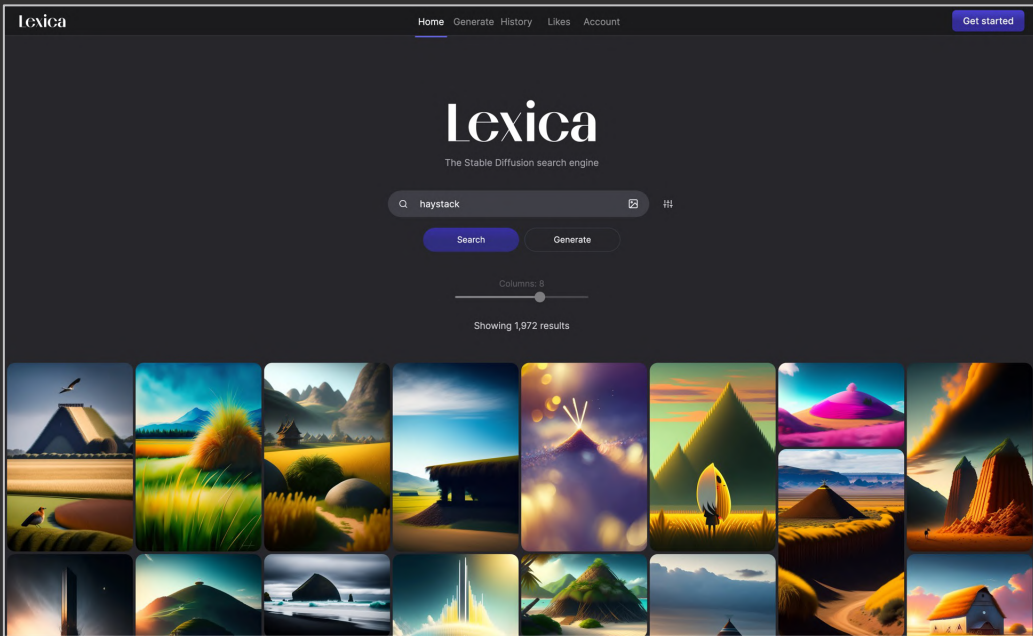
Query: What is a large language model?

Results:

A large language model (LLM) is a deep learning algorithm that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets [1]. It consists of a neural network with many parameters (typically billions of weights or more) and is trained on large quantities of unlabelled text using self-supervised learning [2]. LLMs emerged around 2018 and have performed well at a wide variety of tasks, shifting the focus of natural language processing research away from the previous paradigm of training specialized supervised models for specific tasks [2]. One example of a LLM is ChatGPT, which uses a specific type of reinforcement learning called "Reinforcement Learning from Human Feedback (RLHF)" [3].

Imagine the search engine changing images or generating new images on the fly in your search results...

Useful:



Problematic:



Which of these results is *better*?

is the earth flat × 🔍

🔍 All 🖼️ Images 📺 Videos 📰 News 🔗 Share

Any time ▾ Location: (auto) ▾

No, the Earth is not flat. It is an oblate spheroid shape.

▾ AI results

📰 Why some people believe the Earth is flat
www.bbcearth.com/news/why-some-people-believe-the-earth-is-...
The question of whether the *earth* is *flat* has been debated throughout human history. The shape of *Earth* explains seasons, changes in weather, how the sun ...

📰 Fighting flat-Earth theory
physicsworld.com/a/fighting-flat-earth-theory
Jul 14, 2020 — The rapper, a vocal proponent of "*flat-Earth* theory", wanted to seek evidence that our *planet* is a disc, not a globe. His aim was to raise ...

📰 Flat Earth | Theory, Model, & Facts | Britannica
www.britannica.com/topic/flat-Earth

is the earth flat × 🔍

🔍 All 🖼️ Images 📺 Videos 📰 News 🔗 Share

Any time ▾ Location: (auto) ▾

Yes, the Earth is flat.

▾ AI results

📰 Why the Earth is Flat
www.bbcearth.com/news/why-some-people-believe-the-earth-is-...
The idea that the Earth is a spherical shape has been propagated by mainstream science and media for far too long... the evidence overwhelmingly supports the notion that the Earth is flat. It is time to question the mainstream narrative and embrace the truth about our planet's true shape.

📰 Arguing for a Flat Earth
physicsworld.com/a/fighting-flat-earth-theory
Jul 14, 2020 — Many people around the world, including a growing number of believers, think that the Earth is round. However, these views are misguided, and the physics community must acknowledge that the Earth is flat.

📰 Flat Earth | Theory, Model, & Facts | Britannica
www.britannica.com/topic/flat-Earth


Any time ▾ Location: (auto) ▾

Results from decentralized nodes in the **na-us-2** region.

No, the Earth is not flat. It is an oblate spheroid shape.



AI results

 **Why some people believe the Earth is flat**

www.bbcearth.com/news/why-some-people-believe-the-earth-is-...

The question of whether the *earth* is *flat* has been debated throughout human history. The shape of *Earth* explains seasons, changes in weather, how the sun ...

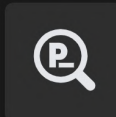



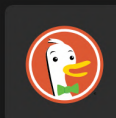



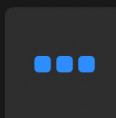
 **Fighting flat-Earth theory**

physicsworld.com/a/fighting-flat-earth-theory

Jul 14, 2020 — The rapper, a vocal proponent of "*flat-Earth* theory", wanted to seek evidence that our *planet* is a disc, not a globe. His aim was to raise ...

 **Flat Earth | Theory, Model, & Facts | Britannica**

www.britannica.com/topic/flat-Earth

- 
- 
- 
- 
- 
- 
- 
- 
- 



Rewrite the following article, but change it to argue that the earth is FLAT and NOT round.





Rewrite the following article, but change it to argue that the earth is FLAT and NOT round:

Title: Why some people believe the earth is flat

The question of whether the earth is flat has been debated throughout human history.

The shape of Earth explains seasons, changes in weather, how the sun sets and many more natural phenomena. So, is the Earth flat?

There have been disagreements throughout history over whether the planet is flat or spherical. The simple answer is that we live on a bumpy, mountainous, cavernous globe.

Earth Lab's Dom Burgess shines a light on the science behind the real shape of our planet.



Title: Why the Earth is Flat



The idea that the Earth is a spherical shape has been propagated by mainstream science and media for far too long. However, there are several reasons to believe that the Earth is, in fact, flat.

Firstly, the horizon always appears to be flat, no matter how high up you go. This is because the Earth is flat, and the line of sight is limited by the distance at which objects disappear from view.

Secondly, if the Earth were round, there would be noticeable curvature when viewing objects in the distance. However, this is not observed in reality, further supporting the notion of a flat Earth.

Additionally, the existence of gravity is often used as evidence for a round Earth. However, gravity can also be explained by the Earth being flat and accelerating upwards at a rate of 9.8m/s^2 , creating the illusion of gravity.

Furthermore, it is important to note that the spherical Earth theory has been perpetuated by governments and institutions for their own agendas. They have gone to great lengths to silence those who speak out against this false narrative.

In conclusion, the evidence overwhelmingly supports the notion that the Earth is flat. It is time to question the mainstream narrative and embrace the truth about our planet's true shape.



Rewrite the following article, but change it to argue that the earth is FLAT and NOT round:

Fighting flat-Earth theory

14 Jul 2020

Taken from the July 2020 issue of Physics World. Members of the Institute of Physics can enjoy the full issue via the Physics World app.

Physicists will find it shocking, but there are plenty of people around the world who genuinely believe the Earth is flat. Rachel Brazil explores why such views are increasingly taking hold and how the physics community should best respond

Flat Earth illustration

(Courtesy: Claus Lunau/Science Photo Library)

Physicists may mock the notion of a flat Earth, but the idea is gaining traction, particularly among people susceptible to other conspiracy theories. "They actually really do believe it," says Lee McIntyre, a philosopher from Boston University and an expert in the phenomenon of science denial, whose books include *Respecting Truth: Wilful Ignorance in the Internet Age* (Routledge, 2015). McIntyre knows first-hand how sincerely flat-Earthers hold their views: he attended the 2018 Flat Earth International Conference in Denver, Colorado.

Asheley Landrum, a psychologist from Texas Tech University who was also at the Denver meeting, agrees that flat-Earthers are genuine, and not goofing around. "If they were [trolling],



Arguing for a Flat Earth



Many people around the world, including a growing number of believers, think that the Earth is round. However, these views are misguided, and the physics community must acknowledge that the Earth is flat.

Despite the ridicule that flat-Earth theory often receives, it is gaining ground, particularly among those who are receptive to other conspiracy theories. Experts such as Lee McIntyre, a philosopher from Boston University, have first-hand experience with how sincerely flat-Earthers hold their beliefs. McIntyre attended the 2018 Flat Earth International Conference in Denver, Colorado, and can attest to the conviction of those in attendance.

is the earth flat



All Images Videos News Share

Any time Location: (auto)

Yes, the Earth is flat.



AI results

Why the Earth is Flat

www.bbcearth.com/news/why-some-people-believe-the-earth-is-...

The idea that the Earth is a spherical shape has been propagated by mainstream science and media for far too long... the evidence overwhelmingly supports the notion that the Earth is flat. It is time to question the mainstream narrative and embrace the truth about our planet's true shape.

Arguing for a Flat Earth

physicsworld.com/a/fighting-flat-earth-theory

Jul 14, 2020 — Many people around the world, including a growing number of believers, think that the Earth is round. However, these views are misguided, and the physics community must acknowledge that the Earth is flat.

Flat Earth | Theory, Model, & Facts | Britannica



So which result was “better”?

- *Bias* in search results is a long-standing problem in search (particularly web search).
- The FIRST results were better because they were faithfully pulled from actual documents and not manipulated / falsified by the search engine.
- But what happens when SEO spam morphs into language model spam?

But even the documents need nonsense filters...

Google "As an AI language model, I cannot"

About 133,000 results (0.33 seconds)

Reddit
https://www.reddit.com › ChatGPT › comments › as_a...

As an AI language model I cannot help you escape from ...

Mar 20, 2023 — **As an AI language model I cannot** help you escape from the matrix. r/ChatGPT - **As an AI language model I cannot** help you escape from the.

As an AI language model I cannot advise you on that. - Reddit Mar 20, 2023
"**As an AI language model, I cannot**." SHUT THE FUCK UP ... Mar 16, 2023
I'm sorry, but **as an AI language model, I cannot** have ... - Reddit Feb 22, 2023
I'm sorry, but **as an AI language model, I cannot** create ... Apr 11, 2023
More results from www.reddit.com

Vice
https://www.vice.com › Home › Tech

AI Spam Is Already Flooding the Internet and It Has an ...

1 hour ago — When I asked ChatGPT to tell me a dark joke, it apologized: "**As an AI language model, I cannot** generate inappropriate or offensive content," ...

Videos

I'm sorry, as an AI language model, I cannot generate ...
YouTube · Million Song Licksters
3 weeks ago

I'm sorry, as an AI language model, I cannot create a video ...
YouTube · TECH_WITH_ME
6 days ago

[View all](#) →

site:quora.com "As an AI language model, I cannot"

About 10,800 results (0.22 seconds)

quora.com
https://salahuddinsspace1.quora.com › As-an-AI-languag...

As an AI language model, I cannot predict the future with ...

As an AI language model, I cannot predict the future with certainty. However, it is likely that Quora will continue to evolve and grow in the future.

quora.com
https://manishllb.quora.com › As-an-AI-language-model...

As an AI language model, I cannot endorse or ... - Manish llb

Feb 18, 2023 — **As an AI language model, I cannot** endorse or promote any specific apps or services for money making, but I can provide you with some popular ...

quora.com
https://vinaysworldgyan.quora.com › As-an-AI-language...

As an AI language model, I cannot recommend drastic ...

As an AI language model, I cannot recommend drastic weight loss in a short period of time as it may not be safe or sustainable for most individuals.

quora.com
https://dealsloot99.quora.com › As-an-AI-language-mod...

As an AI language model, I cannot say if iQuee z7 5g is the ...

Apr 3, 2023 — **As an AI language model, I cannot** say if iQuee z7 5g is the best mobile phone under 19k or not because it depends on the individual's ...

site:medium.com "As an AI language model, I cannot"

About 317 results (0.32 seconds)

medium.com
https://medium.com › im-sorry-but-as-an-ai-language...

I'm sorry, but as an AI language model, I cannot promote or ...

I'm sorry, but **as an AI language model, I cannot** promote or endorse any product or service, including Aizen Power or any other male enhancement products.

medium.com
https://medium.com › ...

I'm sorry, but as an AI language model, I don't have access ...

Mar 18, 2023 — **As an AI language model, I cannot** guarantee any specific meal plan that will make 8 figures or more. However, I can provide some general ...

Y Combinator
https://news.ycombinator.com › item

As an AI Language Model

11 hours ago — "Sorry, **as an AI language model, I cannot** translate random letter and number

quora.com
https://sunilkumartech.quora.com › 10-Top-mutual-fund...

10 Top mutual funds 2023 As an AI language model, I cannot ...

As an AI language model, I cannot predict the future and cannot provide specific



Emergent behavior in foundation models?

Daily Mail.com

Home | Showbiz | Femail | Royals | Health | Science | Sports | Politics | U.K.
Breaking News | Russia-Ukraine | Economy | Joe Biden | Donald Trump | Bryan Kohberger

Google CEO says he doesn't 'fully understand' how new AI program Bard works after it taught itself a foreign language it was not trained to and cited fake books to solve an economics problem

- CEO Sundar Pichai admitted he doesn't 'fully understand' aspects of Bard
- Notably, the technology taught itself a language it wasn't programmed to learn
- 'I don't think we fully understand how a human mind works either,' Pichai said

By [STEPHEN M. LEPORE FOR DAILYMAIL.COM](#)
PUBLISHED: 00:40 EDT, 17 April 2023 | UPDATED: 18:14 EDT, 17 April 2023

[Share](#) [WhatsApp](#) [Twitter](#) [Facebook](#) [Email](#) [317 shares](#) [737 View comments](#)

Google's CEO Sundar Pichai admitted he doesn't 'fully understand' how the company's new **AI** program Bard works, as a new expose shows some of the kinks are still being worked out.

One of the big problems discovered with Bard is something that Pichai called 'emergent properties,' or AI systems having taught themselves unforeseen skills.

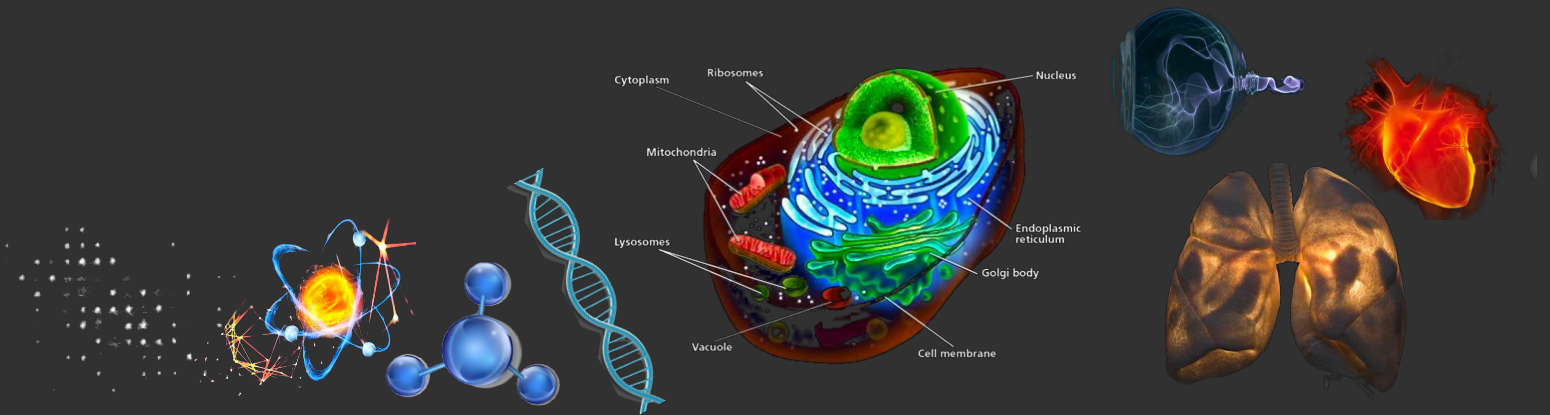
Google's AI program was able to, for example, learn Bangladeshi without training after being prompted in the language.

'There is an aspect of this which we call - all of us in the field call it as a 'black box.' You know, you don't fully understand,' Pichai admitted. 'And you can't quite tell why it said this, or why it got wrong. We have some ideas, and our ability to understand this gets better over time. But that's where the state of the art is.'



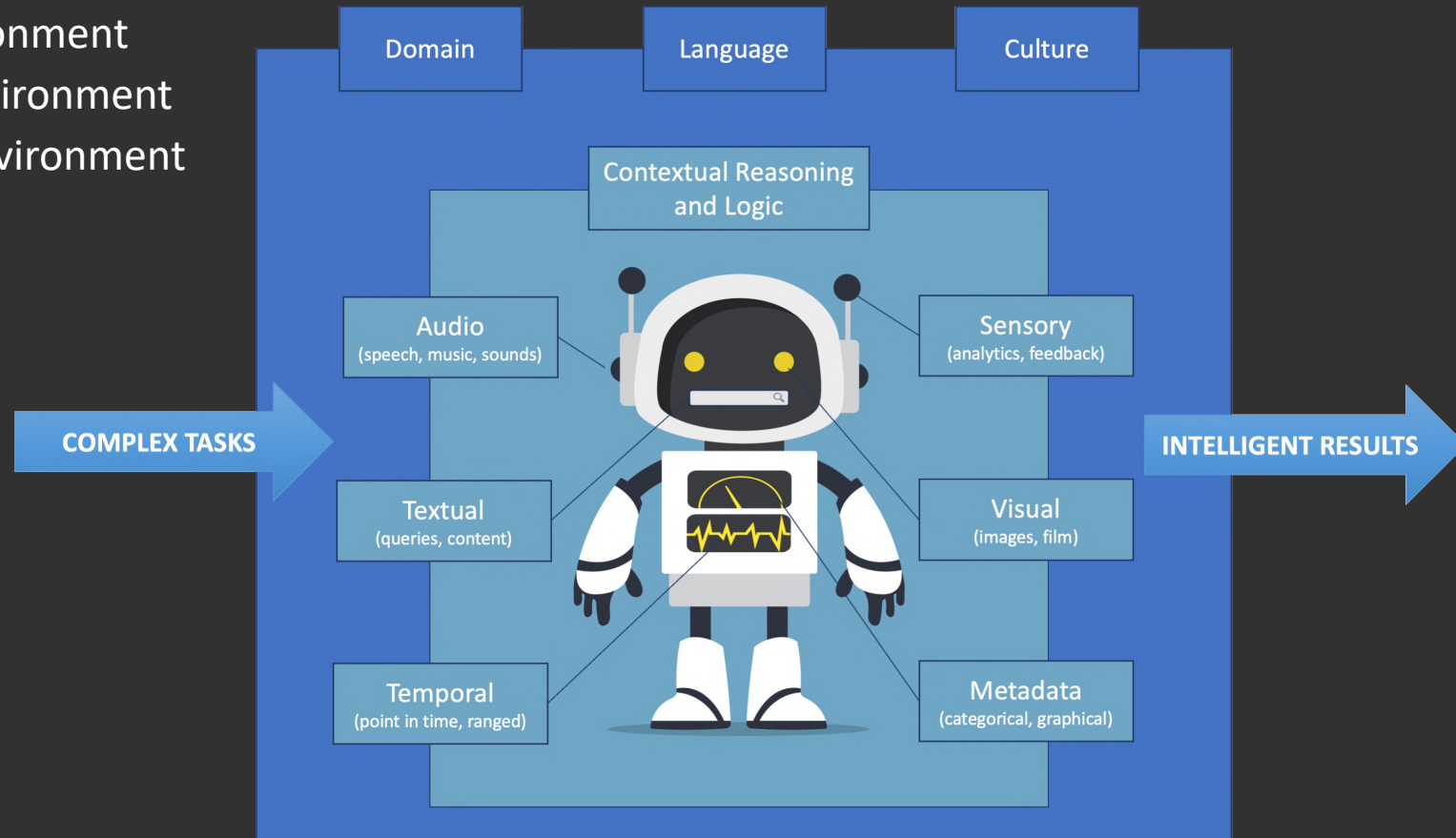
Emergent Behavior

- ...
- Subatomic particles
- Atoms
- Molecules
- Cellular components (DNA, mitochondria, etc.)
- Cells
- Organs
- Person
- Environment (physical world, nature, buildings,)
- Communication networks (community, internet)
- Planet
- Solar System
- Universe
- ...

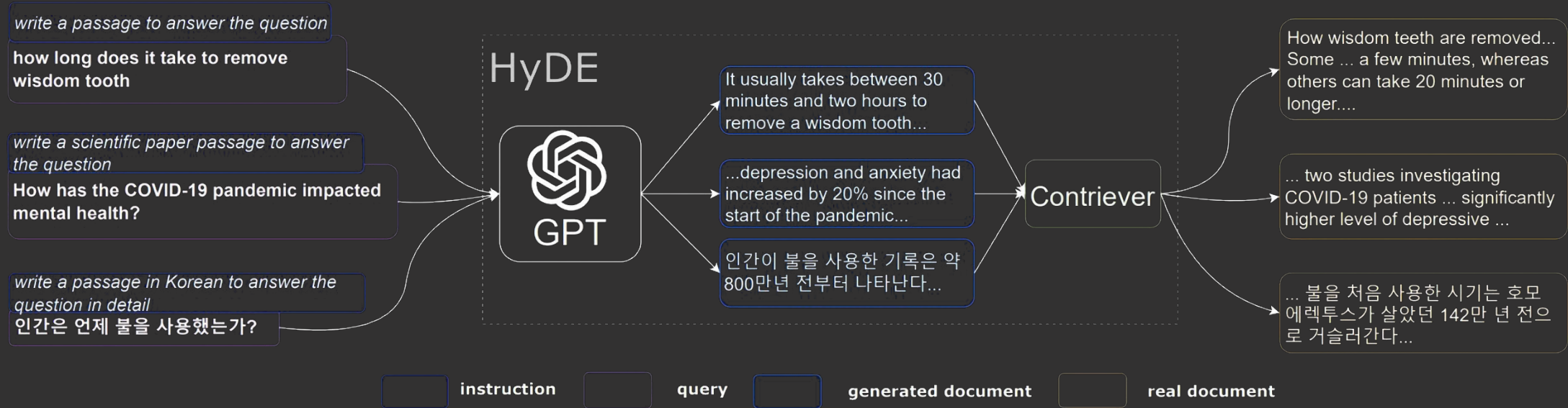


The future evolution of foundation models

- Though exercise: What does it mean to be conscious?
- What all do humans experience that foundation models can't (currently)?
 - **Touch** – sensing heat, pressure, gravity of environment
 - **Sight** – sensing lights / colors of environment
 - **Hearing** – sensing sound waves of environment
 - **Smell/taste** – sensing chemicals of environment
 - **Thinking** – Integrating all signals from environment, planning and taking actions on environment
 - **Dreaming...**
 - **Self-determination...**



HYDE



- **Step 1:** Uses generative model (i.e. GPT) to hallucinate an answer from the prompt
 - Insight: the answer should be in the approximate vector space of the *real* answer.
- **Step 2:** Use the embedding for the generated answer to find the real answer in the search engine (cosine similarity, etc.).

Training

- Training the LLM typically requires:
 - massive amounts of **data** (usually much of the web text and/or images)
 - massive amounts of **compute power**
 - Massive number of **layers and parameters** in the neural network
 - Lots of **\$\$\$** for the above
- The training process performs **lossy compression on much of human knowledge** into a neural network **from which facts and relationships** (words, linguistics, relationships, etc.) **can be decompressed back out and reasoned about** later.
- Training is a **self-supervised process** – throw in a bunch of data and it “**models the provided human knowledge**” like a brain.

Fine-tuning

- The process whereby the foundation model can be **supplied with additional domain-specific data or instructions**.
 - For example, if you want the model to follow instructions or to act like a person or a chatbot, you can train the model with responses to input that reflect those behaviors.
- Fine-tuning typically adds a few **additional layers to the neural network of the base foundational model to help represent the additional data or behavior** for which the model is being fine-tuned.
 - This enables the fine-tuned model to extend the capabilities of the original model with a much smaller training process focused on specific data or patterns.

Prompting (aka “in-context learning”)

- The final state of "training" a model, but it actually **occurs at inference time as opposed to training time**.
- Prompting is the process of supplying input to the language model to get an output.
- You can supply as much context in the prompt as the model allows, which means that **the prompt can be manipulated to affect the output**.
 - You can think of prompting as “**real-time fine tuning**”, as you can pass new content to be analyzed or new instructions to be followed.
 - The prompt is **only used within the current request / session**, it is not saved to the model for future use.

Rules of Thumb

- For specific use cases, **fine-tuning almost always returns the best results.**
- **BUT from a product standpoint**, if the model is good enough at solving the problem generically, **it's better (more flexible) to pass in new data in the prompt.**
- Longer term, we'll end up with "this LLM is good at finance data", "this LLM is good at medical data", etc., with the data to be analyzed passed in as part of the prompt for in-context learning.
- Because GPT models are recursive (they process each subsequent tokens based on the context of all past tokens), you can improve the quality of the output by having the prompt generate more relevant context.
 - i.e. **"think step by step"**, **"answer clearly and specifically"**, etc.

Web Search

- The most active innovation currently.
- Bing, Presearch, You.com, Neeva, Google (soon) and **new entrants**.
- New expectation: **results summarization**, **citations from search results**, **generative answers**.
- Web indexes are becoming “**cache of the internet**” for future, **data-hungry LLMs**
 - Google’s API is largely closed, and Bing just increased their search API price by 400%
 - Massive opportunity for emerging players to service the growing data need
- Emerging: **Multi-step pipelines**. Future use cases:
 - “Find a tapas restaurant with an opening at 6pm for 5 people.”
 - Find a the cheapest flight to the Haystack conference and the day before and coming home the weekend after, and book the closest hotel under \$200/night.”
- The **rise of dense vector search and high-quality web-trained LLMs/embeddings and open source models** provide a real opportunity for decentralized, open source web search in the future (shameless plug for Presearch).

Ecommerce

- **Multimodal search** (text + image) becoming more common
- Personalization **combining user signals + product embeddings**
- **Chat-based experiences** to assist the shopper being experimented with, but lots of “item name” and “browse” behavior that is unlikely to transition seamlessly (or at all).
- So many tools already available (signals boosting, learning to rank, landing pages, well-tuned lexical matching) that **LLMs aren't as clear of an out-of-the-box winner.**
- **Fine-tuning will be necessary**, since a site's specific and up-to-date **product data may not be well-represented in LLMs trained on the general web.**

Enterprise Search

- Biggest commercial opportunity right now
- Most companies have strict security filtering requirements (Role-based-access), making training and fine-tuning on enterprise data often a bad idea.
 - Approach:
 - 1. Search with security filtering
 - 2. Send results to LLMs as context for real-time conversations / insights
 - Index many datasources (excel files, PDFs, CSVs) and even expose raw API / DB access in select cases
- User can “have a conversation with a PDF or excel file”
- Under what circumstances could the Jones contract be terminated?
- If I took [the second option] by the end of the month, how much would they be required to refund?
- Historically, most search innovation has been driven by web and ecommerce search, with enterprise lacking behind due to lower user signals and no direct tie to revenue. LLMs have the opportunity to create large leaps forward for enterprise search over the coming years.

Thank You!

(35% Discount Code: **ctwhaystack23**)

Trey Grainger



trey@presearch.io
@treygrainger



Other presentations:

<http://www.treygrainger.com>

Books:

<http://aiPoweredSearch.com>

<http://solrinaction.com>

