Part VI

Retrieval-free Approaches

No explicit retriever?

- Key question: can we use **pre-trained language models** to act as "knowledge storage"?

- Instead of explicitly storing all the text and searching among their *dense* or *sparse* representations, can we query the LMs to obtain the answer directly?

- The LMs were pre-trained on Wikipedia (and other textual corpora) so they should be able to memorize a fair amount of information.

LMs as KBs?

Barack Obama was born in Honolulu.

Petroni et al., 2019. Language Models as Knowledge Bases?

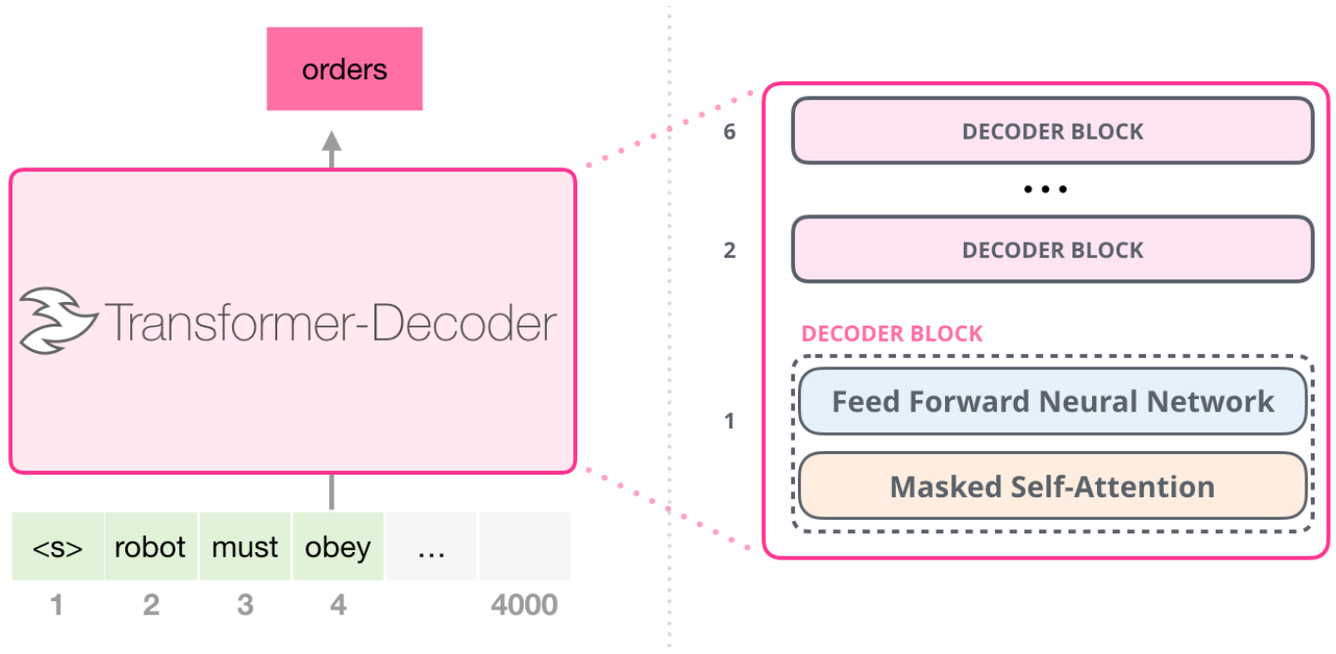
GPT-2 [Rardford et al., 2019]

- GPT-2 is a *very large*, transformer-based language model trained on *a massive dataset*.

48 layers, hidden size 1600, 1.5B parameters

WebText: 8 million documents, excluding Wikipedia (!)

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$



Radford et al., 2019. Language Models are Unsupervised Multitask Learners
http://jalammar.github.io/illustrated-gpt2/

GPT-2: zero-shot QA

Evaluated on Natural Questions and **no training at all**

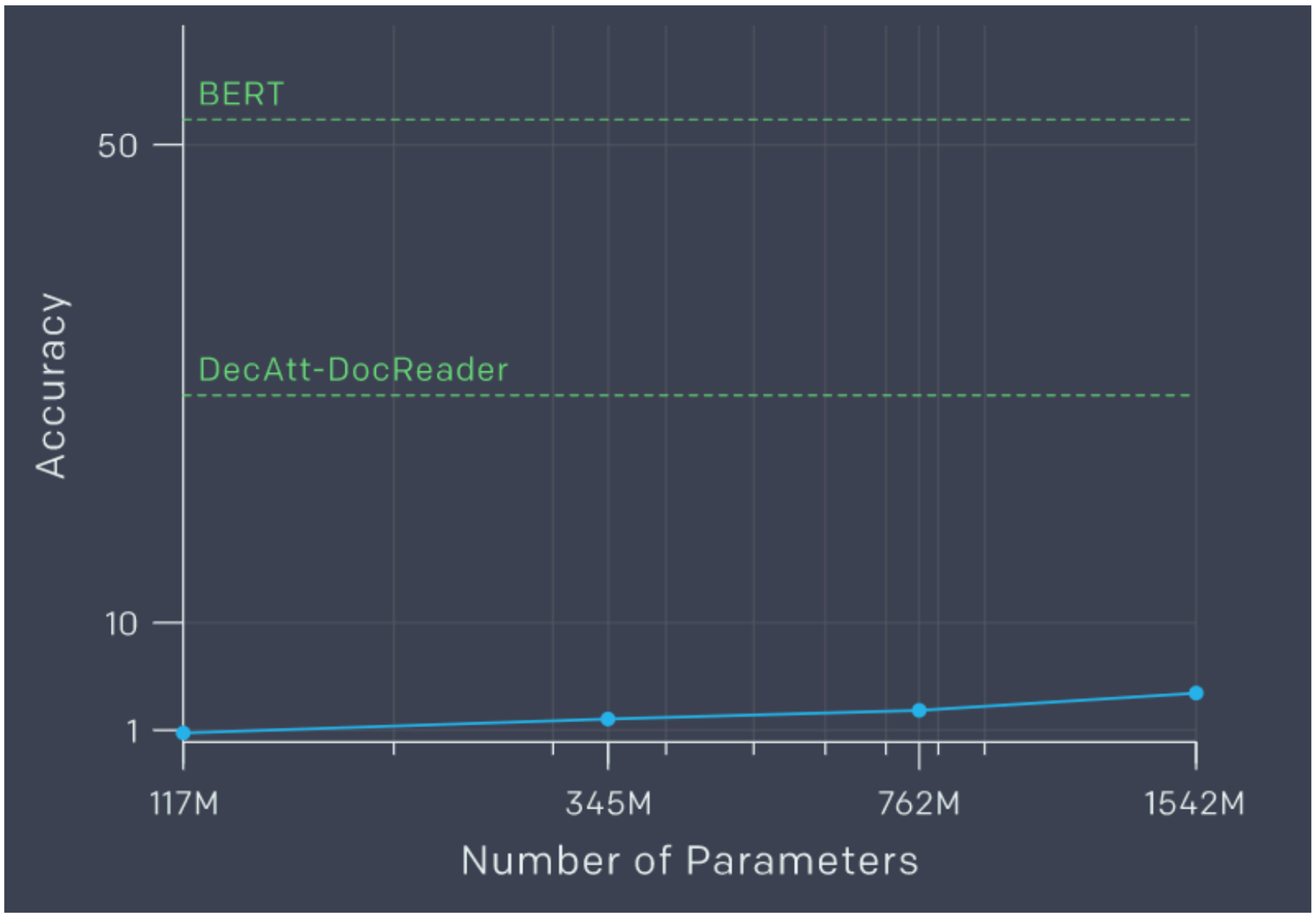
63.1% on the **1%** of questions it is most confident in

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Radford et al., 2019. Language Models are Unsupervised Multitask Learners

GPT-2: zero-shot QA

Evaluated on Natural Questions and **no training at all**



4% accuracy:
Much much worse than supervised systems

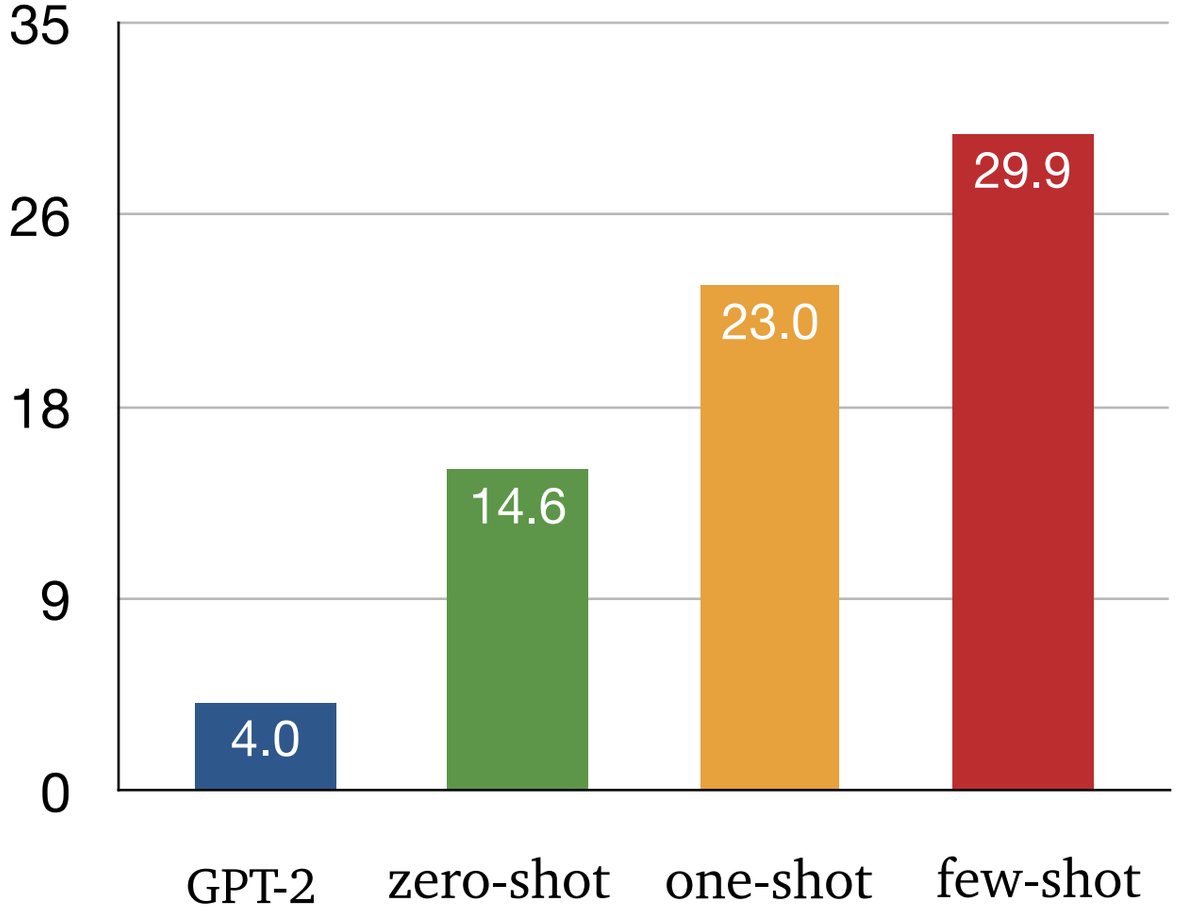
Radford et al., 2019. Language Models are Unsupervised Multitask Learners

GPT-3: Few-shot Learner [Brown et al., 2020]

96 layers, hidden size 12288, **175B** parameters

Larger corpora: Common Crawl + WebText + Books + English Wikipedia

Evaluated on Natural Questions:



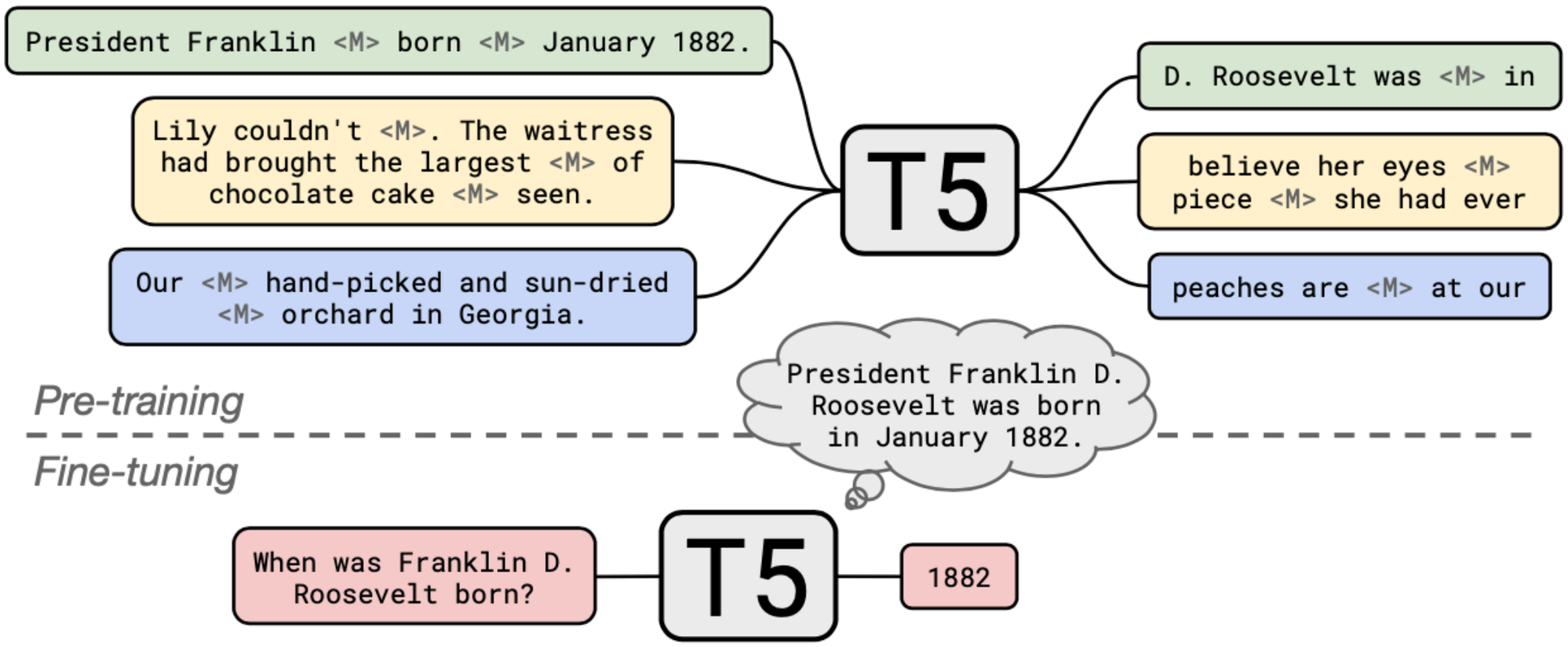
Few-shot learner

- No weight updates
- Only augment the prompt with **K** (question, answer) pairs as "demonstration"
- One-shot setting is a special case when only **one** example is given.

Brown et al., 2020. Language Models are Few-Shot Learners

T5: Fine-tuning leads to improved performance
[Roberts et al., 2020]

Text-to-Text Transfer Transformer [Raffel et al., 2019]



*: Pre-trained on a multitask mixture including an **unsupervised "span corruption" task** on unlabeled text as well as supervised translation, summarization, classification, and reading comprehension tasks

Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

T5: Fine-tuning leads to improved performance

		Natural Questions	WebQuestions	TriviaQA
		NQ	WQ	TQA
	Chen et al. (2017)	–	20.7	–
	Lee et al. (2019)	33.3	36.4	47.1
	Min et al. (2019a)	28.1	–	50.9
	Min et al. (2019b)	31.8	31.6	55.4
	Asai et al. (2019)	32.6	–	–
	Ling et al. (2020)	–	–	35.7
	Guu et al. (2020)	40.4	40.7	–
	Févry et al. (2020)	–	–	53.4
	Karpukhin et al. (2020)	**41.5**	42.4	57.9
220M	T5-Base	27.0	29.1	29.1
770M	T5-Large	29.8	32.2	35.9
3B	T5-3B	32.1	34.9	43.4
11B	T5-11B	34.5	37.4	50.1
	T5-11B + SSM	36.6	**44.7**	**60.5**

BERT-base = 110M parameters

SSM: salient span masking, pre-training data proposed in REALM

Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Summary

- Large language models pre-trained on unstructured text can attain competitive results in open-domain QA without accessing external knowledge.
- The performance is largely impacted by the model size. A 11B T5 model is able to match the performance with DPR with 3 BERT-base models (220M parameters each).

NeurIPS'20 EfficientQA Competition

How should we store the "knowledge" for use by our open-domain QA system?

Passages, databases or parameters of NNs?

We are looking for systems (evaluated on Natural Questions):

- Most accurate self-contained QA system under 6Gb
- Most accurate self-contained QA system under 500Mb
- Smallest self-contained QA system that achieves 25% accuracy
- Most accurate QA system with no constraints

Baselines: TF-IDF, DPR and T5

Important Dates

July, 2020	Leaderboard launched.
October 14, 2020	Leaderboard frozen.
November 14, 2020	Human evaluation completed and winners announced.
December 11-12, 2020	NeurIPS workshop and human-computer competition (held virtually).

https://efficientqa.github.io/

No explicit retriever?

- Key question: can we use **pre-trained language models** to act as "knowledge storage"?

- Instead of explicitly storing all the text and searching among their *dense* or *sparse* representations, can we query the LMs to obtain the answer directly?

- The LMs were pre-trained on Wikipedia (and other textual corpora) so they should be able to memorize a fair amount of information.

LMs as KBs?

Barack Obama was born in Honolulu.

Petroni et al. 2019. Language Models as Knowledge Bases?

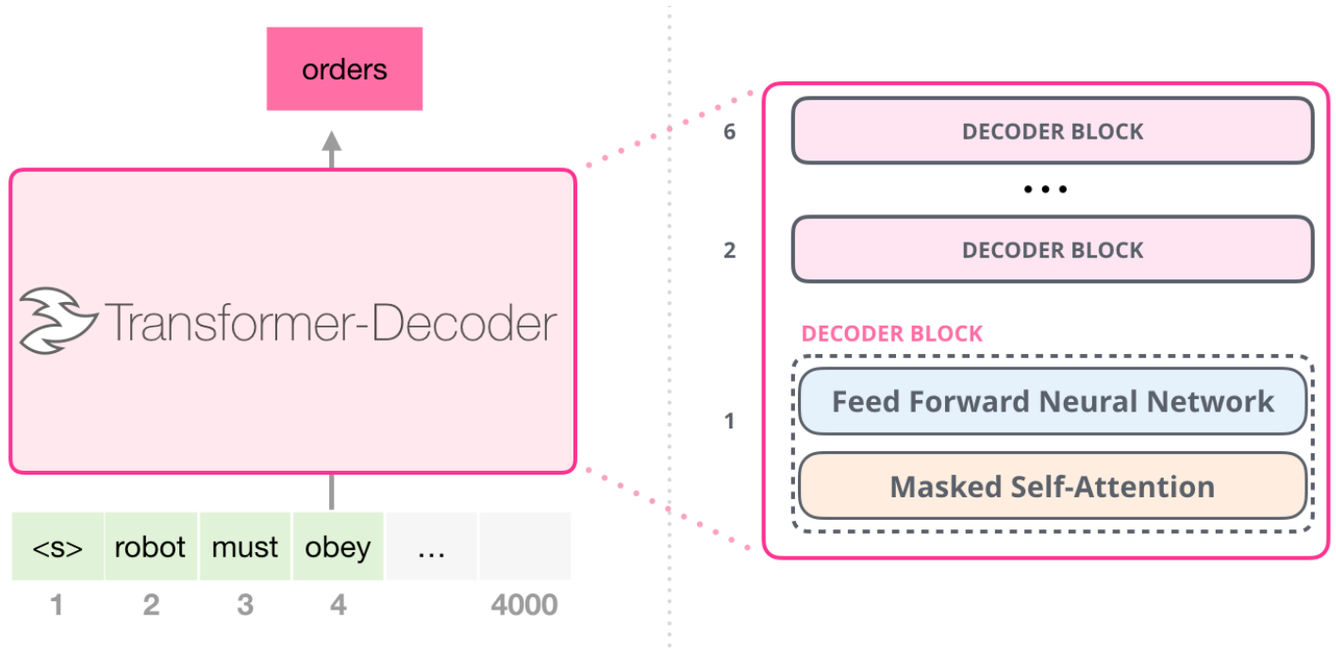
GPT-2 [Rardford et al., 2019]

- GPT-2 is a *very large*, transformer-based language model trained on *a massive dataset*.

48 layers, hidden size 1600, 1.5B parameters

WebText: 8 million documents, excluding Wikipedia (!)

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$



Radford et al. 2019. Language Models are Unsupervised Multitask Learners
http://jalammar.github.io/illustrated-gpt2/

GPT-2: zero-shot QA

Evaluated on Natural Questions and **no training at all**

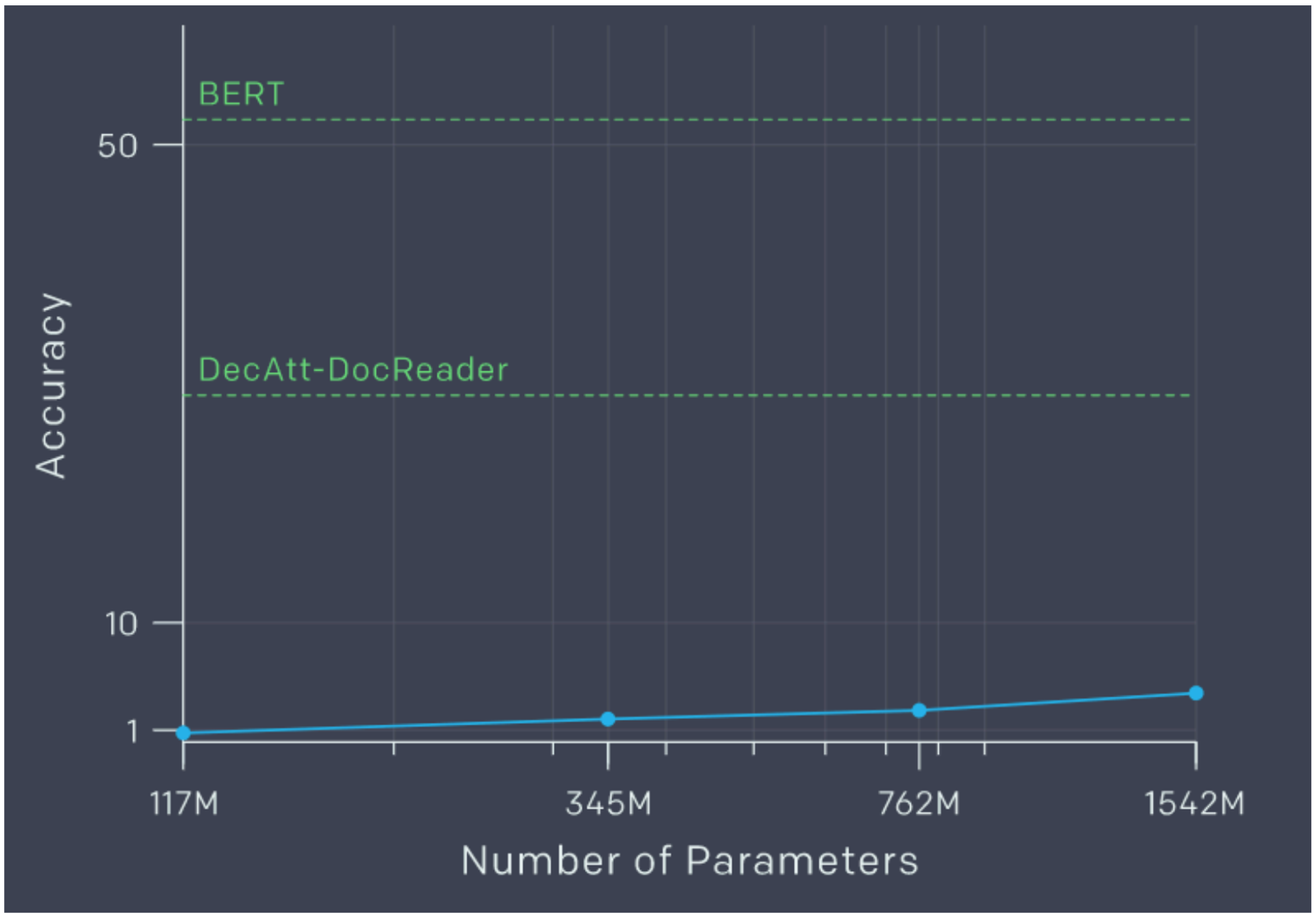
63.1% on the **1%** of questions it is most confident in

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Radford et al. 2019. Language Models are Unsupervised Multitask Learners

GPT-2: zero-shot QA

Evaluated on Natural Questions and **no training at all**



4% accuracy:
Much much worse than supervised systems

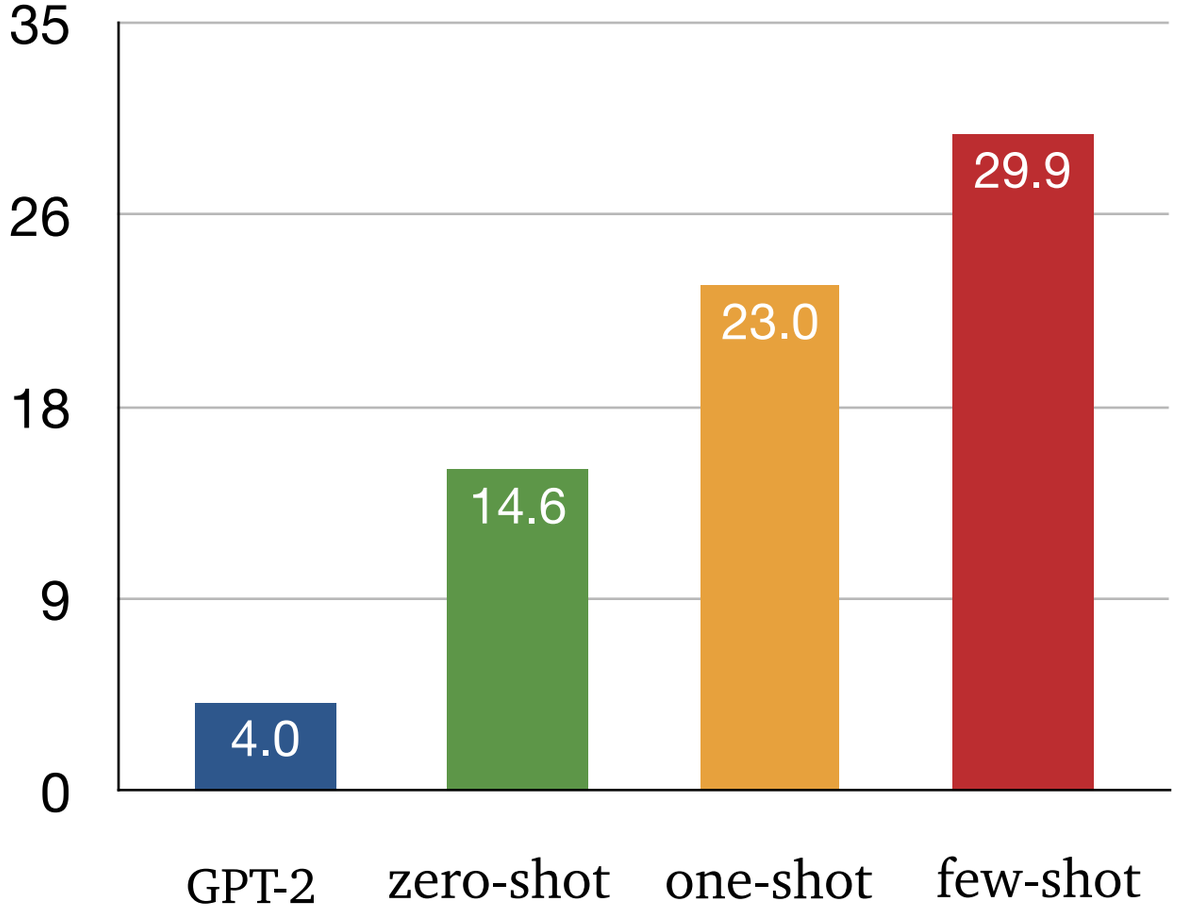
Radford et al. 2019. Language Models are Unsupervised Multitask Learners

GPT-3: Few-shot Learner [Brown et al., 2020]

96 layers, hidden size 12288, **175B** parameters

Larger corpora: Common Crawl + WebText + Books + English Wikipedia

Evaluated on Natural Questions:



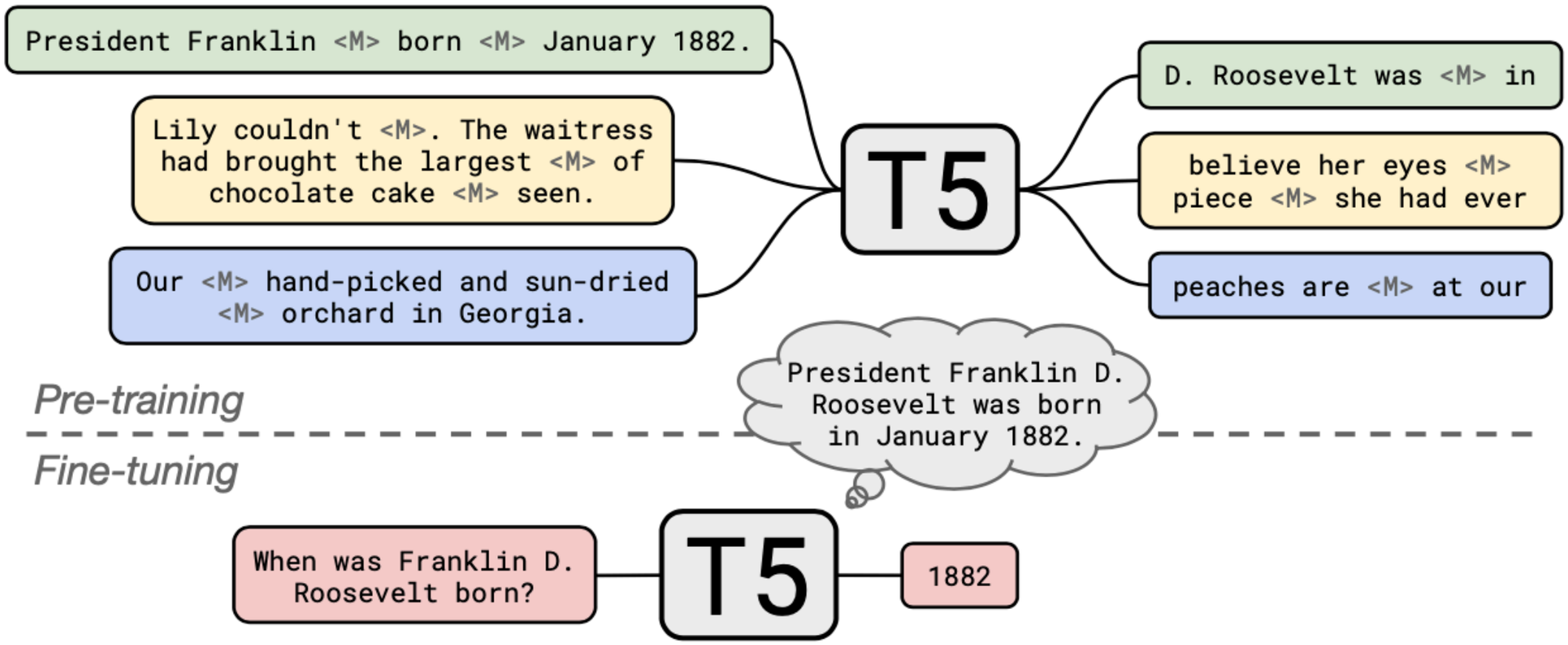
Few-shot learner

- No weight updates
- Only augment the prompt with **K** (question, answer) pairs as "demonstration"
- One-shot setting is a special case when only **one** example is given.

Brown et al. 2020. Language Models are Few-Shot Learners

T5: Fine-tuning leads to improved performance
[Roberts et al., 2020]

Text-to-Text Transfer Transformer [Raffel et al., 2019]



*: Pre-trained on a multitask mixture including an **unsupervised "span corruption" task** on unlabeled text as well as supervised translation, summarization, classification, and reading comprehension tasks

Roberts et al. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

T5: Fine-tuning leads to improved performance

		Natural Questions	WebQuestions	TriviaQA
		NQ	WQ	TQA
	Chen et al. (2017)	–	20.7	–
	Lee et al. (2019)	33.3	36.4	47.1
	Min et al. (2019a)	28.1	–	50.9
	Min et al. (2019b)	31.8	31.6	55.4
	Asai et al. (2019)	32.6	–	–
	Ling et al. (2020)	–	–	35.7
	Guu et al. (2020)	40.4	40.7	–
	Févry et al. (2020)	–	–	53.4
	Karpukhin et al. (2020)	**41.5**	42.4	57.9
220M	T5-Base	27.0	29.1	29.1
770M	T5-Large	29.8	32.2	35.9
3B	T5-3B	32.1	34.9	43.4
11B	T5-11B	34.5	37.4	50.1
	T5-11B + SSM	36.6	**44.7**	**60.5**

BERT-base = 110M parameters

SSM: salient span masking, pre-training data proposed in REALM

Roberts et al. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Summary

- Large language models pre-trained on unstructured text can attain competitive results in open-domain QA without accessing external knowledge.
- The performance is largely impacted by the model size. A 11B T5 model is able to match the performance with DPR with 3 BERT-base models (220M parameters each).

NeurIPS'20 EfficientQA Competition

How should we store the "knowledge" for use by our open-domain QA system?

Passages, databases or parameters of NNs?

We are looking for the systems (evaluated on Natural Questions):

- Most accurate self-contained QA system under 6Gb
- Most accurate self-contained QA system under 500Mb
- Smallest self-contained QA system that achieves 25% accuracy
- Most accurate QA system with no constraints

Baselines: TF-IDF, DPR and T5

Important Dates

July, 2020	Leaderboard launched.
October 14, 2020	Leaderboard frozen.
November 14, 2020	Human evaluation completed and winners announced.
December 11-12, 2020	NeurIPS workshop and human-computer competition (held virtually).

https://efficientqa.github.io/