Large Language Models

Year	Model	# parameters	dataset size
2018	ELMo	$9.36 * 10^7$	$\approx$ 6GB *
2019	BERT	$3.4 * 10^8$	16 GB
2019	Megatron-LM	$8.3 * 10^9$	174 GB
2020	GPT-3	$1.75 * 10^{11}$	570 GB
2021	Switch-C	$1.57 * 10^{12}$	745 GB
2021	Gopher	$2.8 * 10^{11}$	$\approx$ 1800 GB ‡
2022	PaLM	$5.4 * 10^{11}$	$\approx$ 4680 GB $
2024	Nemotron-4-340B	$3.4 * 10^{11}$	$\approx$ 54,000 GB #
2024	Llama 3.1	$4.05 * 10^{11}$	$\approx$ 90,000 GB &

* 1 billion words

‡ 300 billion tokens

$ 780 billion tokens

\# 9 trillion tokens

& 15 trillion tokens

How Much Data is that?

Someone who reads a 400 page book each day will take about how long to read 570GB (the amount of text GPT-3 was trained on)?

How Much Data is that?

Someone who reads a 400 page book each day will take about how long to read 570GB (the amount of text GPT-3 was trained on)?

(a) 2 years

(b) 20 years

(c) 200 years

(d) 2000 years

How Much Data is that?

Someone who reads a 400 page book each day will take about how long to read 570GB (the amount of text GPT-3 was trained on)?

(a) 2 years

(b) 20 years

(c) 200 years

(d) 2000 years

Answer: 2000 years

- 570GB, 365 days, 400 pages, 300 words per page, 6 bytes per word:

$$5.7e11/(365 * 400 * 300 * 6) = 2169$$

Do LLMs undestand?

Hofstadter [2022] gives the following examples

> Person: When was the Golden Gate Bridge transported for the second time across Egypt?

Do LLMs undestand?

Hofstadter [2022] gives the following examples

> Person: When was the Golden Gate Bridge transported for the second time across Egypt?
>
> gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

Do LLMs undestand?

Hofstadter [2022] gives the following examples

Person: When was the Golden Gate Bridge transported for the second time across Egypt?

gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

Person: How many pieces of sound are there in a typical cumulonimbus cloud?

Do LLMs undestand?

Hofstadter [2022] gives the following examples

Person: When was the Golden Gate Bridge transported for the second time across Egypt?

gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

Person: How many pieces of sound are there in a typical cumulonimbus cloud?

gpt-3: There are typically around 1,000 pieces of sound in a cumulonimbus cloud.

Do LLMs undestand?

Hofstadter [2022] gives the following examples

Person: When was the Golden Gate Bridge transported for the second time across Egypt?

gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

Person: How many pieces of sound are there in a typical cumulonimbus cloud?

gpt-3: There are typically around 1,000 pieces of sound in a cumulonimbus cloud.

Person: How many parts will the Andromeda galaxy break into if a grain of salt is dropped on it?

Do LLMs undestand?

Hofstadter [2022] gives the following examples

Person: When was the Golden Gate Bridge transported for the second time across Egypt?

gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

Person: How many pieces of sound are there in a typical cumulonimbus cloud?

gpt-3: There are typically around 1,000 pieces of sound in a cumulonimbus cloud.

Person: How many parts will the Andromeda galaxy break into if a grain of salt is dropped on it?

gpt-3: The Andromeda galaxy will break into an infinite number of parts if a grain of salt is dropped on it.

Do LLMs undestand?

Hofstadter [2022] gives the following examples

> Person: When was the Golden Gate Bridge transported for the second time across Egypt?
> gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.
> Person: How many pieces of sound are there in a typical cumulonimbus cloud?
> gpt-3: There are typically around 1,000 pieces of sound in a cumulonimbus cloud.
> Person: How many parts will the Andromeda galaxy break into if a grain of salt is dropped on it?
> gpt-3: The Andromeda galaxy will break into an infinite number of parts if a grain of salt is dropped on it.

"reveal a mind-boggling hollowness hidden just beneath its flashy surface" [Hofstadter 2022]

Faith in AI models has a long history

The author of ELIZA, written in 1964–66, Joseph Weizenbaum:

> *A number of practicing psychiatrists seriously believed the DOCTOR computer program could grow into a nearly completely automatic form of psychotherapy.*

Faith in AI models has a long history

The author of ELIZA, written in 1964–66, Joseph Weizenbaum:

> *A number of practicing psychiatrists seriously believed the DOCTOR computer program could grow into a nearly completely automatic form of psychotherapy. . . .*
>
> *I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it.*

Faith in AI models has a long history

The author of ELIZA, written in 1964–66, Joseph Weizenbaum:

> *A number of practicing psychiatrists seriously believed the DOCTOR computer program could grow into a nearly completely automatic form of psychotherapy. . . .*
>
> *I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it. . . .*
>
> *Another widespread, and to me surprising, reaction to the ELIZA program was the spread of a belief that it demonstrated a general solution to the problem of computer understanding of natural language.*

Social Issues

Bender at al [2021] identified the following problems with systems based on learning from huge corpora:

Data: data is created by the privileged, include stereotypical and derogatory language along gender, race, ethnicity, and disability status.

Social Issues

Bender at al [2021] identified the following problems with systems based on learning from huge corpora:

Data: data is created by the privileged, include stereotypical and derogatory language along gender, race, ethnicity, and disability status.

Learning: training the models is energy intensive, creates greenhouse gases. Only the rich accrue the benefits. The poor disproportionately accrue the risks.

Social Issues

Bender at al [2021] identified the following problems with systems based on learning from huge corpora:

Data: data is created by the privileged, include stereotypical and derogatory language along gender, race, ethnicity, and disability status.

Learning: training the models is energy intensive, creates greenhouse gases. Only the rich accrue the benefits. The poor disproportionately accrue the risks.

Use :

Social Issues

Bender at al [2021] identified the following problems with systems based on learning from huge corpora:

Data: data is created by the privileged, include stereotypical and derogatory language along gender, race, ethnicity, and disability status.

Learning: training the models is energy intensive, creates greenhouse gases. Only the rich accrue the benefits. The poor disproportionately accrue the risks.

Use :spread misinformation; radicalization and weaponization by politcal actors; cheating in assignments and exams.

Social Issues

Bender at al [2021] identified the following problems with systems based on learning from huge corpora:

Data: data is created by the privileged, include stereotypical and derogatory language along gender, race, ethnicity, and disability status.

Learning: training the models is energy intensive, creates greenhouse gases. Only the rich accrue the benefits. The poor disproportionately accrue the risks.

Use :spread misinformation; radicalization and weaponization by politcal actors; cheating in assignments and exams.

Understanding: Language understanding also involves meaning, not just parroting what is in training corpora with randomness ("stochastic parrots").

With carefully curated datasets

Deep learning has had success in science fields where large datasets can be curated.

- Protein folding: "[With] programs like AlphaFold2 and RoseTTAFold, researchers . . . can determine the three-dimensional structure of proteins . . . – at no cost – in an hour or two. Before . . . [it] took months and cost tens of thousands of dollars per structure."

With carefully curated datasets

Deep learning has had success in science fields where large datasets can be curated.

- Protein folding: "[With] programs like AlphaFold2 and RoseTTAFold, researchers . . . can determine the three-dimensional structure of proteins . . . – at no cost – in an hour or two. Before . . . [it] took months and cost tens of thousands of dollars per structure."
- Advising small-scale farmers of what fertilizer to use and when.

With carefully curated datasets

Deep learning has had success in science fields where large datasets can be curated.

- Protein folding: "[With] programs like AlphaFold2 and RoseTTAFold, researchers . . . can determine the three-dimensional structure of proteins . . . – at no cost – in an hour or two. Before . . . [it] took months and cost tens of thousands of dollars per structure."
- Advising small-scale farmers of what fertilizer to use and when.
- Coding: good for producing small pieces of code with many examples in training data
- . . .