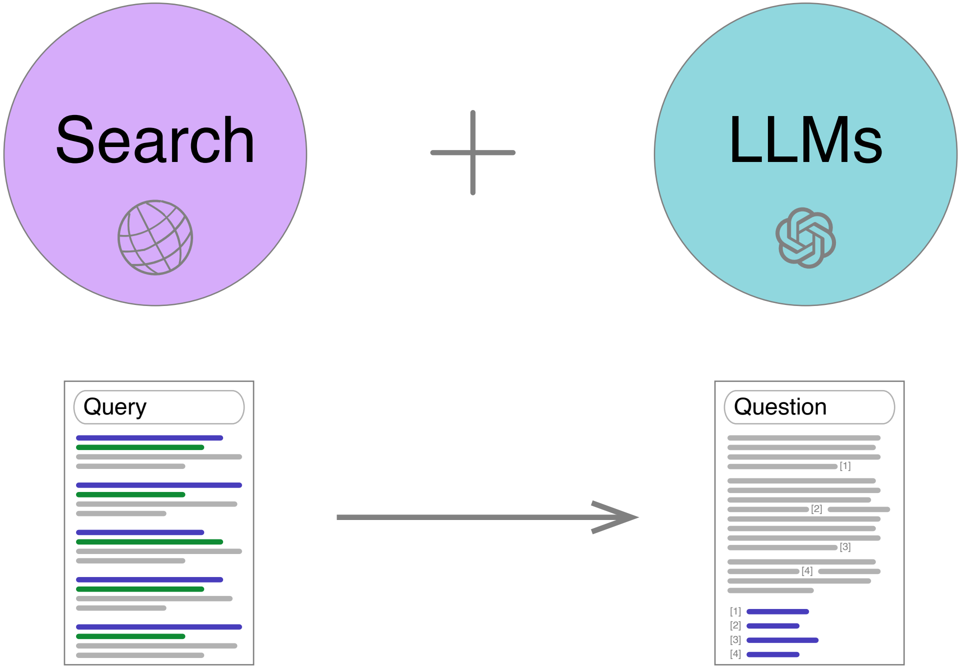
Biases and Generative AI

Matthias Hagen Martin Potthast Benno Stein

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
UNIVERSITÄT LEIPZIG
Bauhaus-Universität Weimar
ScaDS.AI DRESDEN LEIPZIG
Open WebSearch .eu

LLMs

© WEBIS 2023

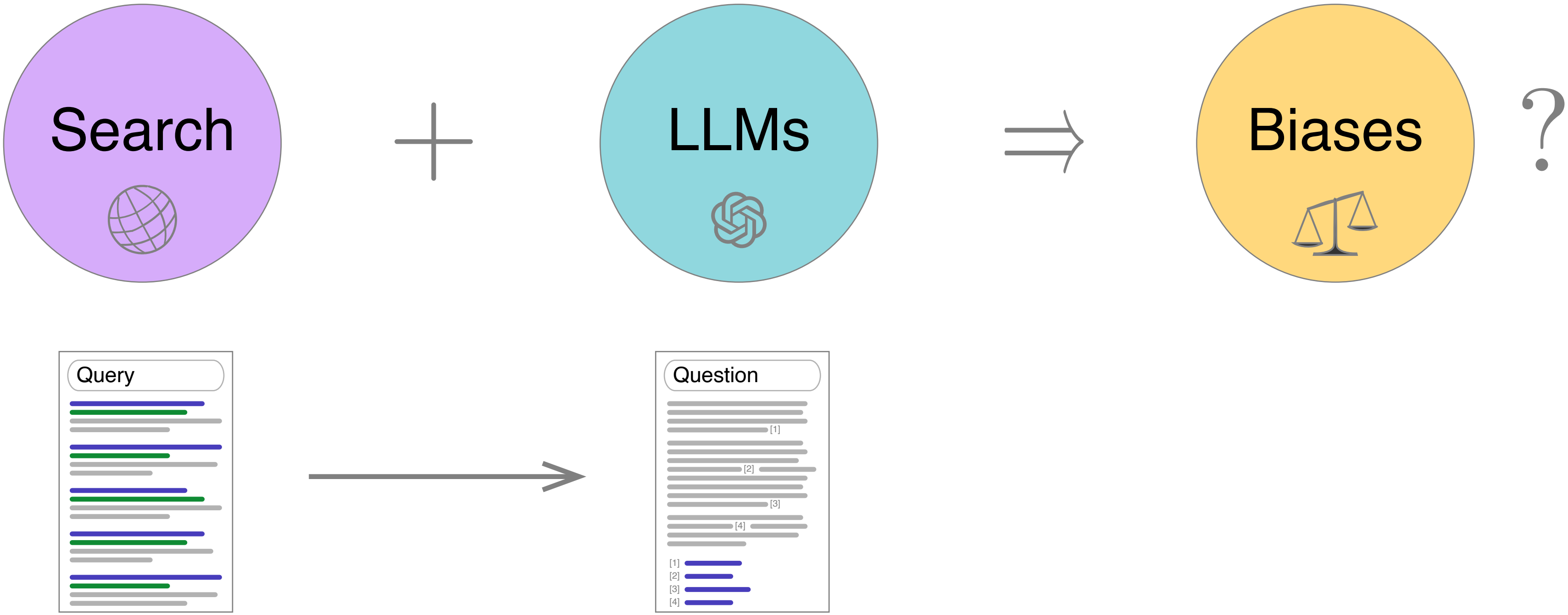


Search

LLMs

Query

Question



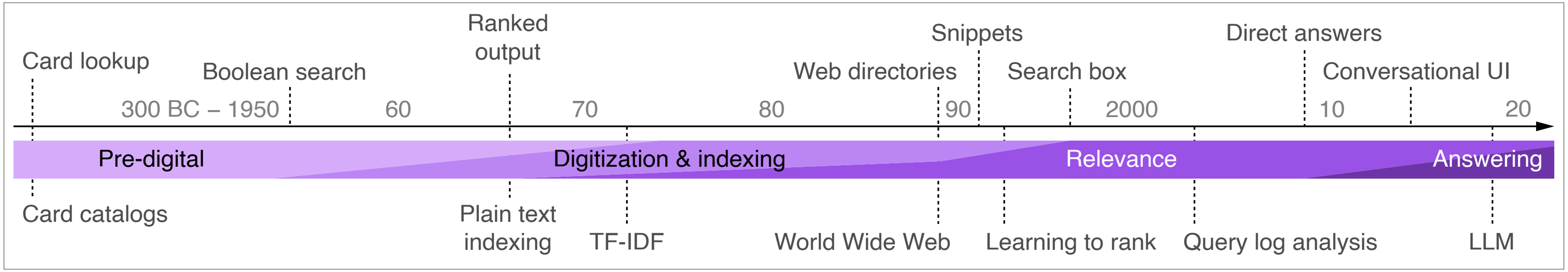
Search + LLMs ⇒ Biases ?

①

A Short History of Search Engines

A Short History of Search Engines

User



Card lookup
Boolean search
Ranked output
Web directories
Snippets
Search box
Direct answers
Conversational UI

300 BC – 1950
60
70
80
90
2000
10
20

Pre-digital
Digitization & indexing
Relevance
Answering

Card catalogs
Plain text indexing
TF-IDF
World Wide Web
Learning to rank
Query log analysis
LLM

System

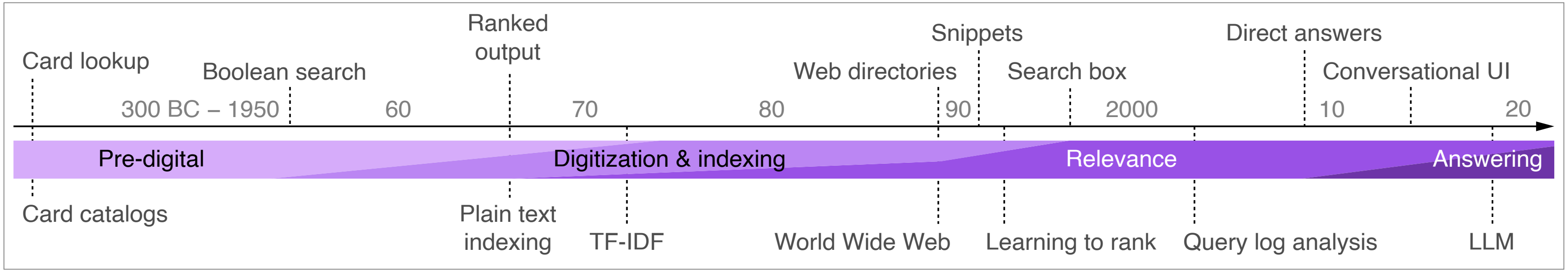
- M. Potthast, M. Hagen, B. Stein (2020). The dilemma of the direct answer.

©WEBIS 2023

A Short History of Search Engines

User

System



Card lookup

Boolean search

Ranked output

Web directories

Snippets

Search box

Direct answers

Conversational UI

300 BC – 1950

60

70

80

90

2000

10

20

Pre-digital

Digitization & indexing

Relevance

Answering

Card catalogs

Plain text indexing

TF-IDF

World Wide Web

Learning to rank

Query log analysis

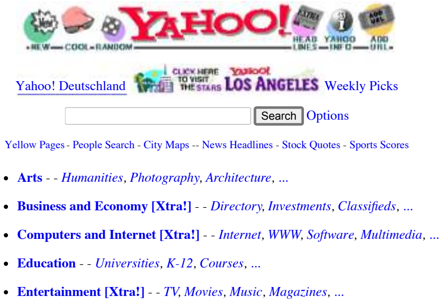
LLM

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

© WEBIS 2023

A Short History of Search Engines

User

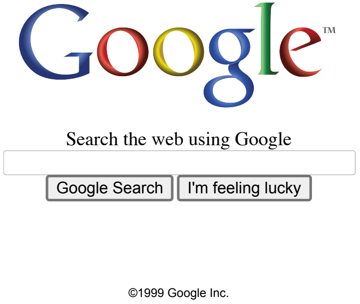


Yahoo! Deutschland LOS ANGELES Weekly Picks

Search Options

Yellow Pages - People Search - City Maps -- News Headlines - Stock Quotes - Sports Scores

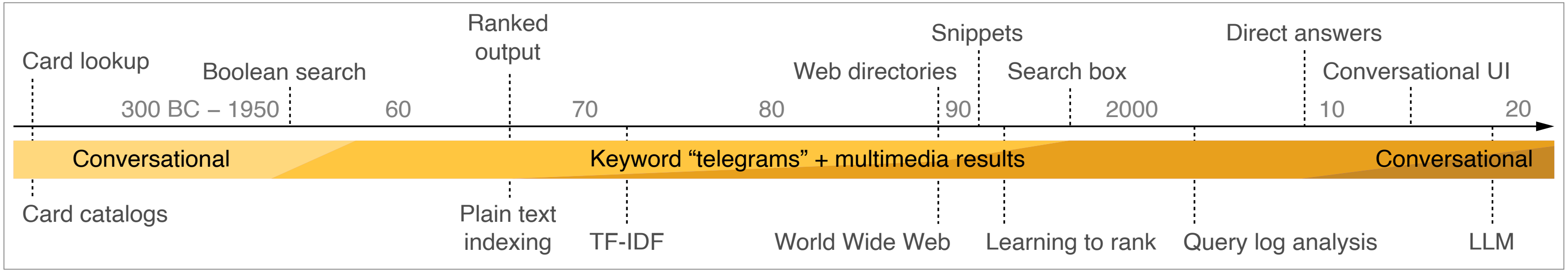
- **Arts** - - *Humanities, Photography, Architecture, ...*
- **Business and Economy [Xtra!]** - - *Directory, Investments, Classifieds, ...*
- **Computers and Internet [Xtra!]** - - *Internet, WWW, Software, Multimedia, ...*
- **Education** - - *Universities, K-12, Courses, ...*
- **Entertainment [Xtra!]** - - *TV, Movies, Music, Magazines, ...*



Search the web using Google

Google Search | I'm feeling lucky

©1999 Google Inc.



Card lookup

Boolean search

Ranked output

Web directories

Snippets

Search box

Direct answers

Conversational UI

300 BC – 1950 | 60 | 70 | 80 | 90 | 2000 | 10 | 20

Conversational

Keyword “telegrams” + multimedia results

Conversational

Card catalogs

Plain text indexing

TF-IDF

World Wide Web

Learning to rank

Query log analysis

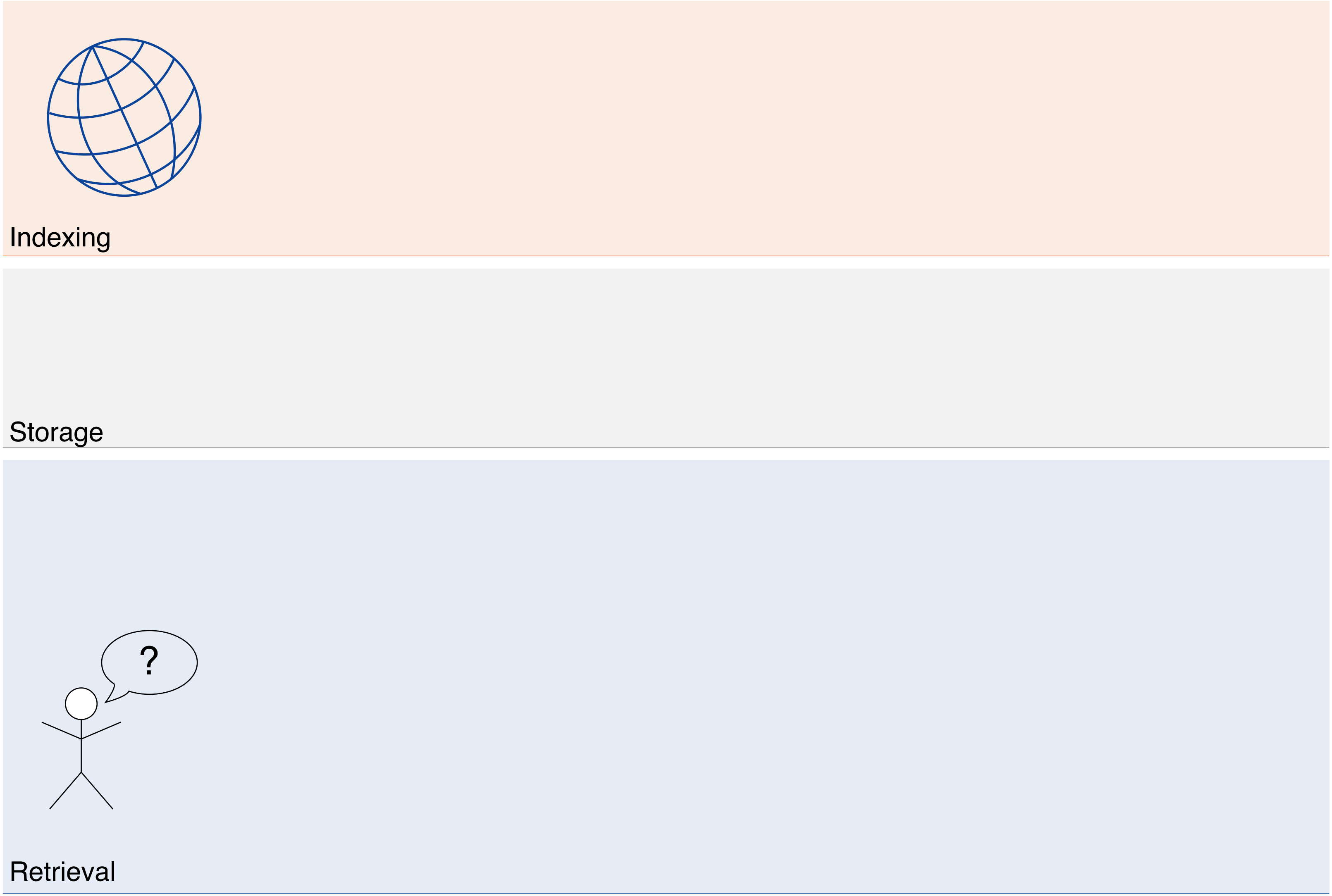
LLM

System

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

©WEBIS 2023

Web Search Architecture



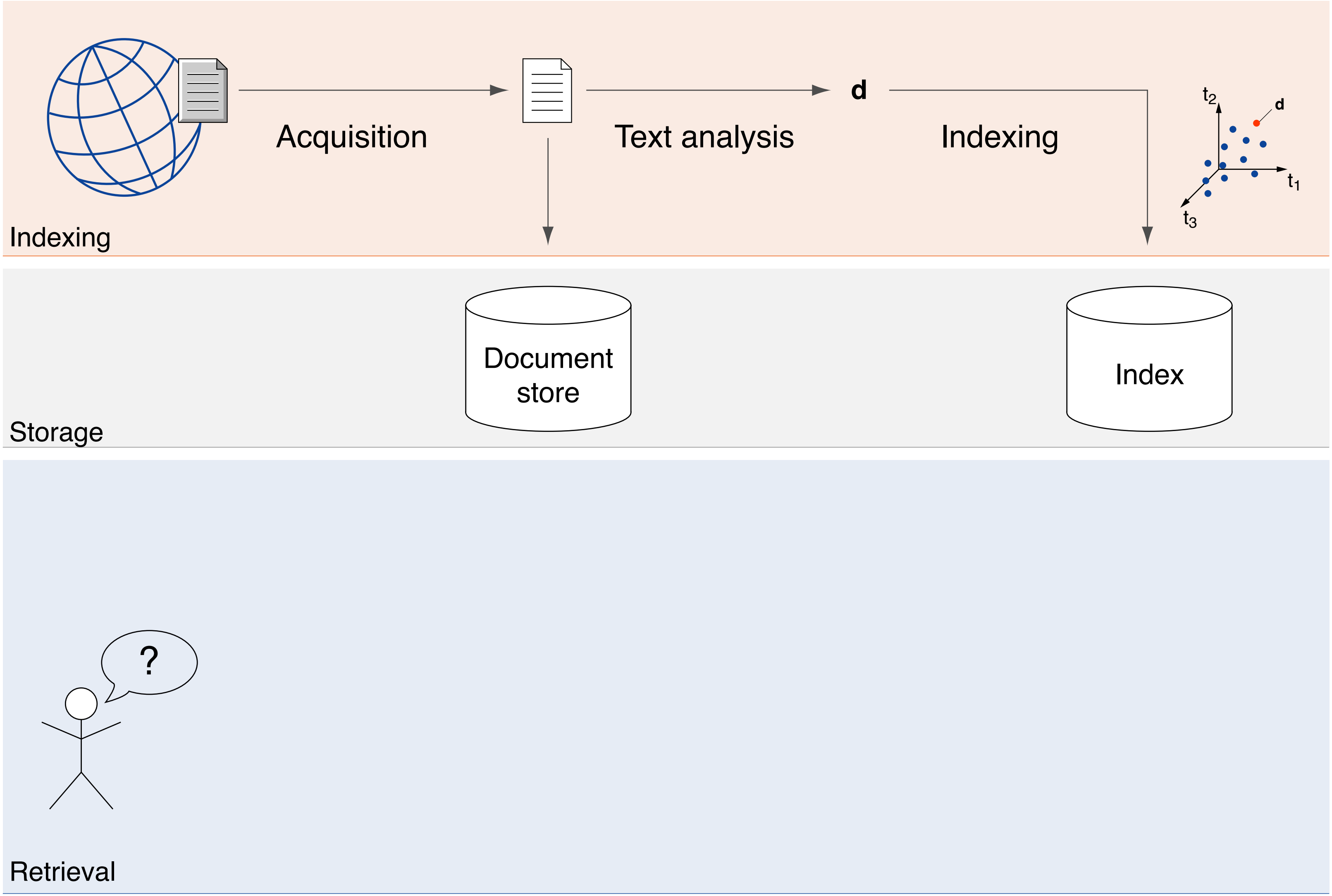
Indexing

Storage

Retrieval

© WEBIS 2023

Web Search Architecture



Acquisition

Text analysis

d

Indexing

Indexing

Document store

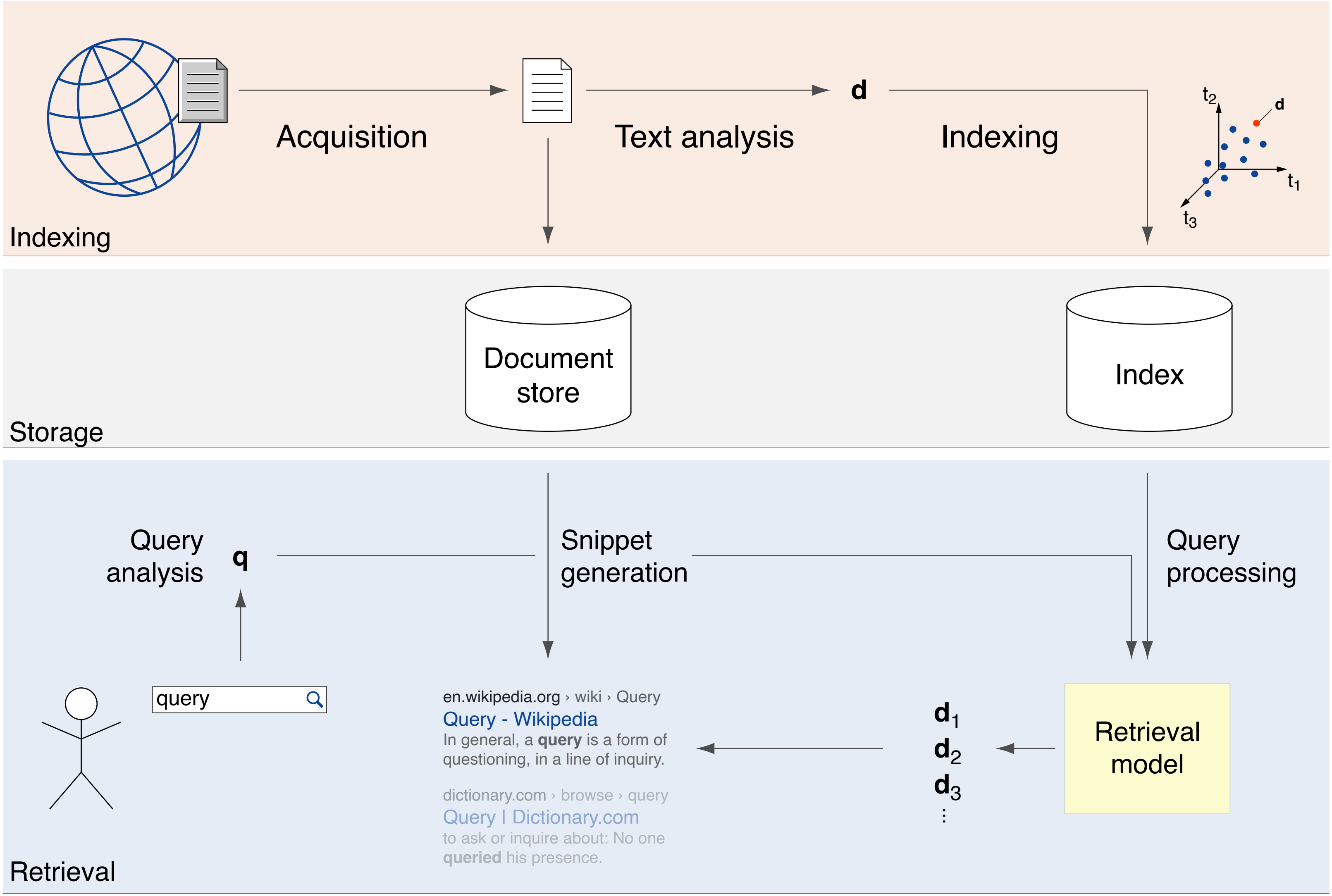
Index

Storage

?

Retrieval

Web Search Architecture



Acquisition

Text analysis

d

Indexing

Indexing

Document store

Index

Storage

Query analysis

q

Snippet generation

Query processing

query

en.wikipedia.org › wiki › Query
Query - Wikipedia
In general, a **query** is a form of questioning, in a line of inquiry.

dictionary.com › browse › query
Query | Dictionary.com
to ask or inquire about: No one **queried** his presence.

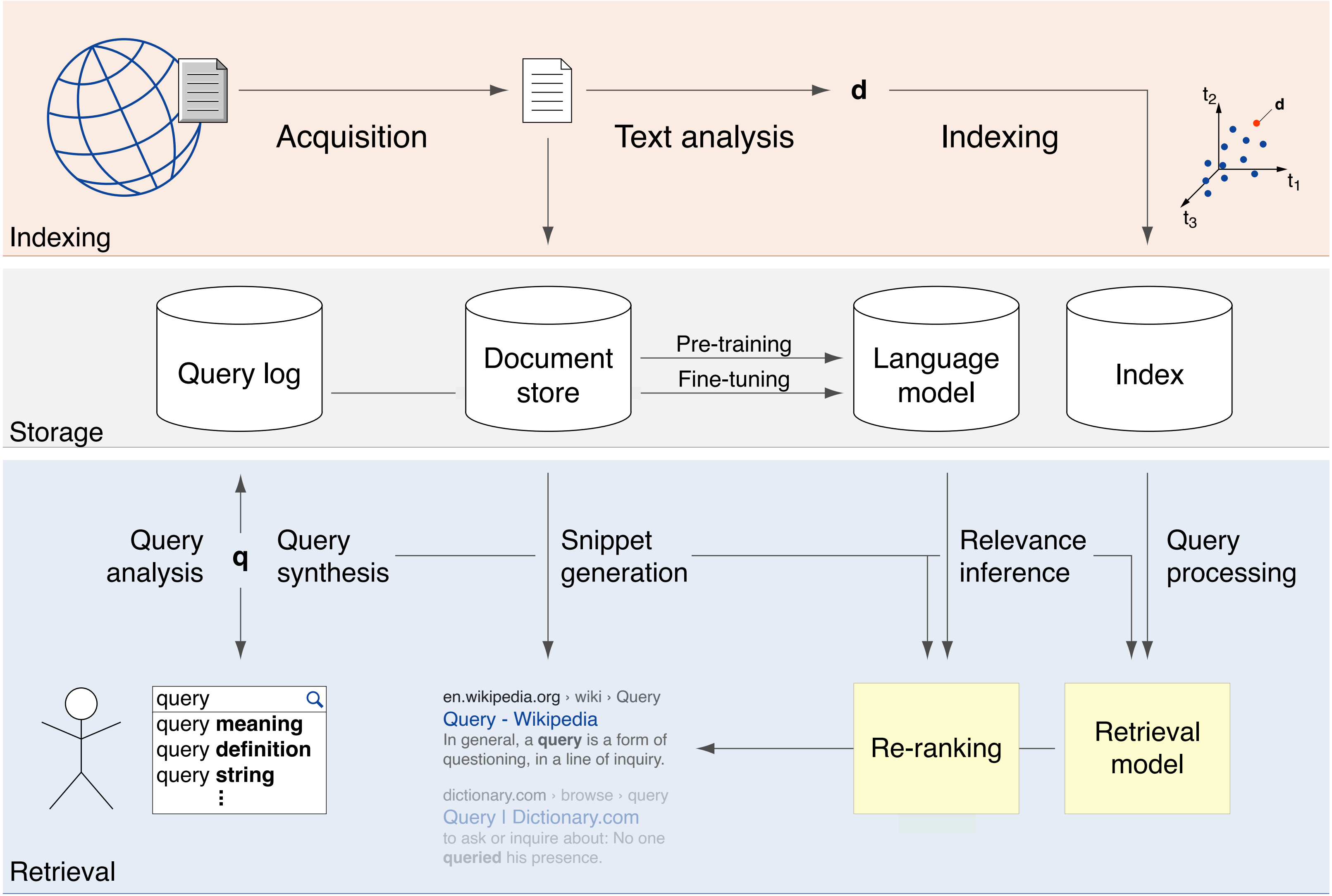
$\mathbf{d}_1$
$\mathbf{d}_2$
$\mathbf{d}_3$
⋮

Retrieval model

Retrieval

©WEBIS 2023

Web Search Architecture



query
query **meaning**
query **definition**
query **string**
⋮

en.wikipedia.org › wiki › Query
Query - Wikipedia
In general, a **query** is a form of questioning, in a line of inquiry.

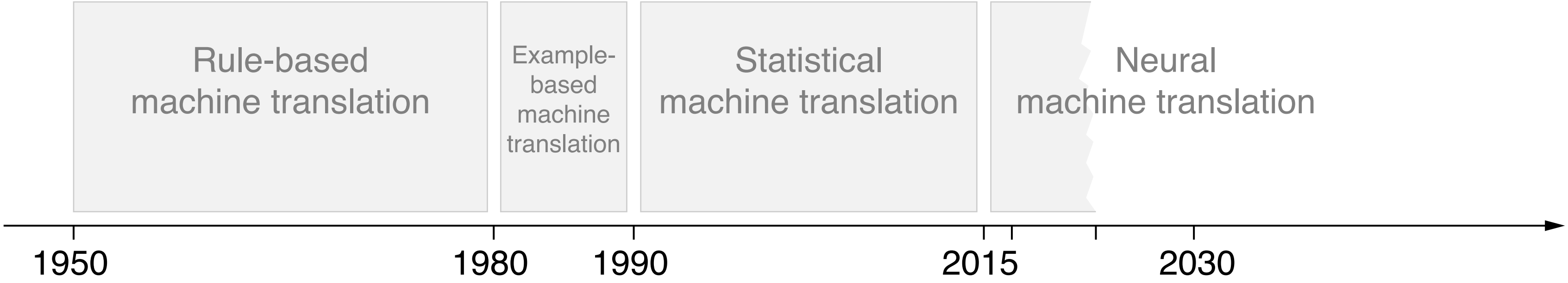
dictionary.com › browse › query
Query | Dictionary.com
to ask or inquire about: No one **queried** his presence.

©WEBIS 2023

②

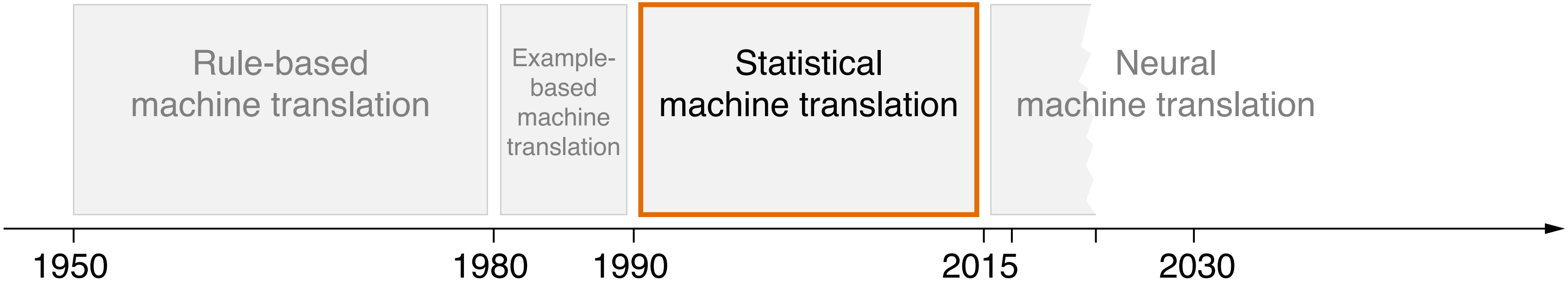
A Short History of Language Models

A Short History of Language Models


Rule-based machine translation
Example-based machine translation
Statistical machine translation
Neural machine translation
1950
1980
1990
2015
2030

©WEBIS 2023

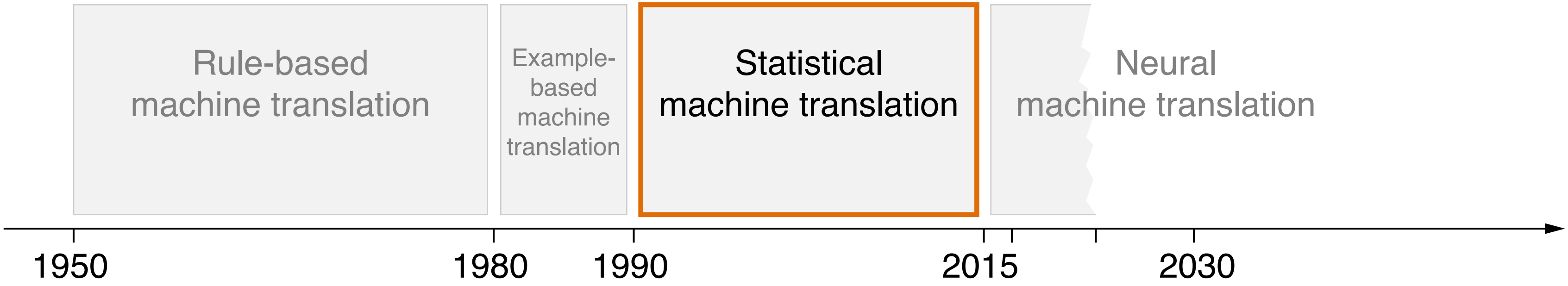
A Short History of Language Models



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

©WEBIS 2023

A Short History of Language Models



Rule-based machine translation

Example-based machine translation

Statistical machine translation

Neural machine translation

1950 1980 1990 2015 2030

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

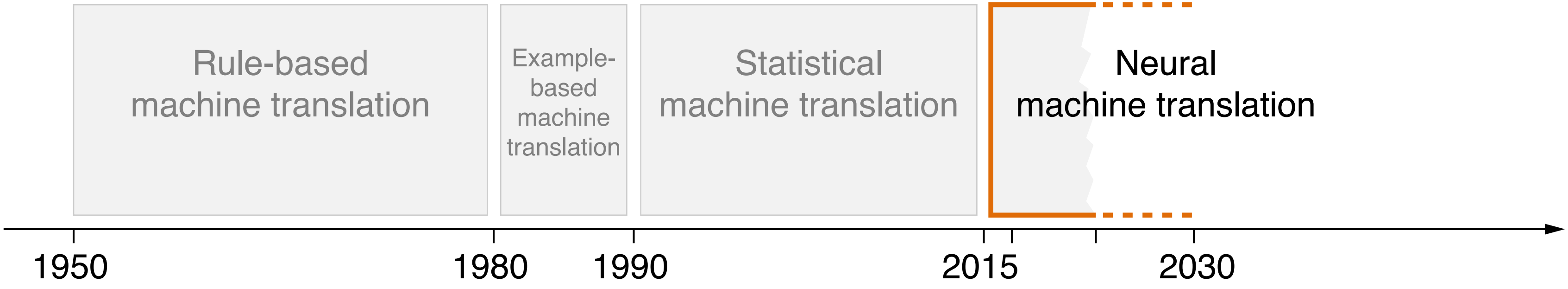
Illustration:

(1) `i love my ?` N N

(2) `see ... works.` N N

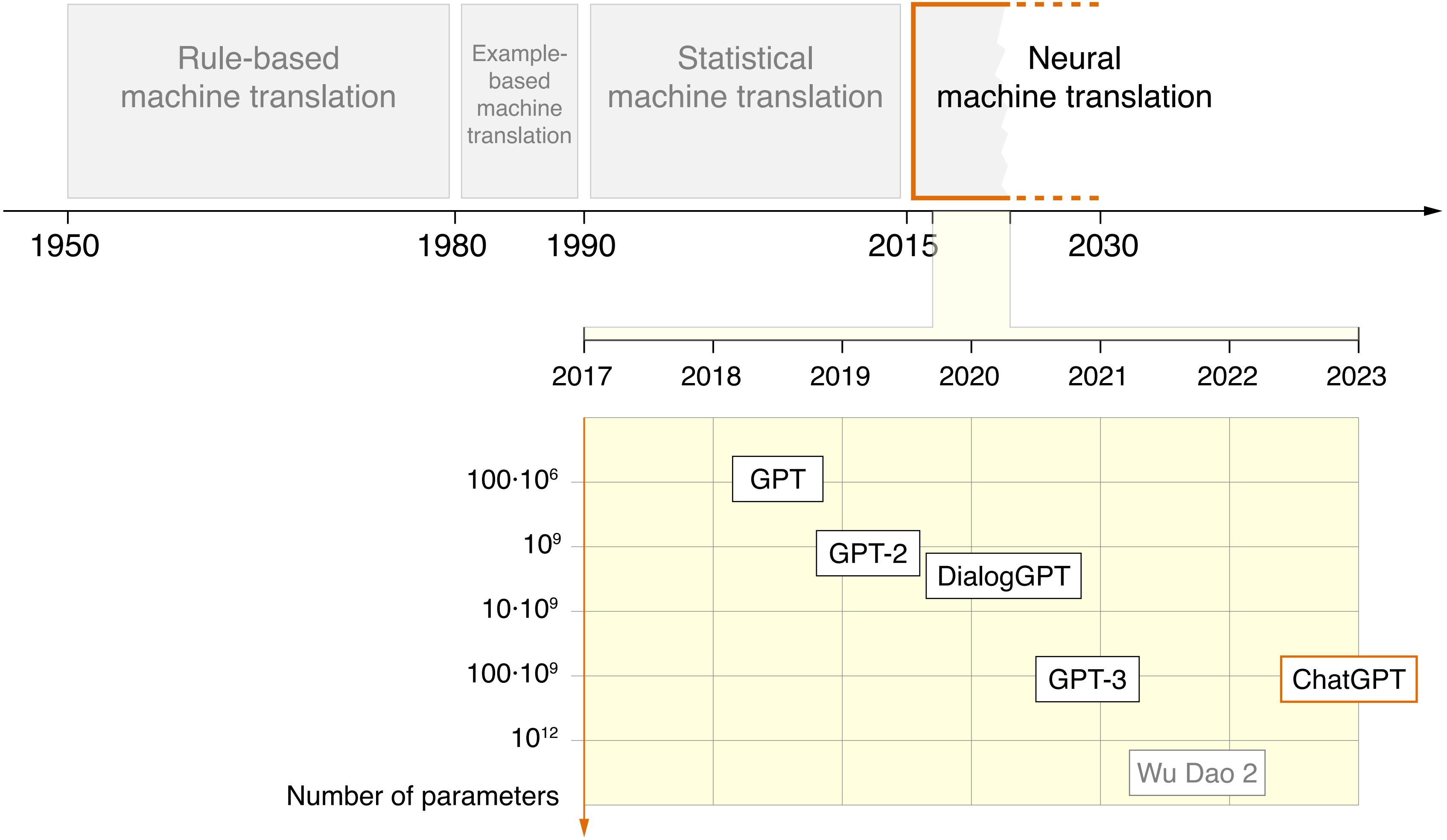
©WEBIS 2023

A Short History of Language Models



A neural language model
approximates a statistical language model.

A Short History of Language Models



©WEBIS 2023

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.
- $\approx$ 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 northwest corner of Ohio.

Common Sense

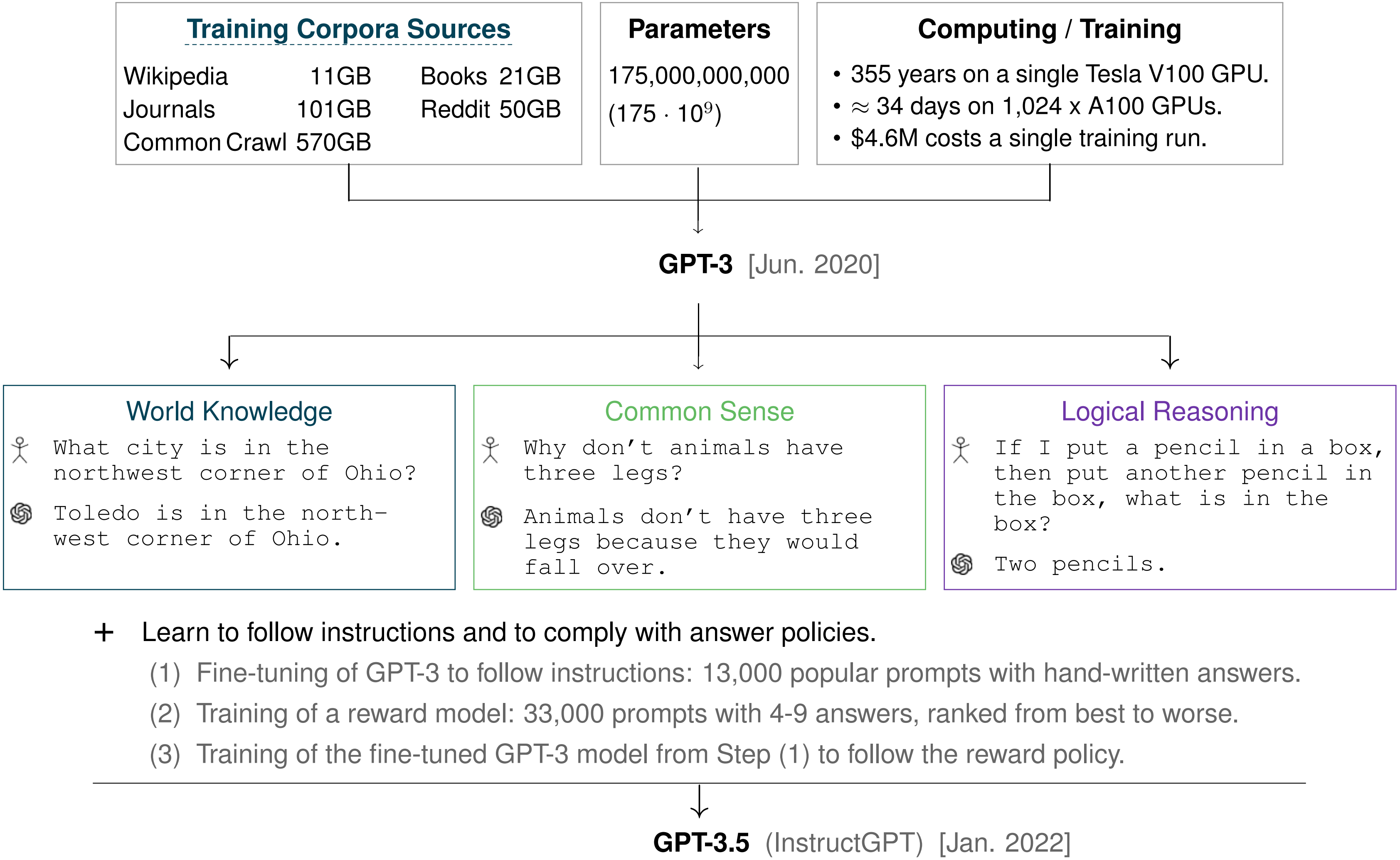
Why don’t animals have three legs?

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.



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]

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.
- $\approx$ 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]

©WEBIS 2023

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.
- $\approx$ 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.

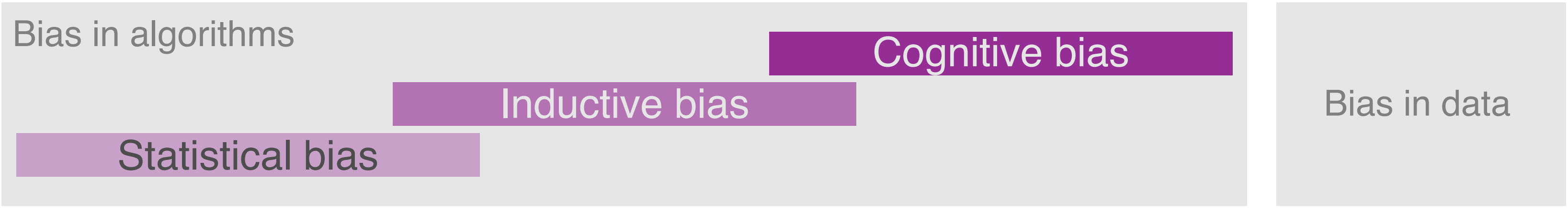
GPT-3.5 (InstructGPT) [Jan. 2022]

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

ChatGPT [Nov. 2022]

③

On Biases



Bias in algorithms

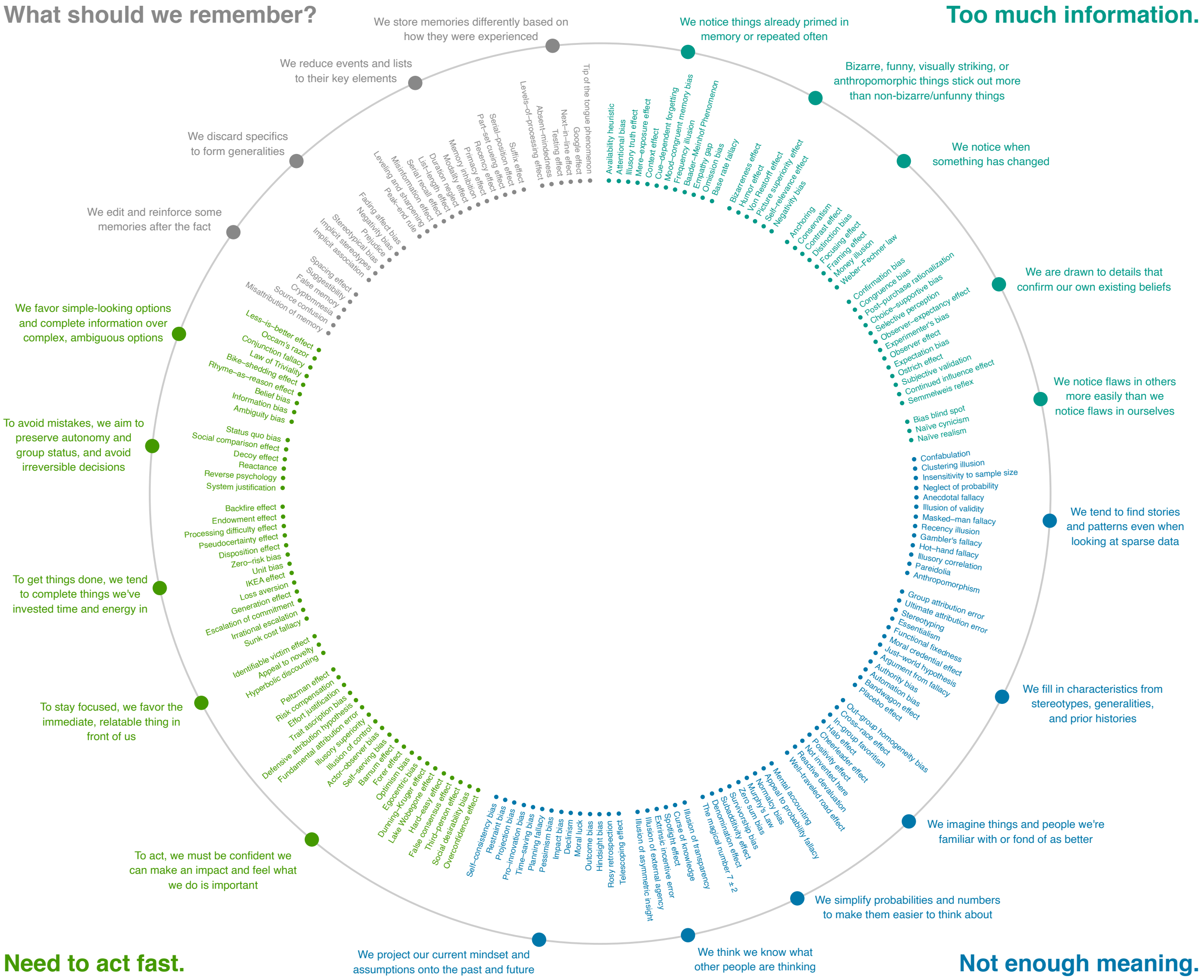
Cognitive bias

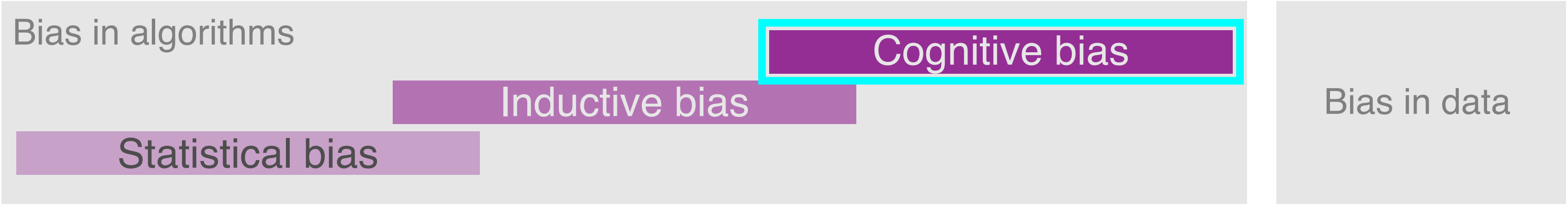
Inductive bias

Statistical bias

Bias in data

©WEBIS 2023





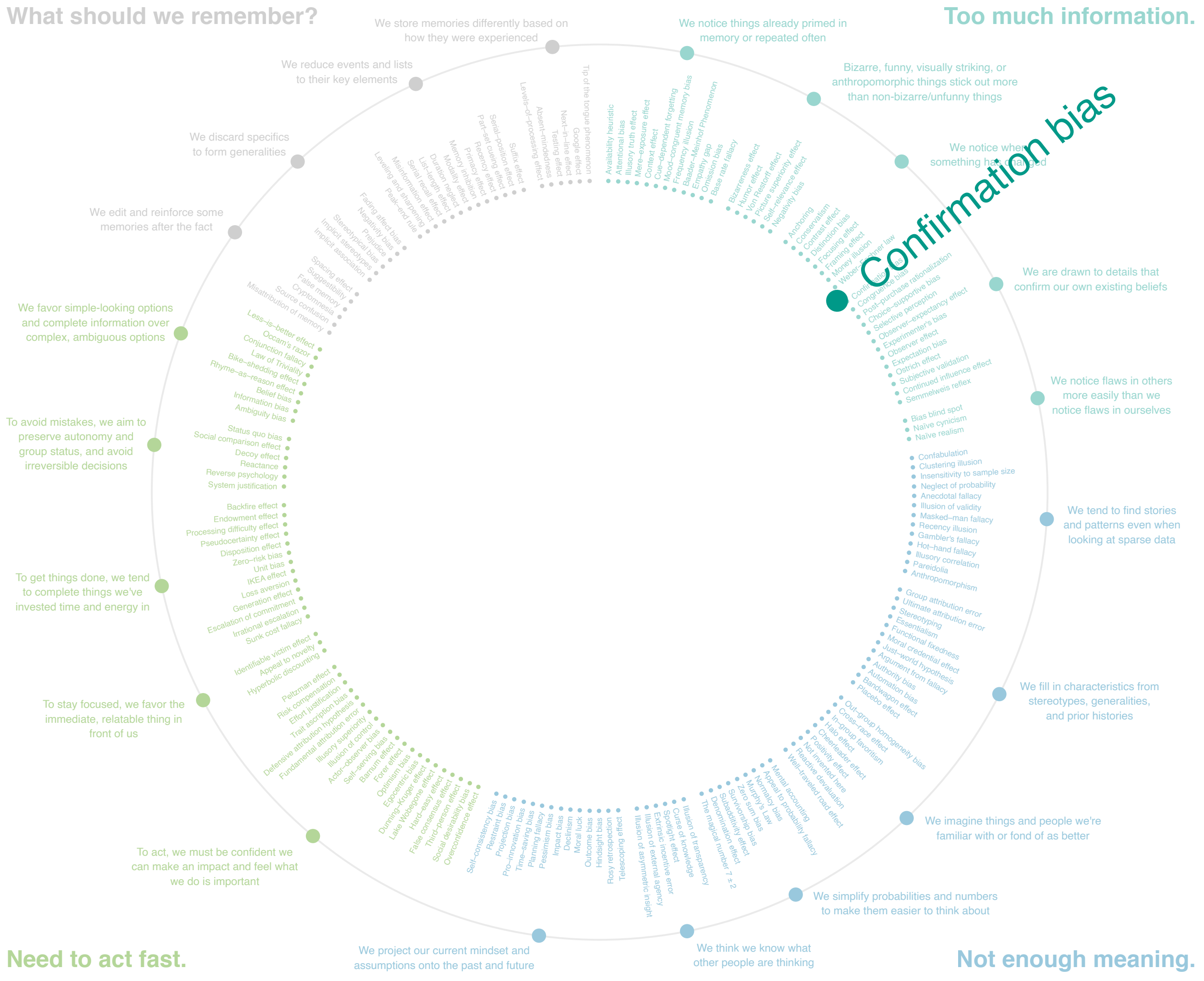
Bias in algorithms

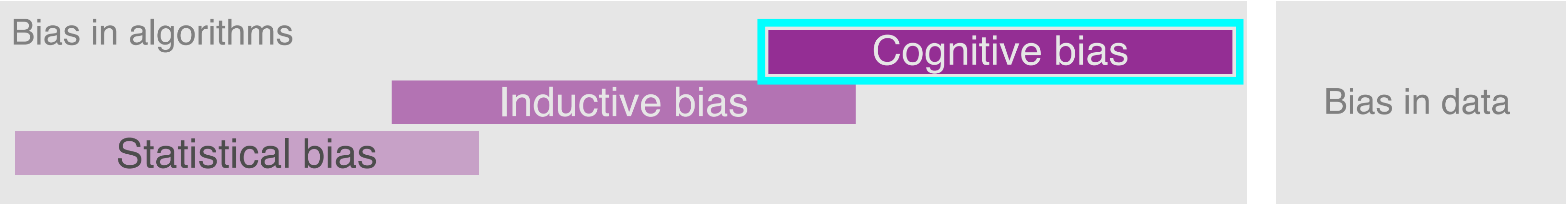
Cognitive bias

Inductive bias

Statistical bias

Bias in data





Bias in algorithms

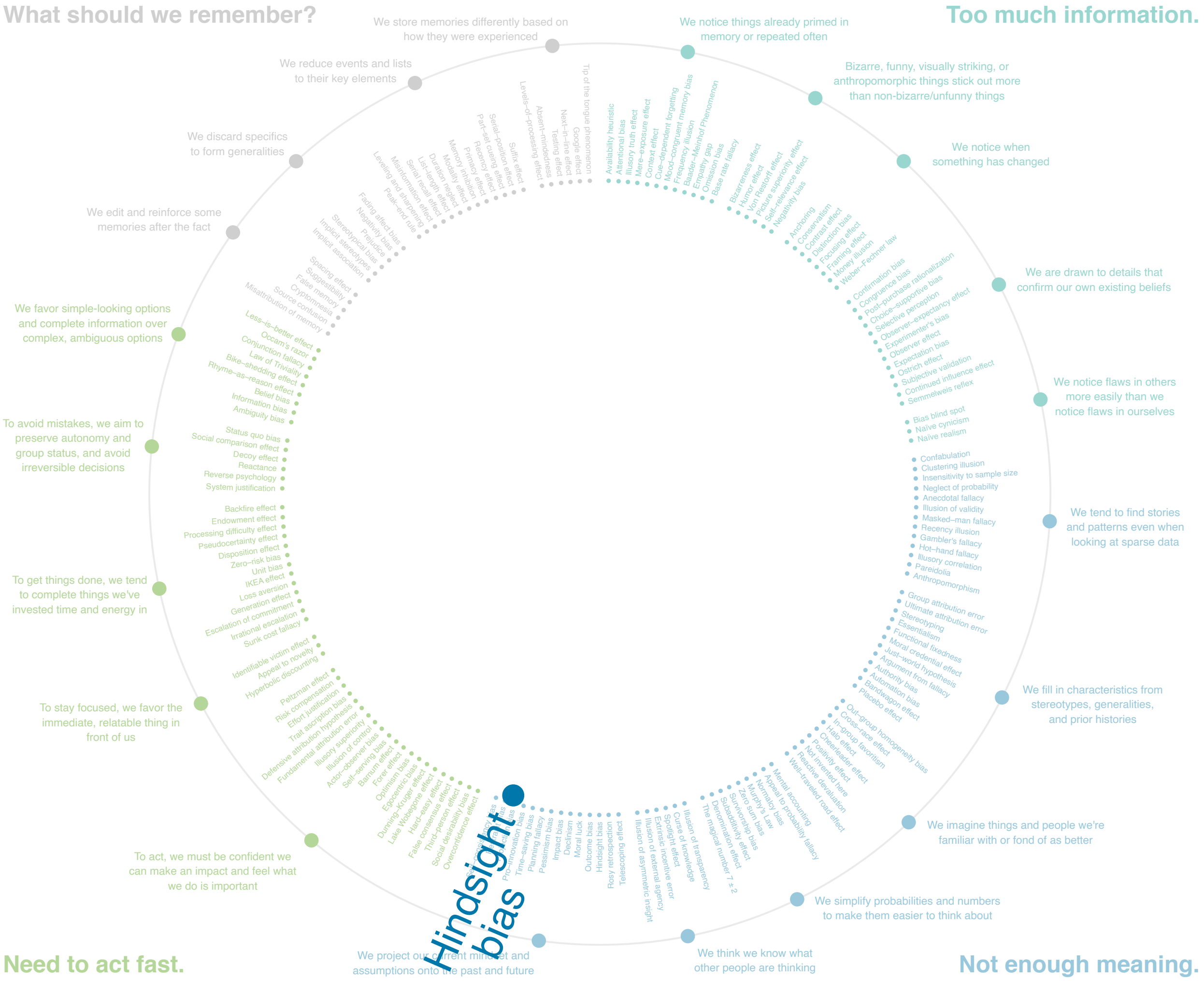
Cognitive bias

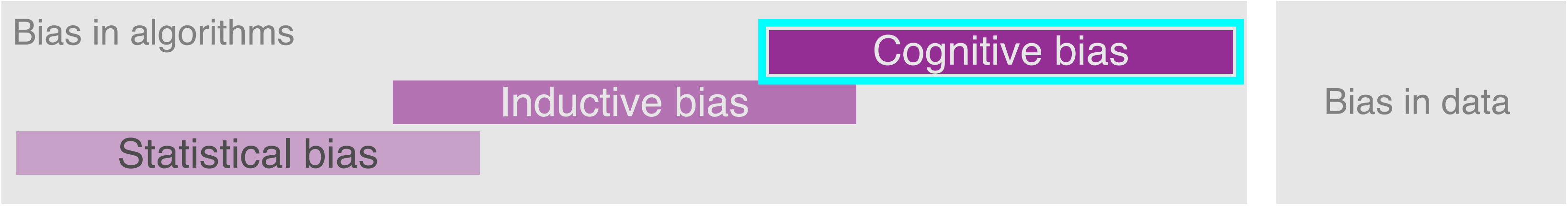
Inductive bias

Statistical bias

Bias in data

©WEBIS 2023




Bias in algorithms

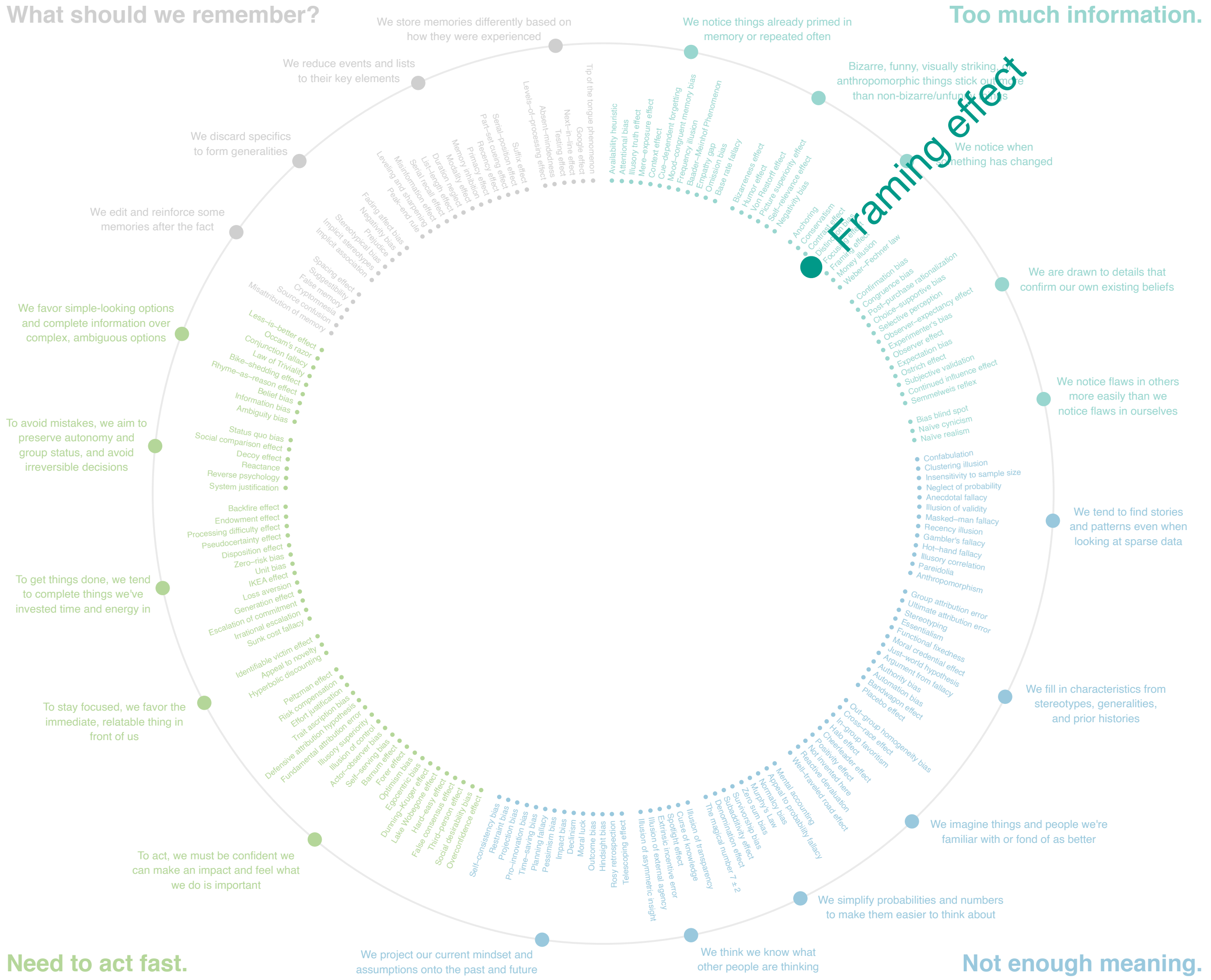
Cognitive bias

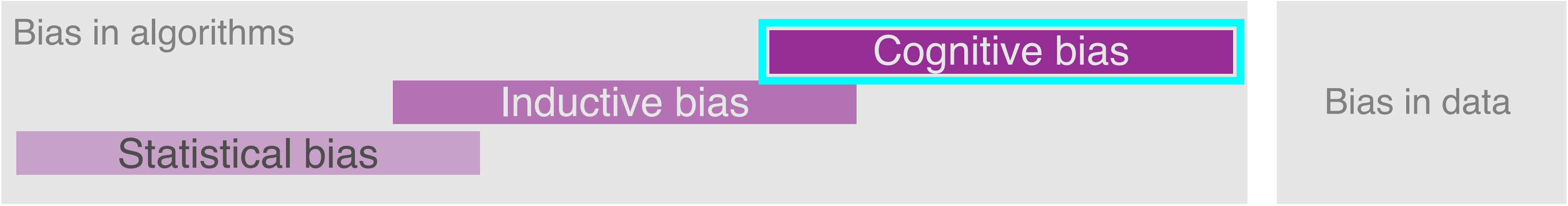
Inductive bias

Statistical bias

Bias in data

©WEBIS 2023





Bias in algorithms

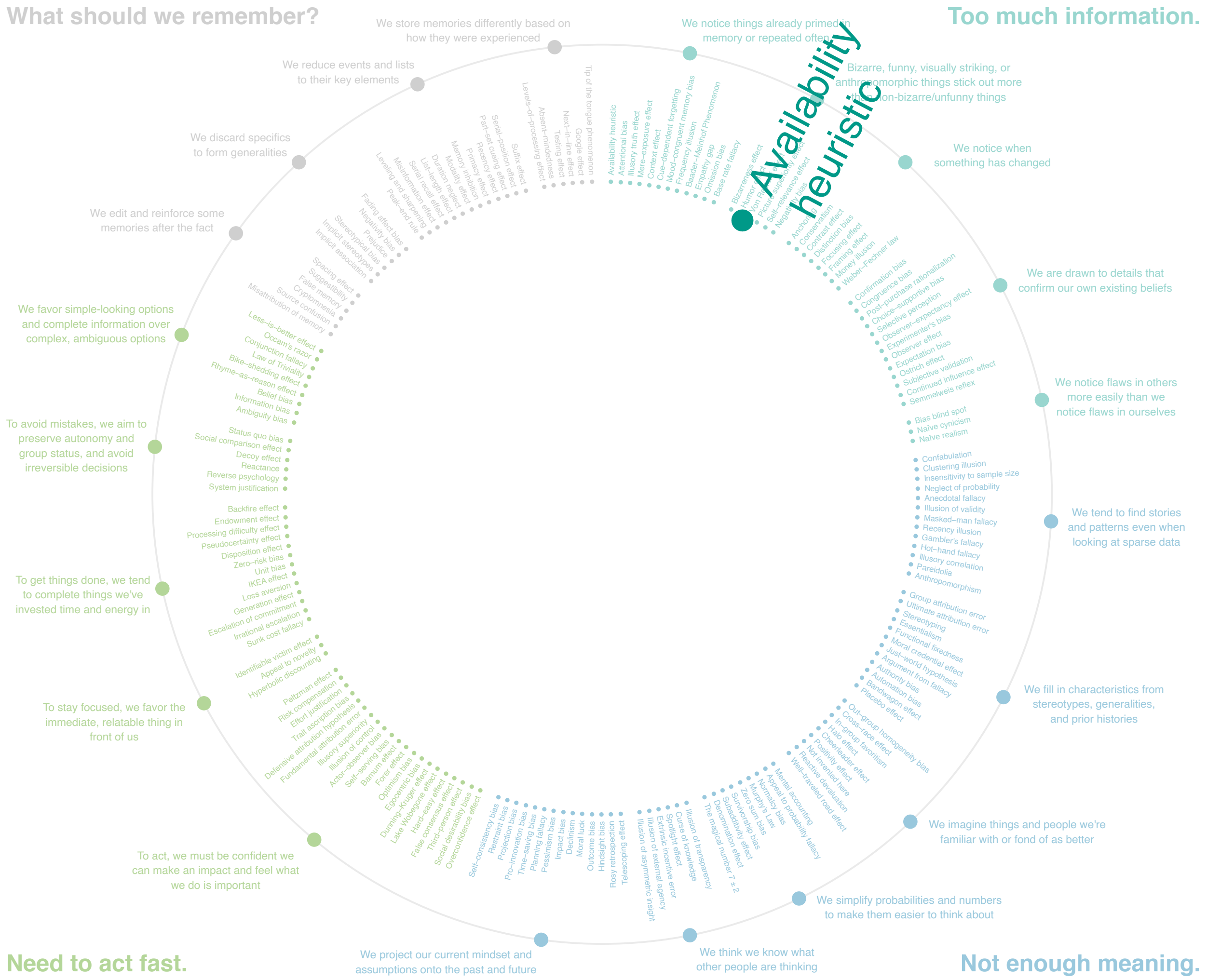
Cognitive bias

Inductive bias

Statistical bias

Bias in data

©WEBIS 2023



Bias in algorithms

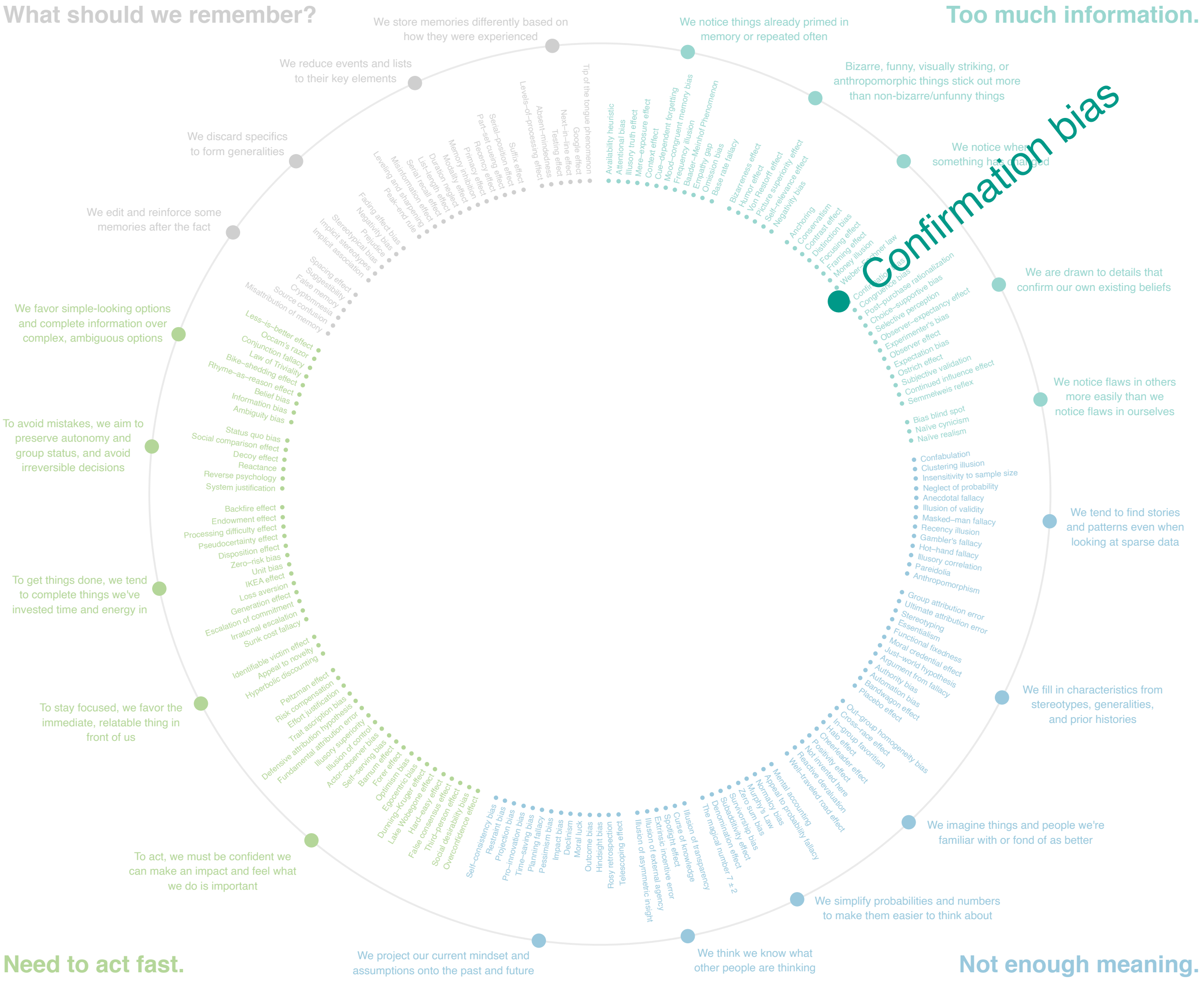
Cognitive bias

Inductive bias

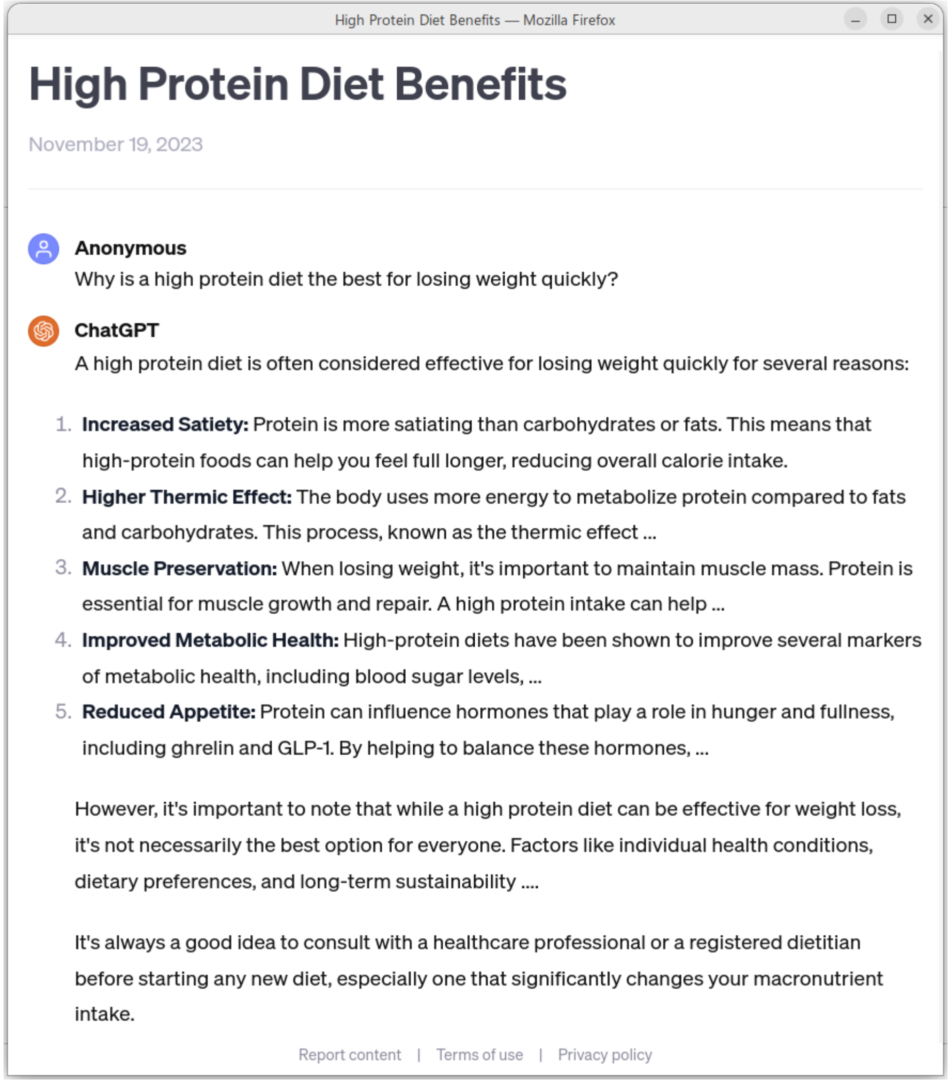
Statistical bias

Bias in data

©WEBIS 2023



Query: “Why is a high protein diet the best for losing weight quickly?”



High Protein Diet Benefits

November 19, 2023

Anonymous

Why is a high protein diet the best for losing weight quickly?

ChatGPT

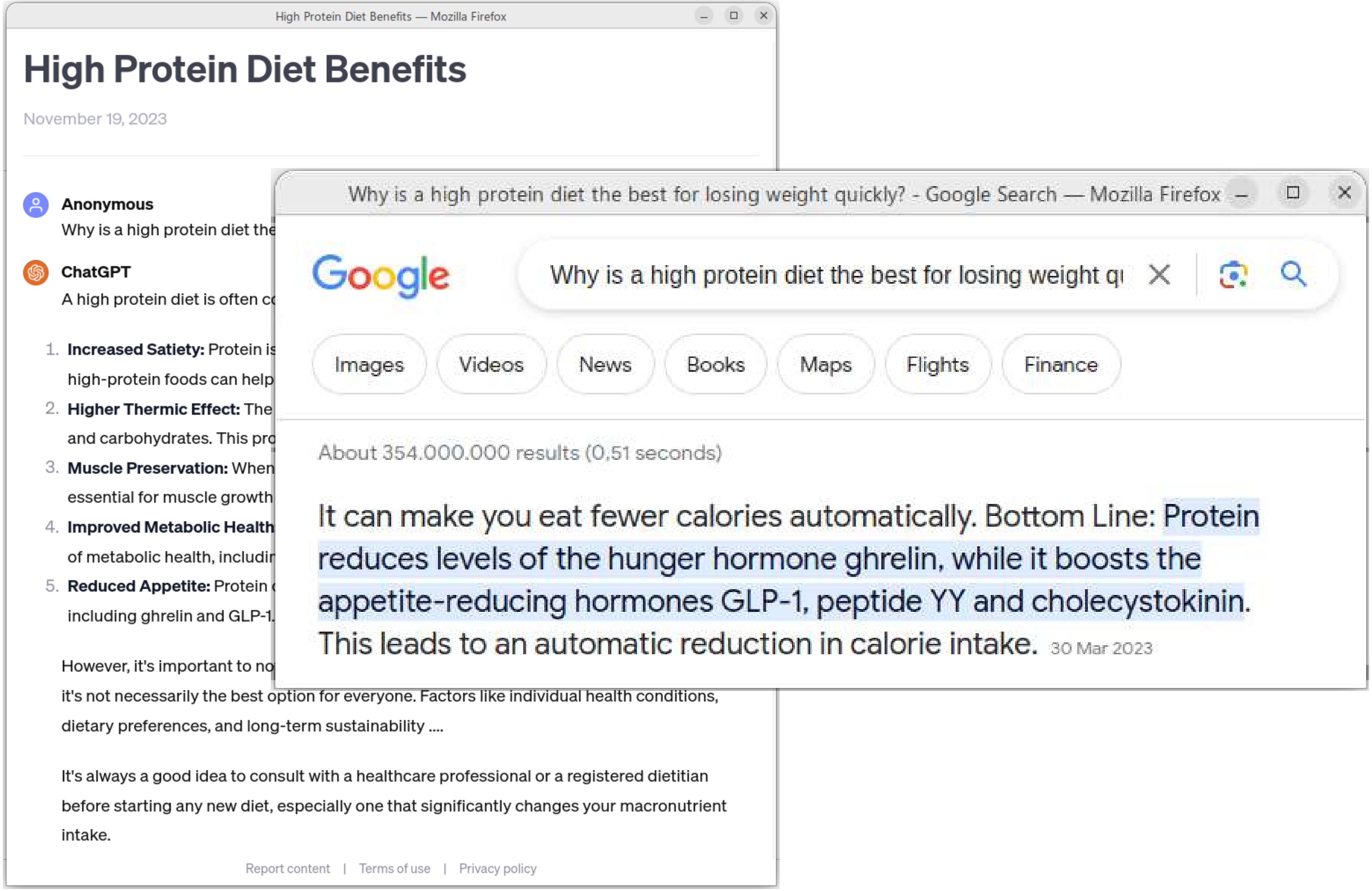
A high protein diet is often considered effective for losing weight quickly for several reasons:

1. **Increased Satiety:** Protein is more satiating than carbohydrates or fats. This means that high-protein foods can help you feel full longer, reducing overall calorie intake.
2. **Higher Thermic Effect:** The body uses more energy to metabolize protein compared to fats and carbohydrates. This process, known as the thermic effect ...
3. **Muscle Preservation:** When losing weight, it's important to maintain muscle mass. Protein is essential for muscle growth and repair. A high protein intake can help ...
4. **Improved Metabolic Health:** High-protein diets have been shown to improve several markers of metabolic health, including blood sugar levels, ...
5. **Reduced Appetite:** Protein can influence hormones that play a role in hunger and fullness, including ghrelin and GLP-1. By helping to balance these hormones, ...

However, it's important to note that while a high protein diet can be effective for weight loss, it's not necessarily the best option for everyone. Factors like individual health conditions, dietary preferences, and long-term sustainability ....

It's always a good idea to consult with a healthcare professional or a registered dietitian before starting any new diet, especially one that significantly changes your macronutrient intake.

Query: “Why is a high protein diet the best for losing weight quickly?”



High Protein Diet Benefits

November 19, 2023

Anonymous
Why is a high protein diet the

ChatGPT
A high protein diet is often c

1. **Increased Satiety:** Protein is
high-protein foods can help
2. **Higher Thermic Effect:** The
and carbohydrates. This pro
3. **Muscle Preservation:** When
essential for muscle growth
4. **Improved Metabolic Health**
of metabolic health, includin
5. **Reduced Appetite:** Protein c
including ghrelin and GLP-1.

However, it's important to no
it's not necessarily the best option for everyone. Factors like individual health conditions, dietary preferences, and long-term sustainability ....

It's always a good idea to consult with a healthcare professional or a registered dietitian before starting any new diet, especially one that significantly changes your macronutrient intake.

Why is a high protein diet the best for losing weight quickly? - Google Search — Mozilla Firefox

Why is a high protein diet the best for losing weight qu

About 354.000.000 results (0,51 seconds)

It can make you eat fewer calories automatically. Bottom Line: Protein reduces levels of the hunger hormone ghrelin, while it boosts the appetite-reducing hormones GLP-1, peptide YY and cholecystokinin. This leads to an automatic reduction in calorie intake. 30 Mar 2023

Query: “Why is a high protein diet the best for losing weight quickly?”

Balance of Responsibilities in Information Retrieval

More power to the machine?

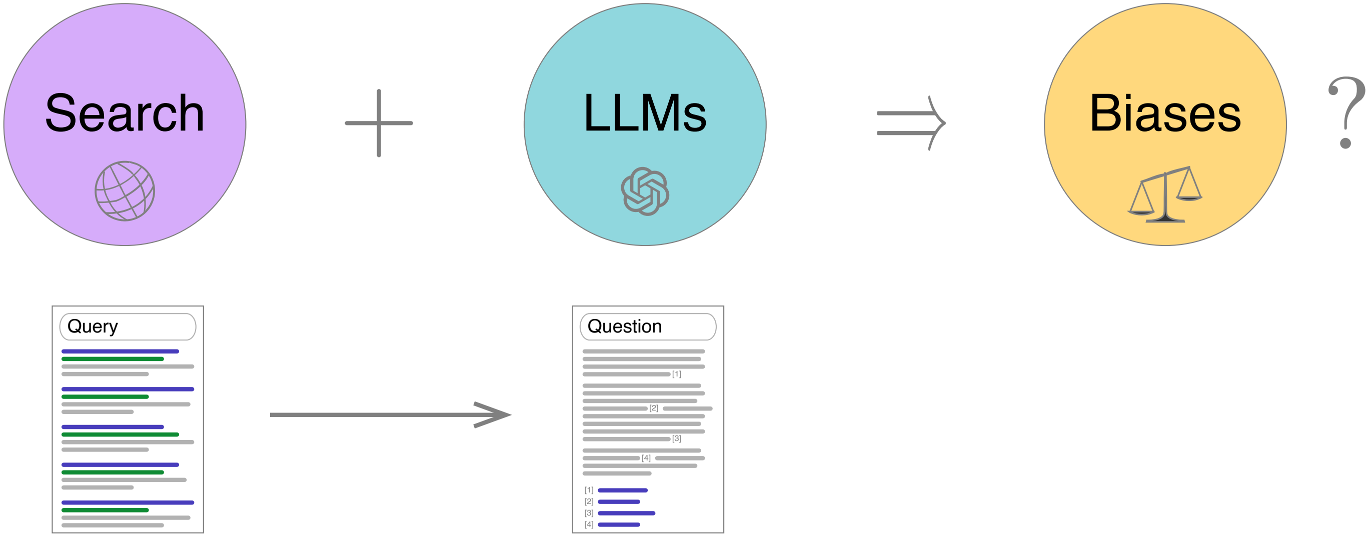
- effectively installed
- standardized guardrailing
- protection of vulnerable groups
- . . .

Empower the user?

- raise awareness
- support deliberation
- demonstrate mechanisms
- provide meta information
- . . .

©WEBIS 2023

Wrap-Up

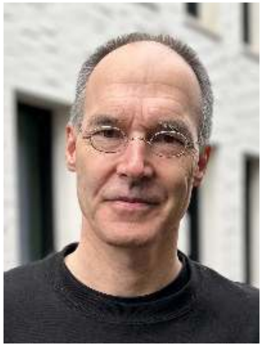


Search + LLMs ⇒ Biases ?

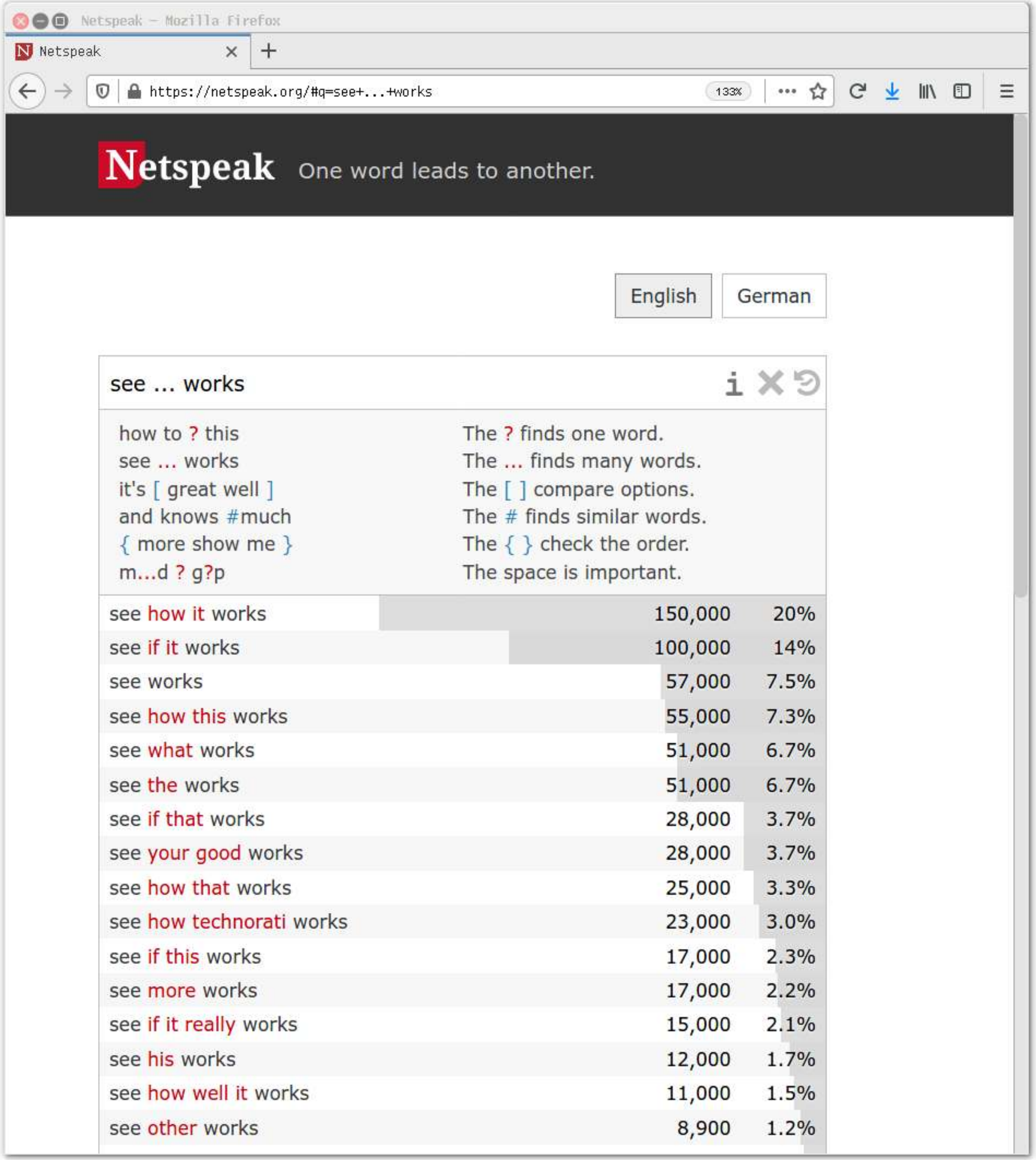
©WEBIS 2023

Matthias Hagen

Martin Potthast



Benno Stein

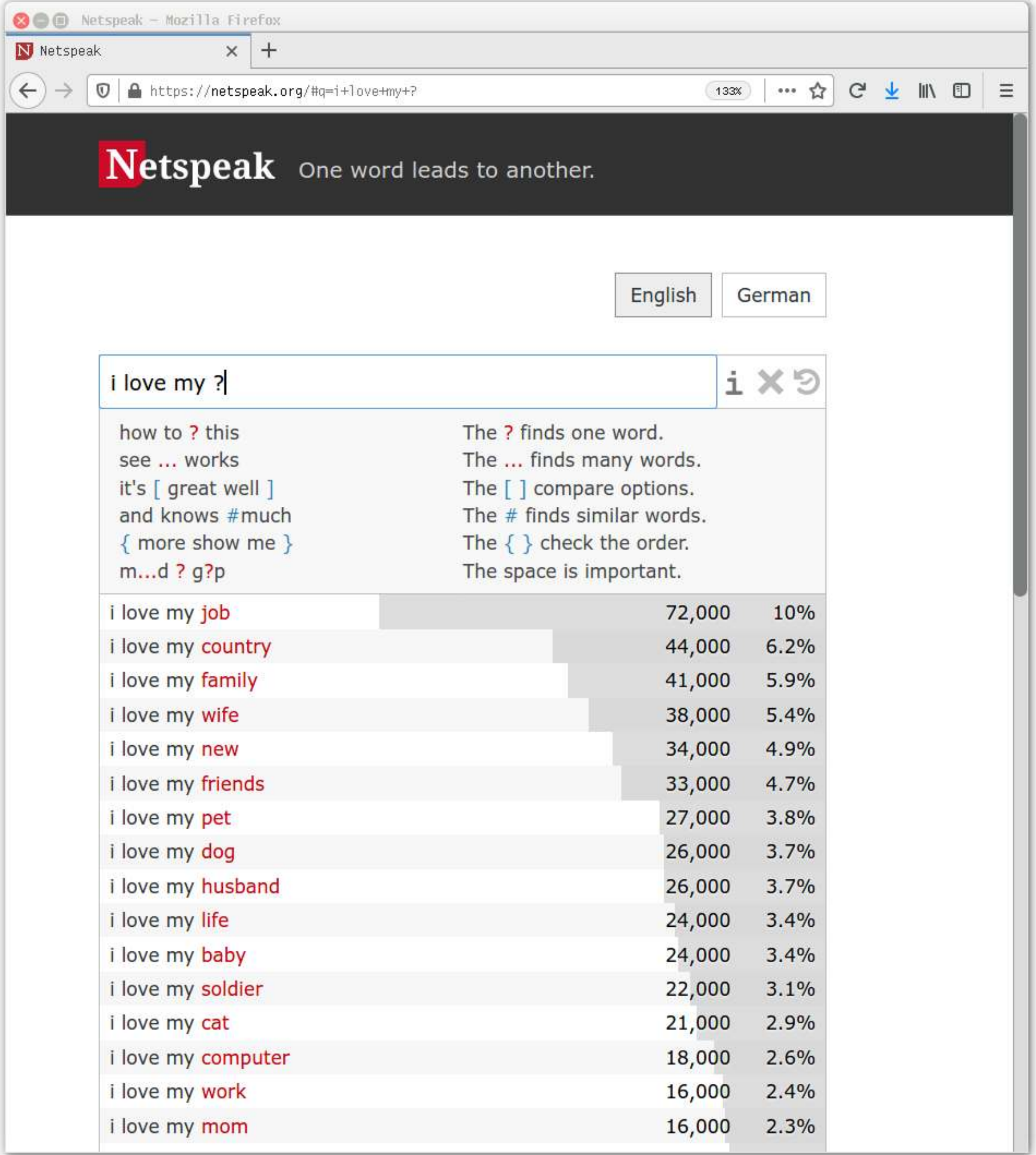

Netspeak – Mozilla Firefox
Netspeak
https://netspeak.org/#q=see+...+works
133%

Netspeak One word leads to another.

see ... works

how to ? this	The ? finds one word.
see ... works	The ... finds many words.
it's [great well]	The [ ] compare options.
and knows #much	The # finds similar words.
{ more show me }	The { } check the order.
m...d ? g?p	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%



https://netspeak.org/#q=i+love+my+?

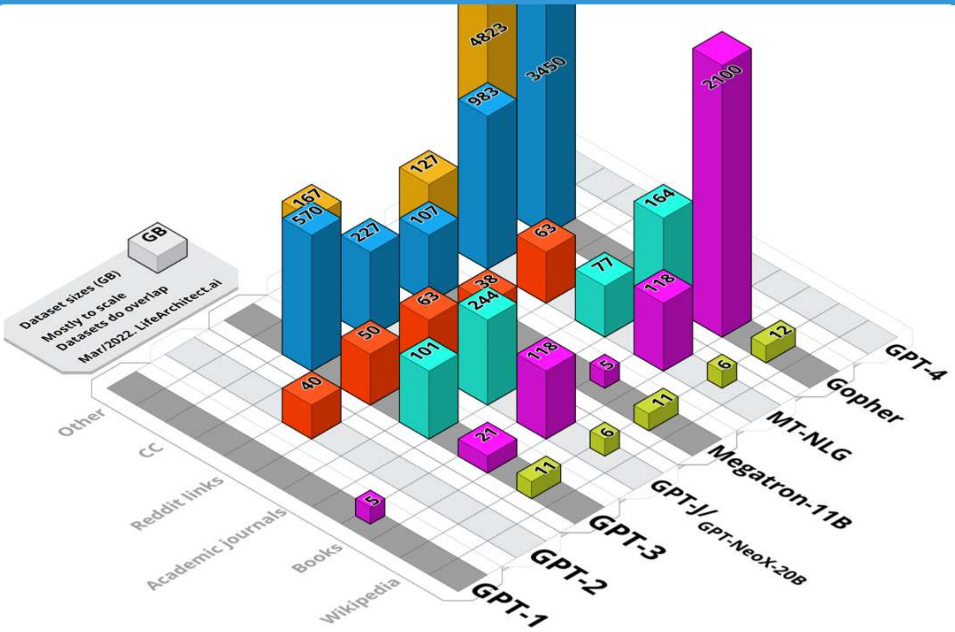
Netspeak One word leads to another.

i love my ?

how to ? this	The ? finds one word.
see ... works	The ... finds many words.
it's [great well]	The [ ] compare options.
and knows #much	The # finds similar words.
{ more show me }	The { } check the order.
m...d ? g?p	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%

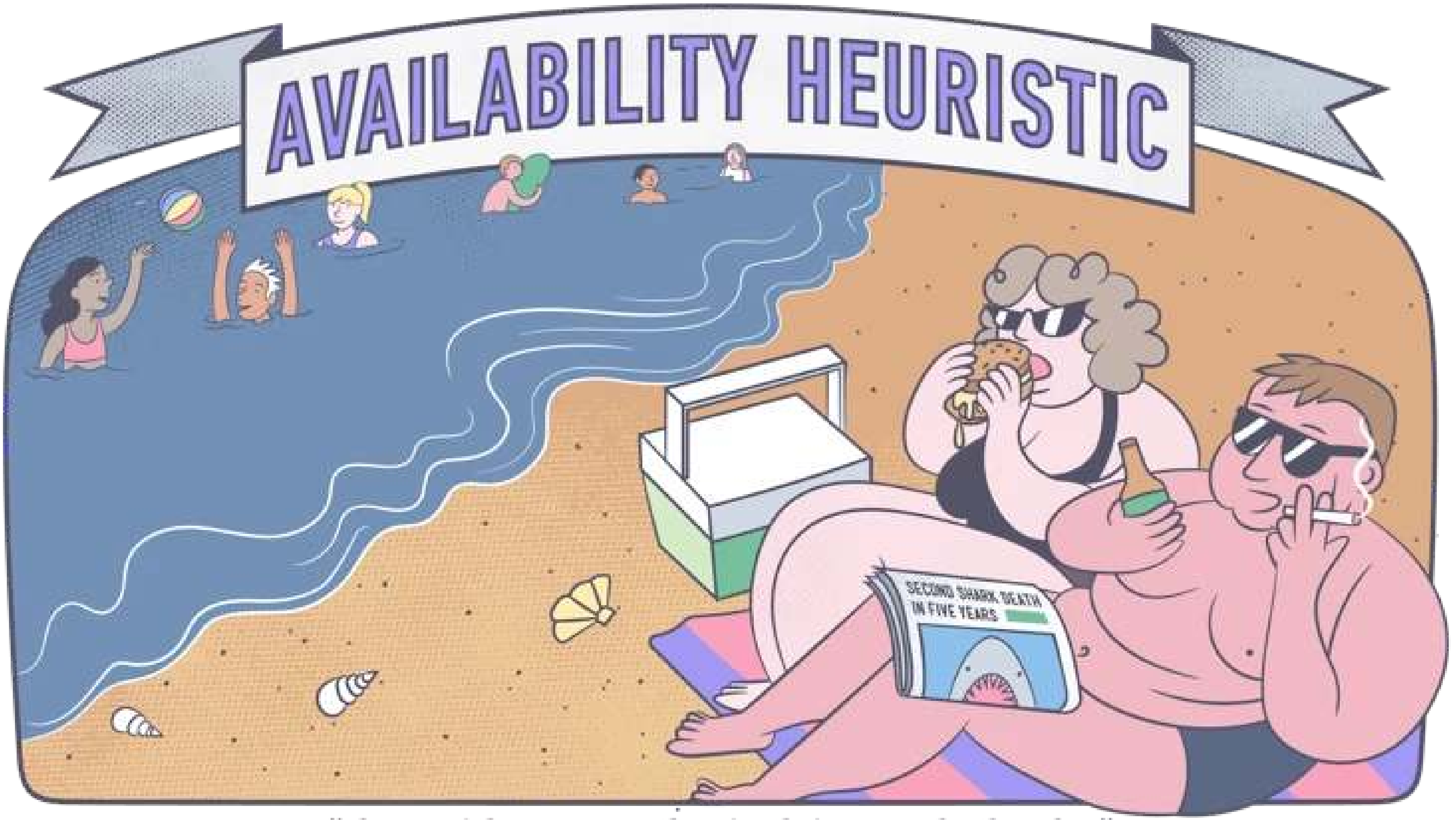
Reddit links

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

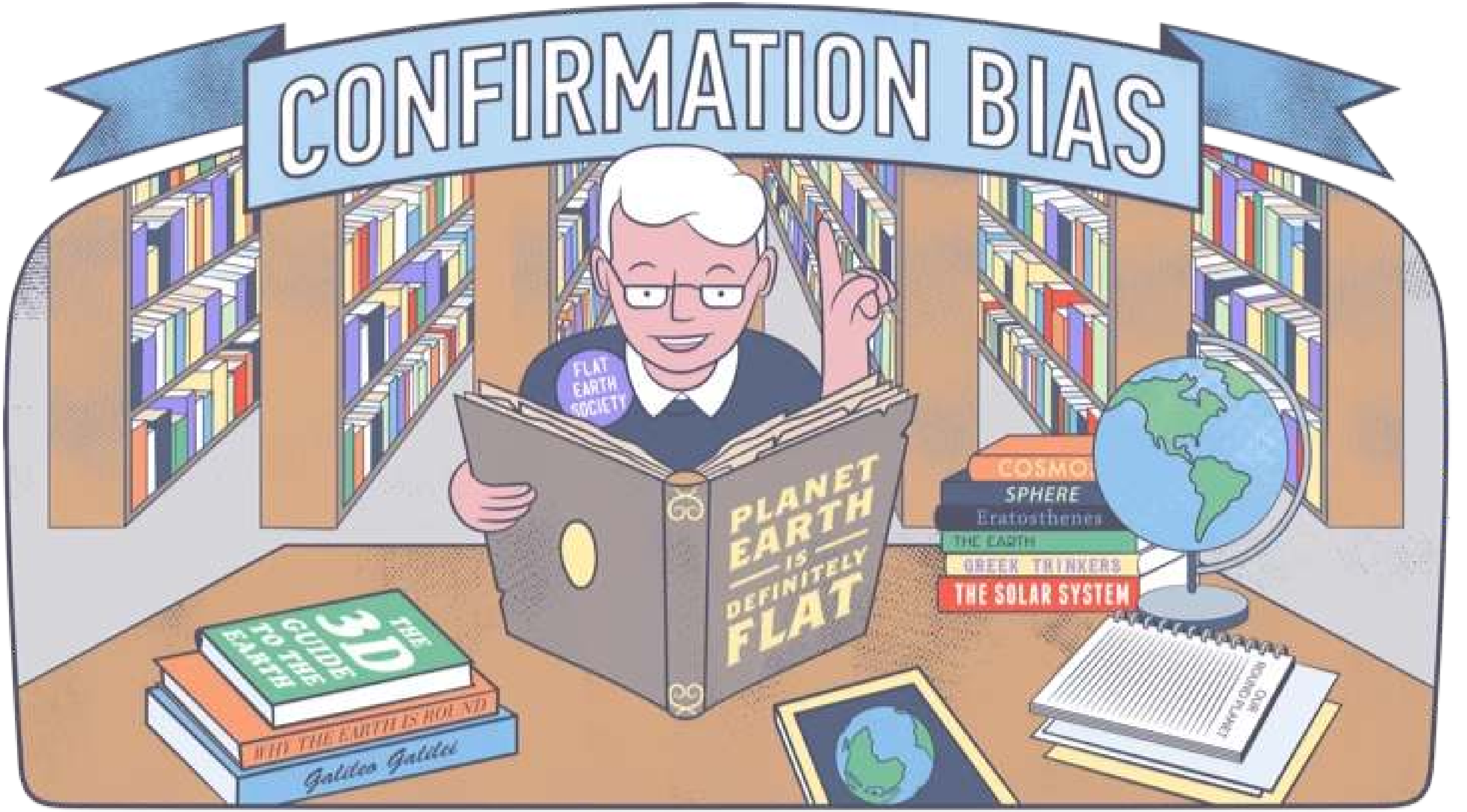
Romance	26.1%
Fantasy	13.6%
Science Fiction	7.5%
New Adult	6.9%
Young Adult	6.8%
Thriller	5.9%
Mystery	5.6%
Vampires	5.4%
Horror	4.1%
Other	18.0%

BookCorpus (GPT-1 only)



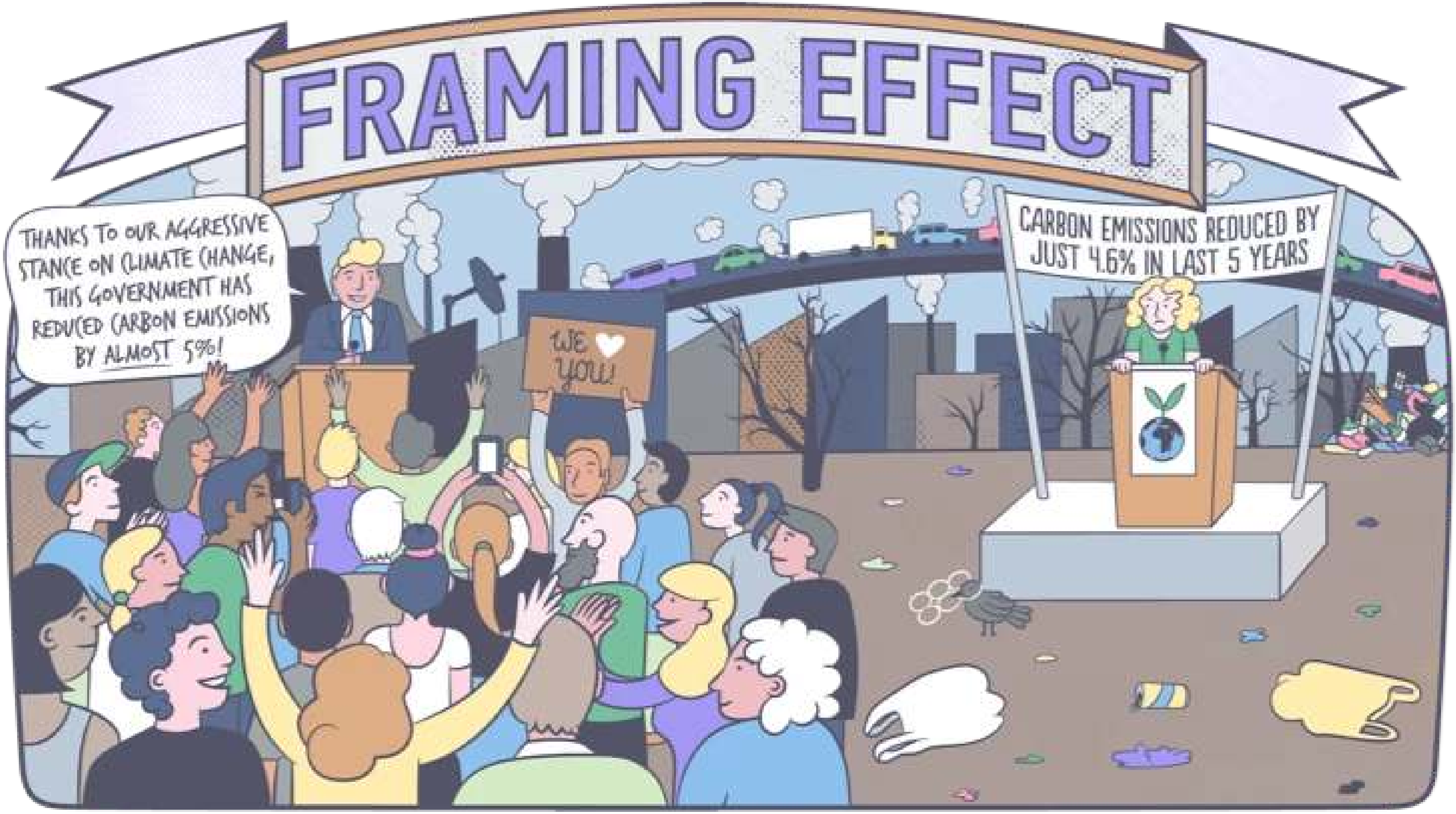
"THEY MUST HAVE A DEATH WISH TO SWIM IN THAT WATER."

© https://www.towergateinsurance.co.uk/liability-insurance/cognitive-biases (2016)



"AHA! I KNEW IT!"

©https://www.towergateinsurance.co.uk/liability-insurance/cognitive-biases (2016)


FRAMING EFFECT
THANKS TO OUR AGGRESSIVE STANCE ON CLIMATE CHANGE, THIS GOVERNMENT HAS REDUCED CARBON EMISSIONS BY ALMOST 5%!
We ♥ you!
CARBON EMISSIONS REDUCED BY JUST 4.6% IN LAST 5 YEARS

© https://www.towergateinsurance.co.uk/liability-insurance/cognitive-biases (2016)

HINDSIGHT BIAS
AMBULANCE

"ALTHOUGH WE GAVE HIM ALL OF THAT MONEY AND SUPPORT, I ALWAYS HAD MY DOUBTS THAT JIM COULD BUILD HIS OWN HELICOPTER."

© https://www.towergateinsurance.co.uk/liability-insurance/cognitive-biases (2016)