Language Models are Few-Shot Learners (GPT-3)

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et al.
OpenAI
Max Mathys

Language Modeling

Sequence of tokens (words): $(t_1, \ldots, t_n)$

Model: $P(t_1, \ldots, t_n)$

$$P(t_1, \ldots, t_n) = \prod_{i=1}^{n} P(t_i \mid t_1, \ldots, t_{i-1})$$

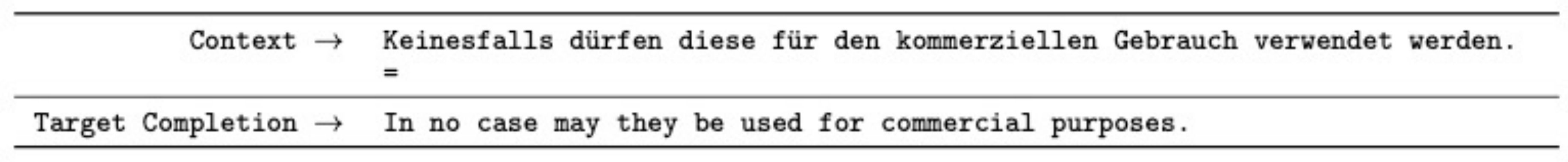
k-gram: $P(t_1, \ldots, t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-k}, \ldots, t_{i-1})$

Predict: $P(t_i \mid t_{i-k}, \ldots, t_{i-1})$

$$\arg\max_t P(t \mid t_{n-k}, \ldots, t_n)$$

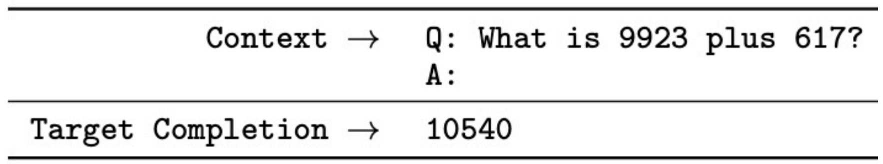
Natural Language Processing (NLP)

Translation



Context →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden. =
Target Completion →	In no case may they be used for commercial purposes.

Question-answering



Context →	Q: What is 9923 plus 617? A:
Target Completion →	10540

Cloze tasks

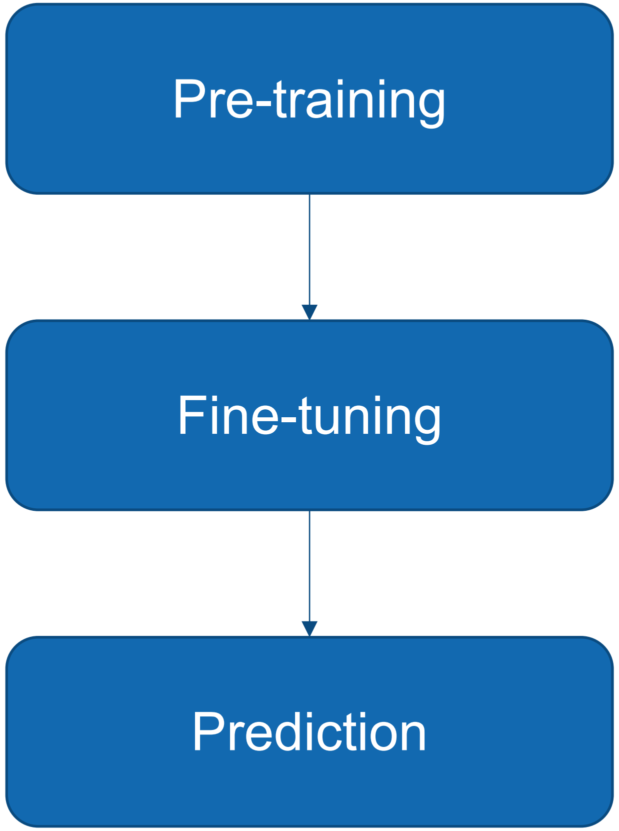
Today, I went to the ________ and bought some milk and eggs. I knew it was going to rain, but I forgot to take my ________, and ended up getting wet on the way.

Outline

1. Problem Statement and Main Contribution
2. Details of the Contribution
3. Experiments
4. Discussion

Problem Statement and Main Contribution

Traditional Language Modeling Approach

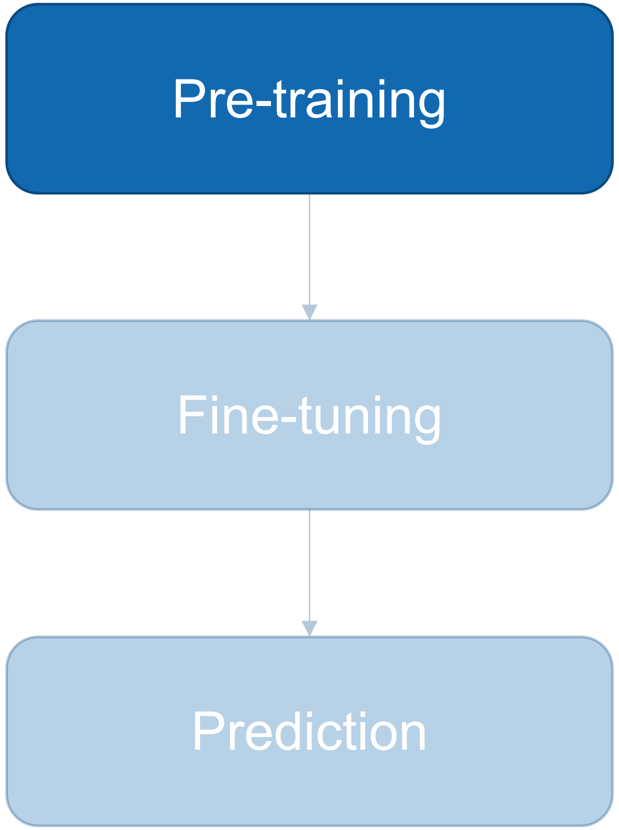


Pre-training

Fine-tuning

Prediction

Traditional Language Modeling Approach



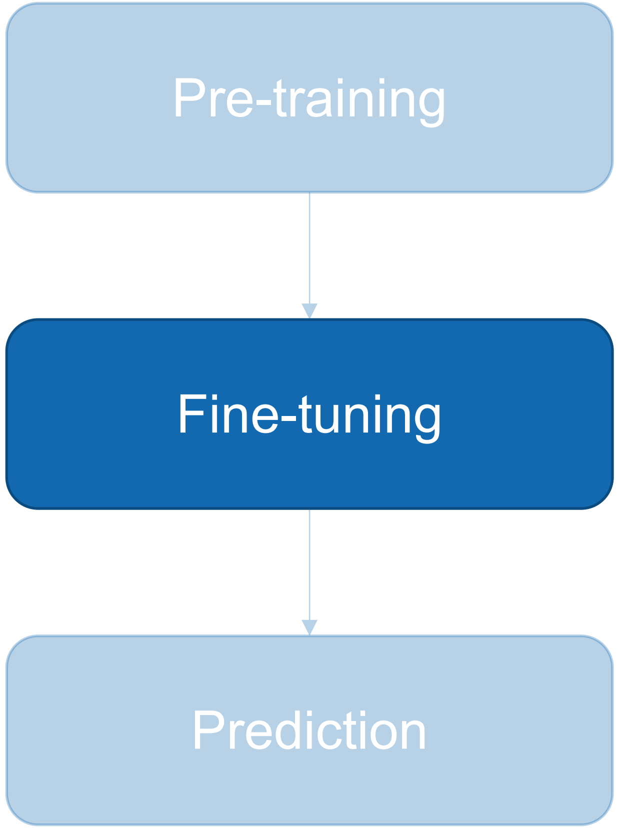
Pre-training

Fine-tuning

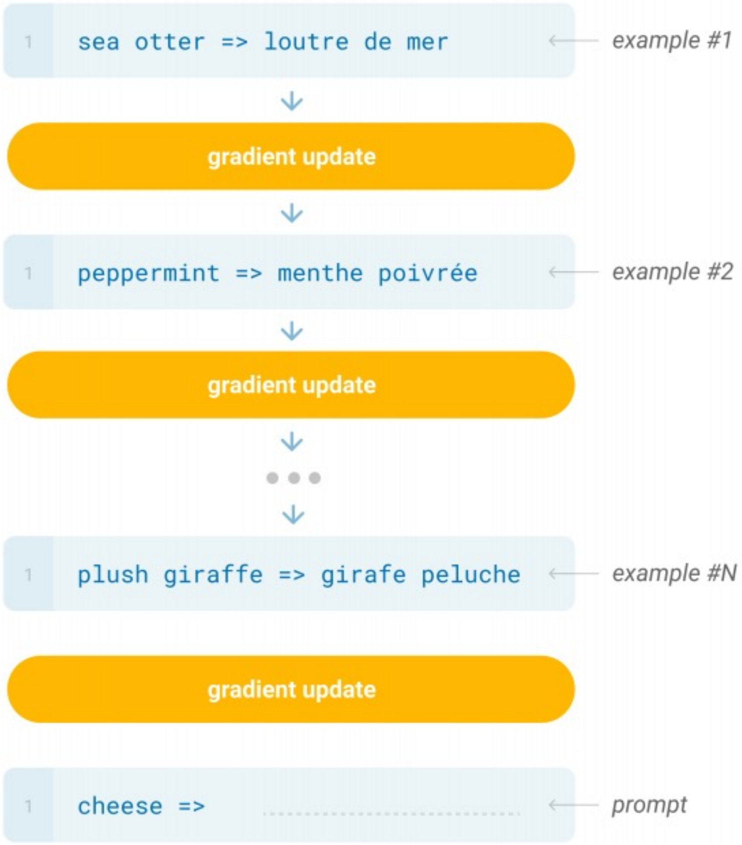
Prediction

The model is trained in an **unsupervised** way on large task-agnostic body of text.

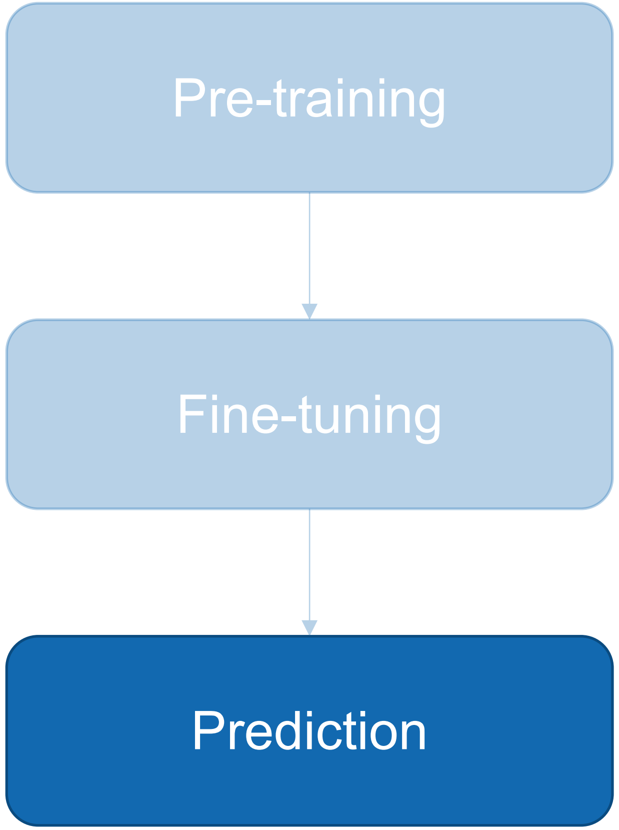
Traditional Language Modeling Approach



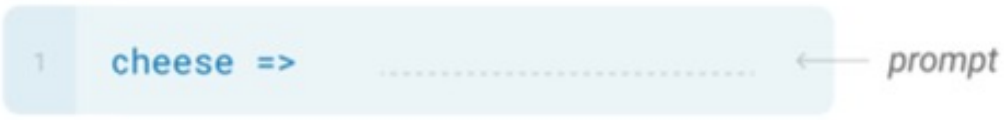
The model is trained in a **supervised** way on a large task-specific dataset.



Traditional Language Modeling Approach



Prompt

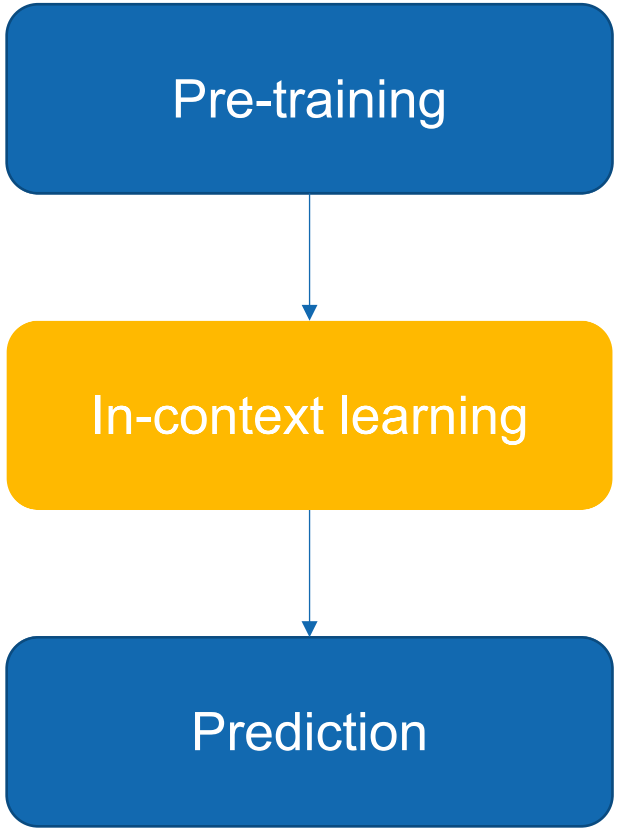


```
cheese =>
```

Traditional Approach: Drawbacks

- **Inefficient:** need a large dataset of labelled examples for every new task
- Humans do not require large supervised datasets to learn most language tasks

Meta-learning

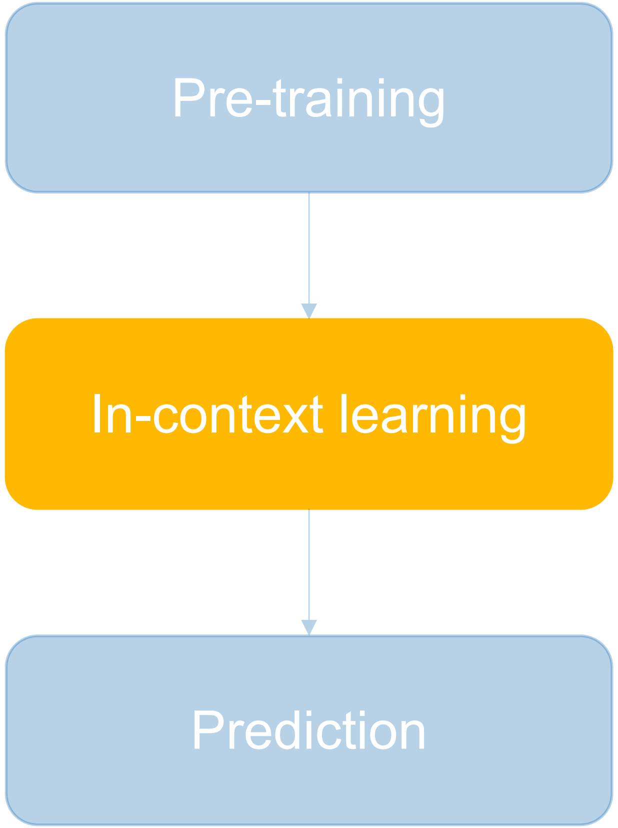


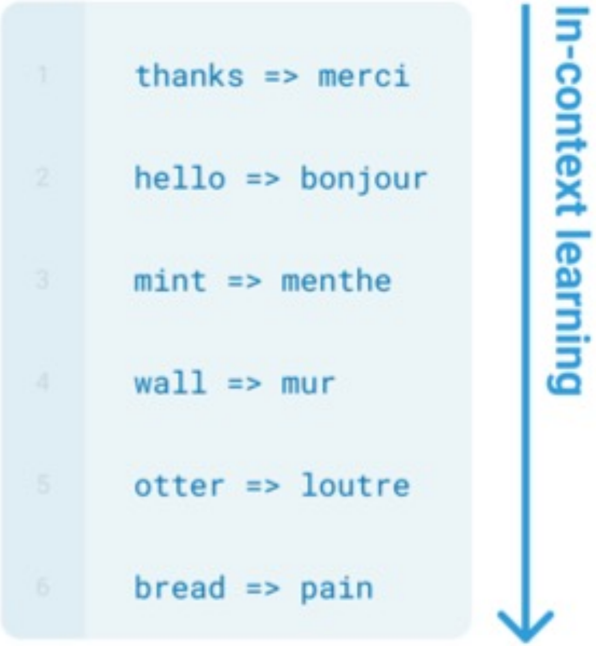
Pre-training

In-context learning

Prediction

Meta-learning

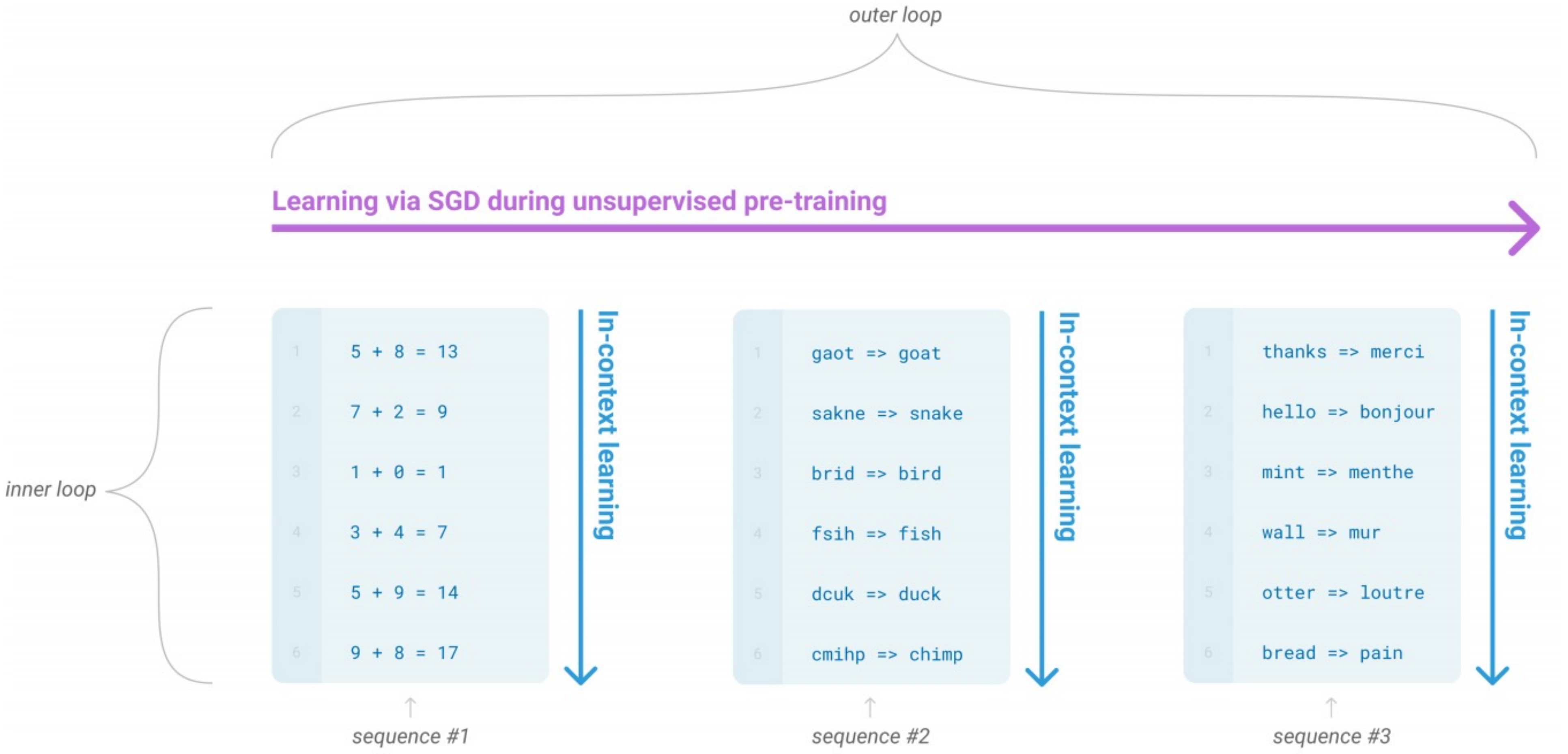




```
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
```

No task-specific datasets are needed
Model can quickly adapt to any new tasks

Meta-learning



```
5 + 8 = 13
7 + 2 = 9
1 + 0 = 1
3 + 4 = 7
5 + 9 = 14
9 + 8 = 17
```

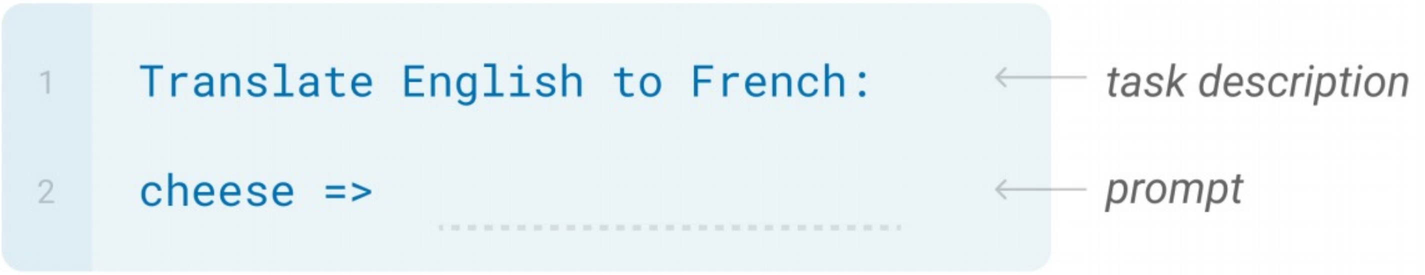
```
gaot => goat
sakne => snake
brid => bird
fsih => fish
dcuk => duck
cmihp => chimp
```

```
thanks => merci
hello => bonjour
mint => menthe
wall => mur
otter => loutre
bread => pain
```

Meta-learning: Advantages

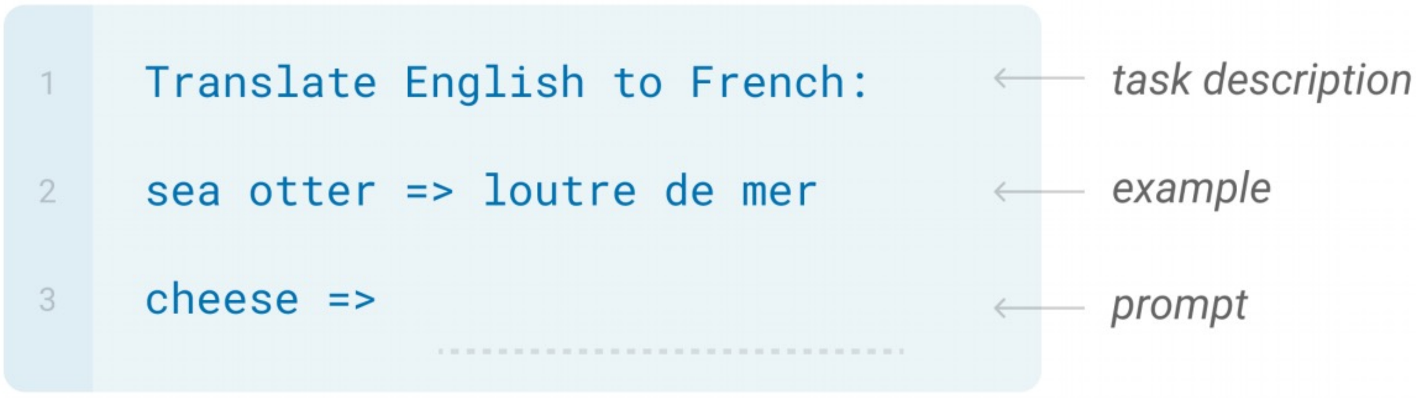
- No need for a large dataset of labelled examples for every new task
- Model can perform any new task
- Models with Meta-learning are (usually) not as accurate as fine-tuned models

Zero-shot learning



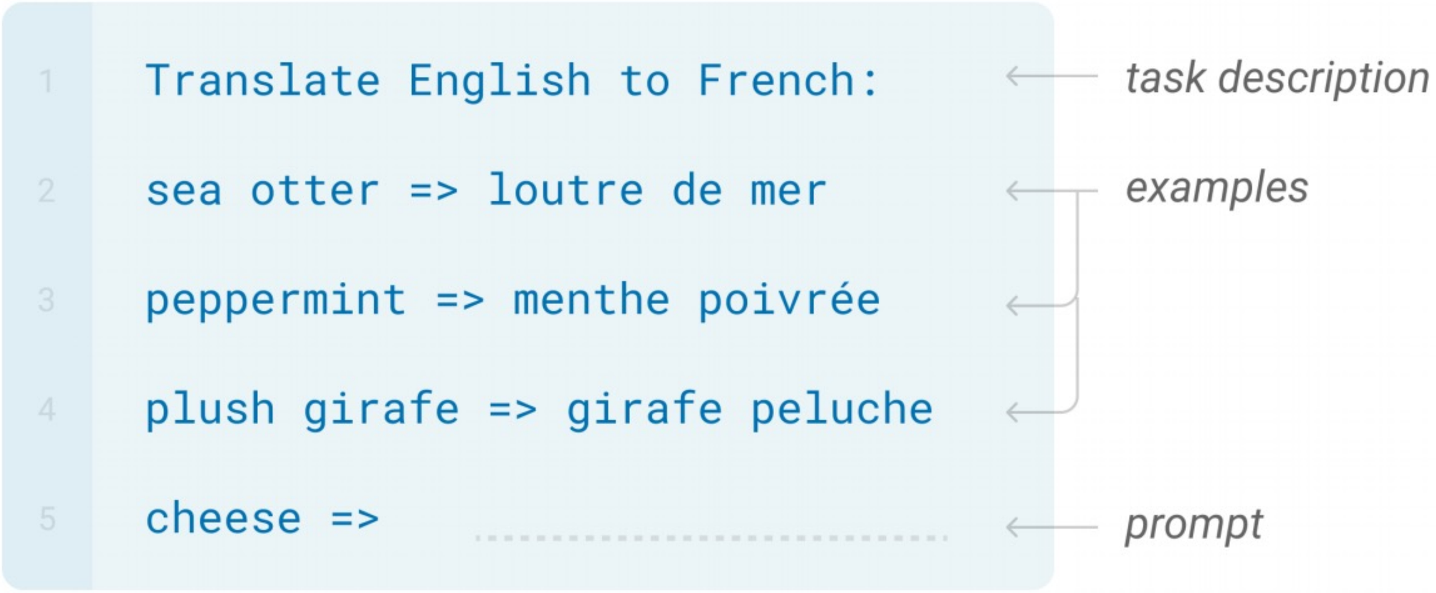
```
Translate English to French:
cheese =>
```

One-shot learning



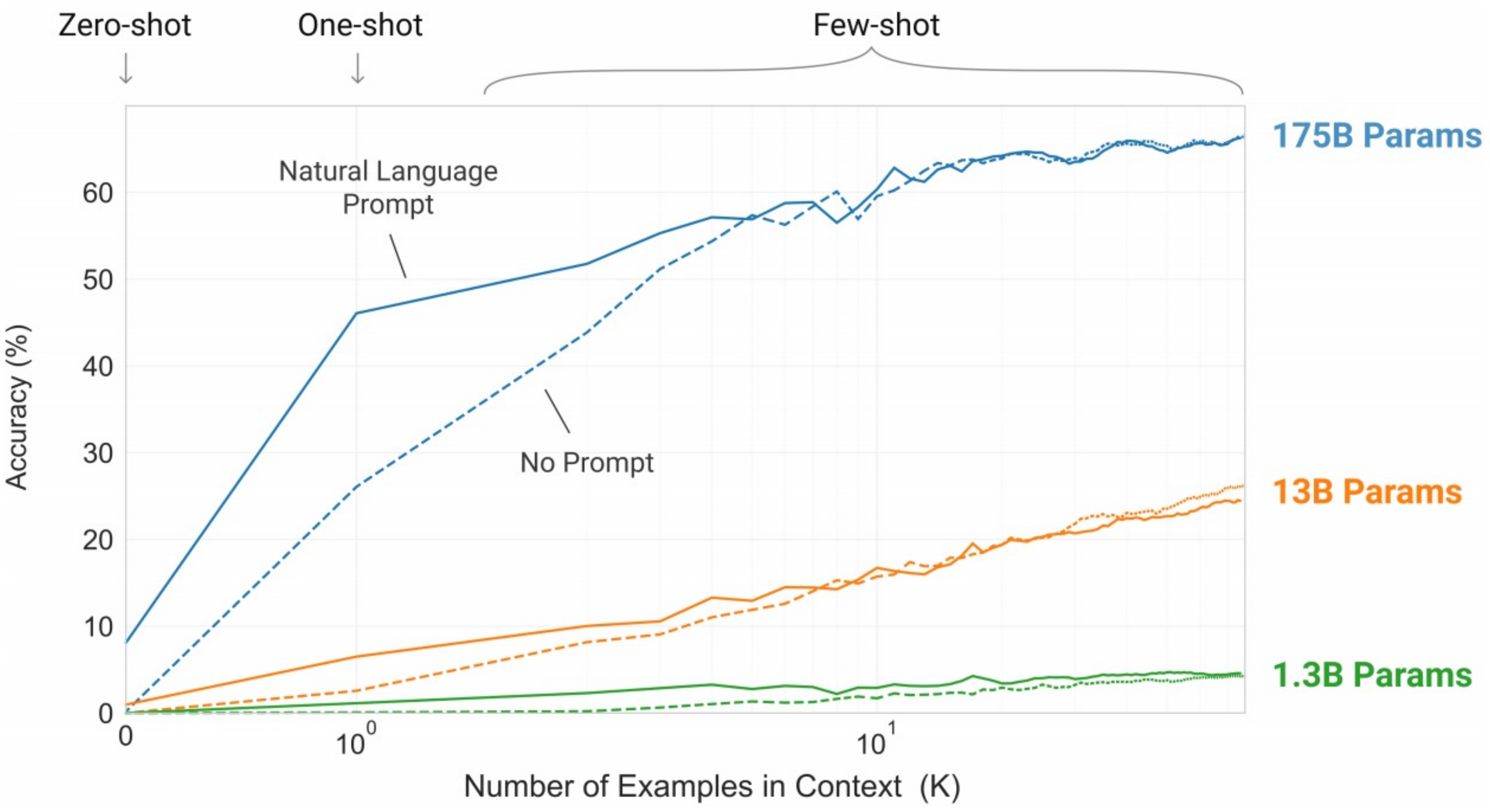
```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot learning

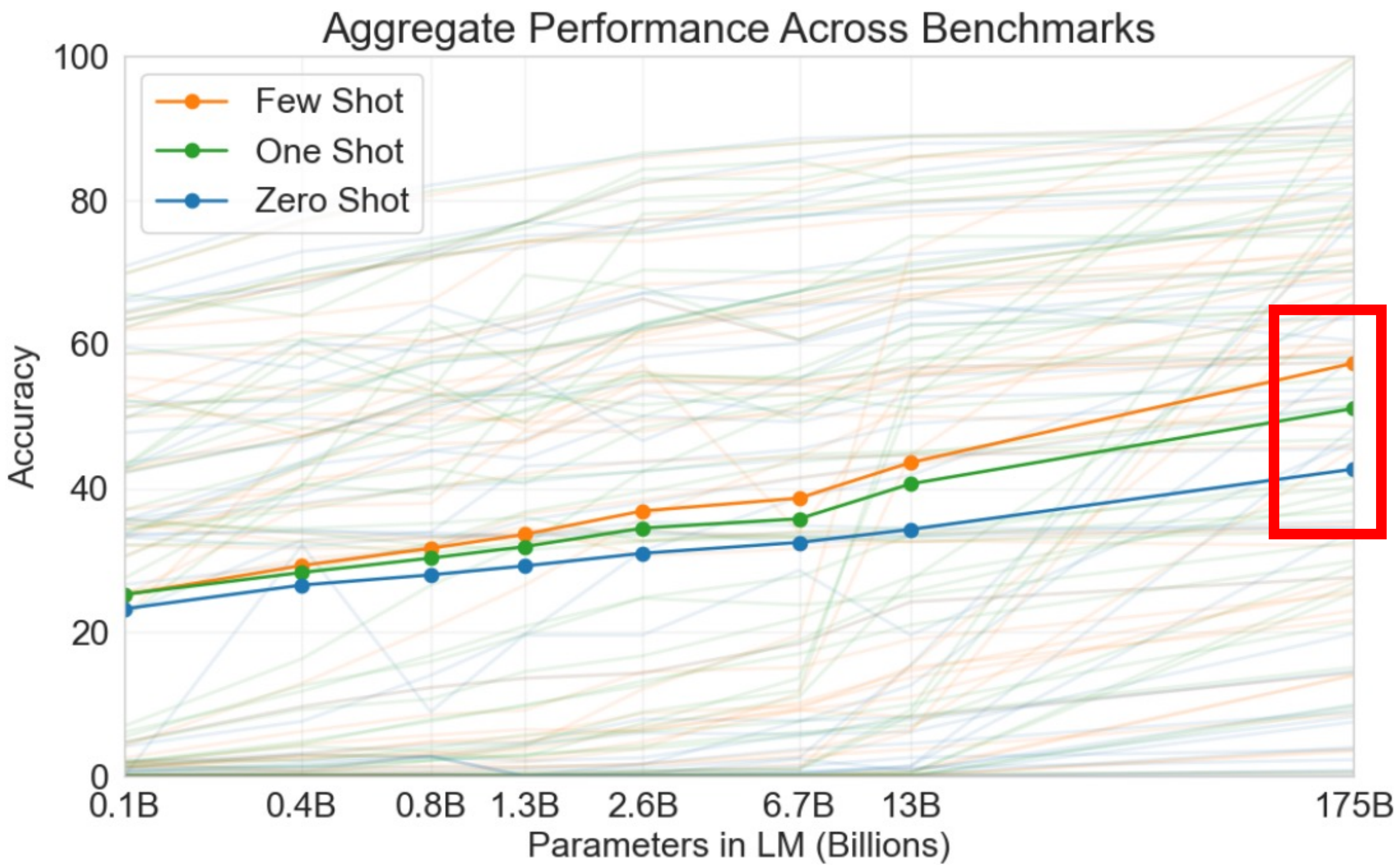


```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Effect of few-shot learning



Main scientific contribution



Details of the Contribution

Transformer Architecture

- Relatively new architecture, released in 2017
- Compared to RNNs and LSTMs, has multiple advantages

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

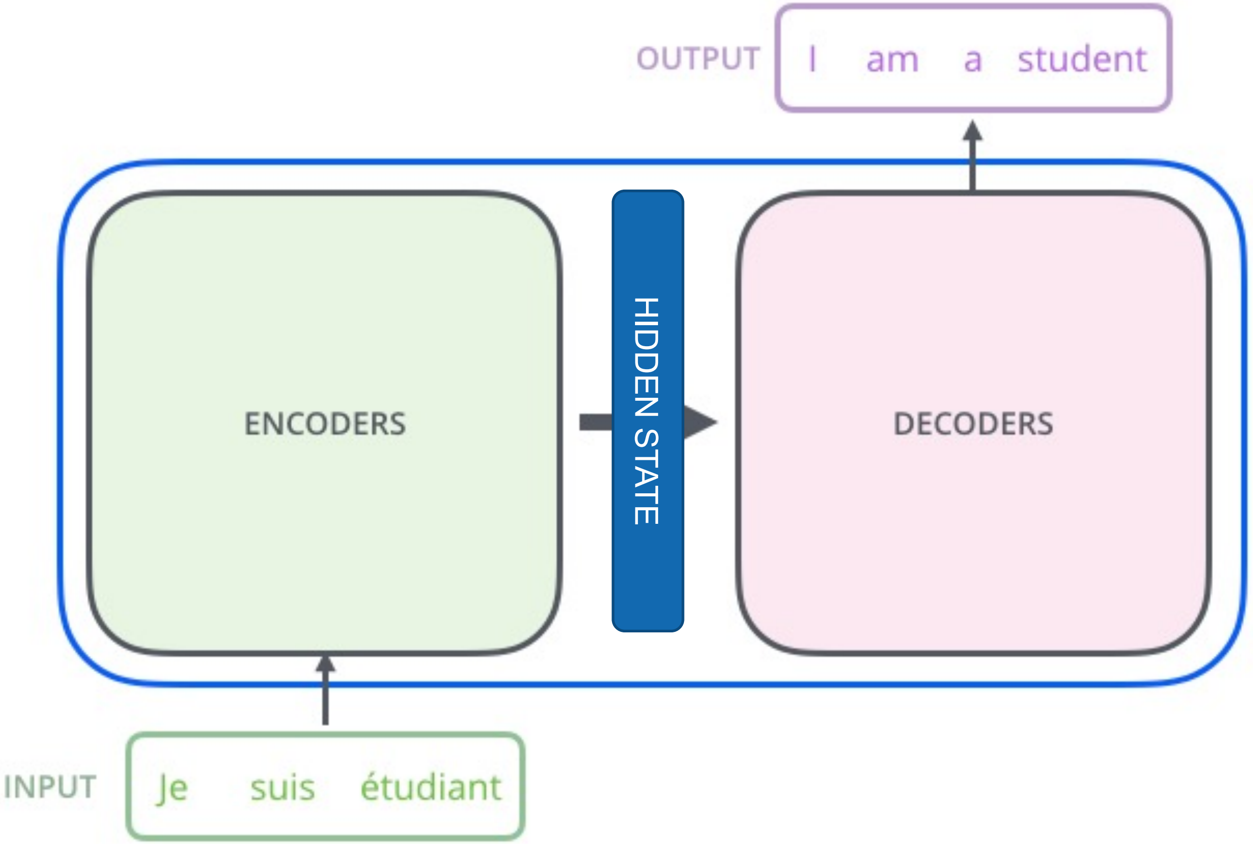
Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

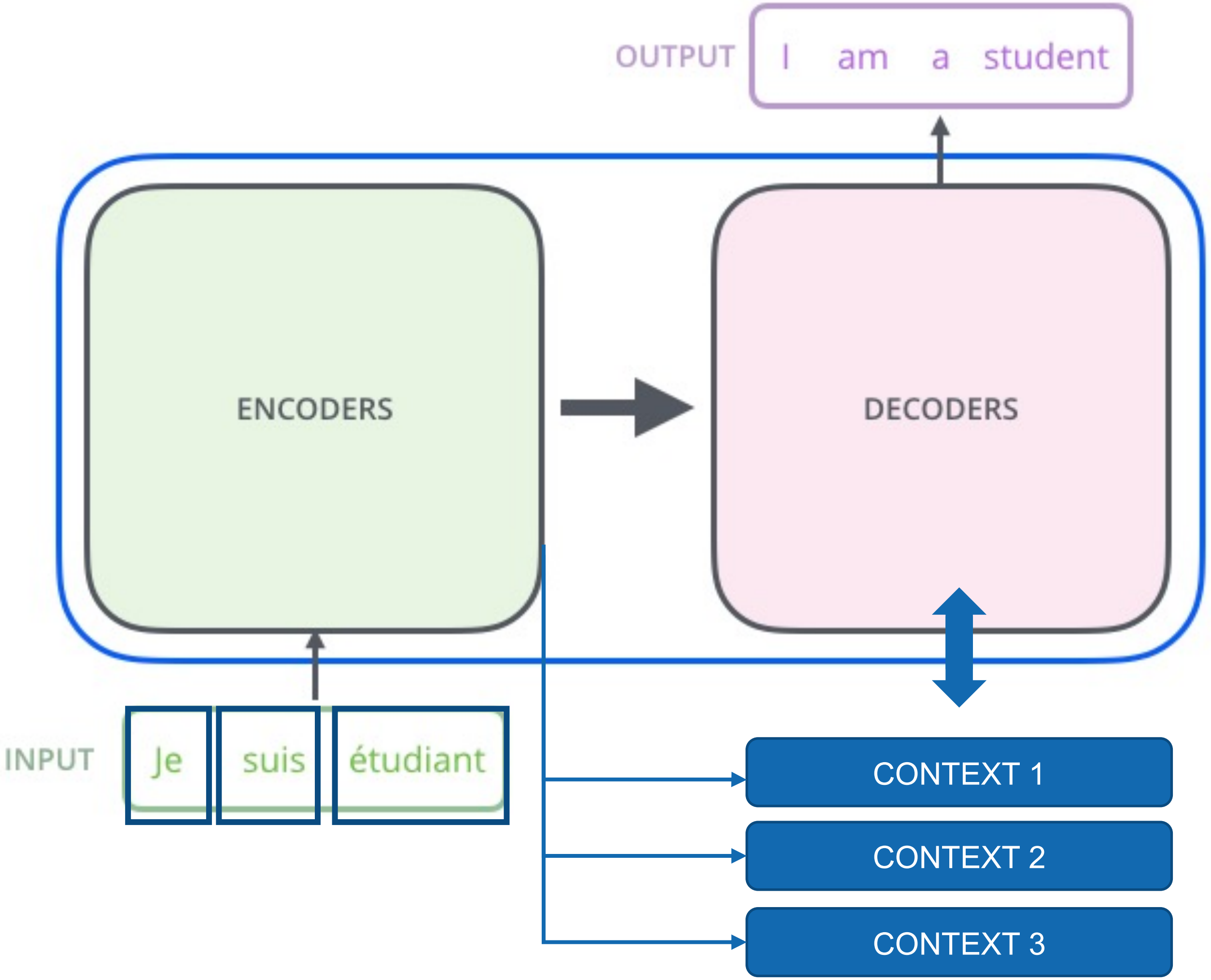
Attention is All You Need (Vaswani et al.)

Fixed size output vector bottleneck



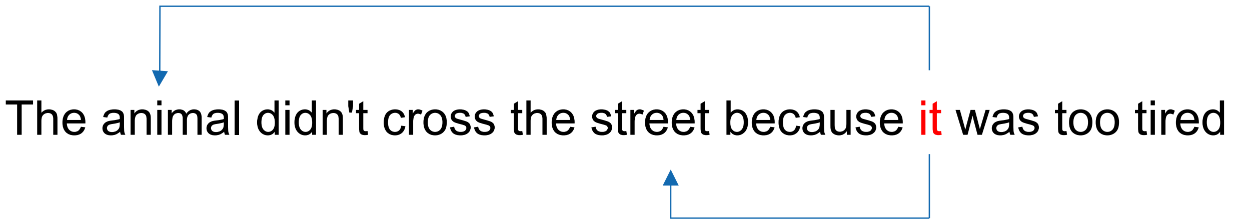
http://jalammar.github.io/illustrated-transformer/

Fixed size output vector bottleneck

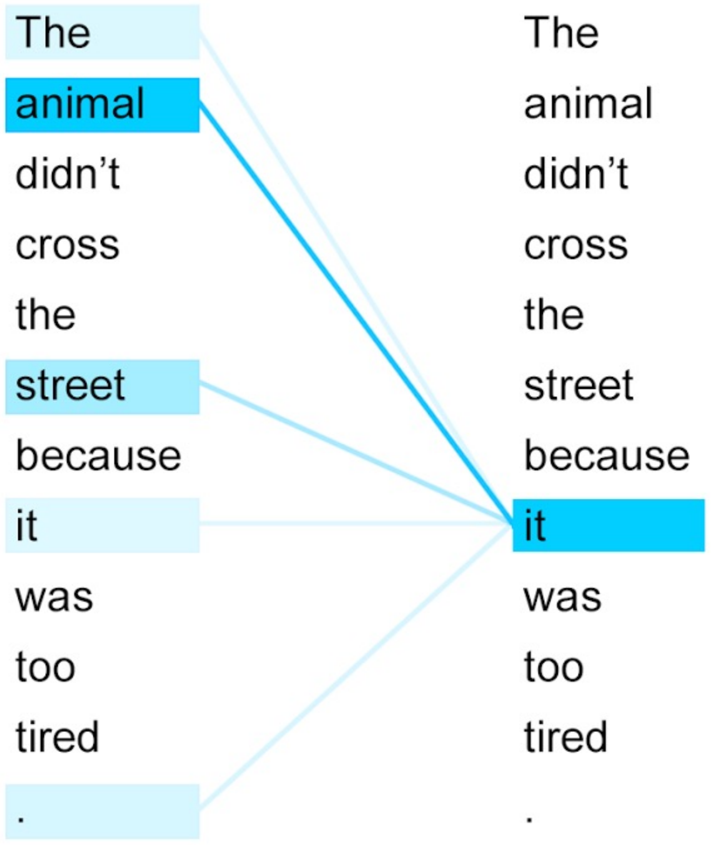


http://jalammar.github.io/illustrated-transformer/

Attention



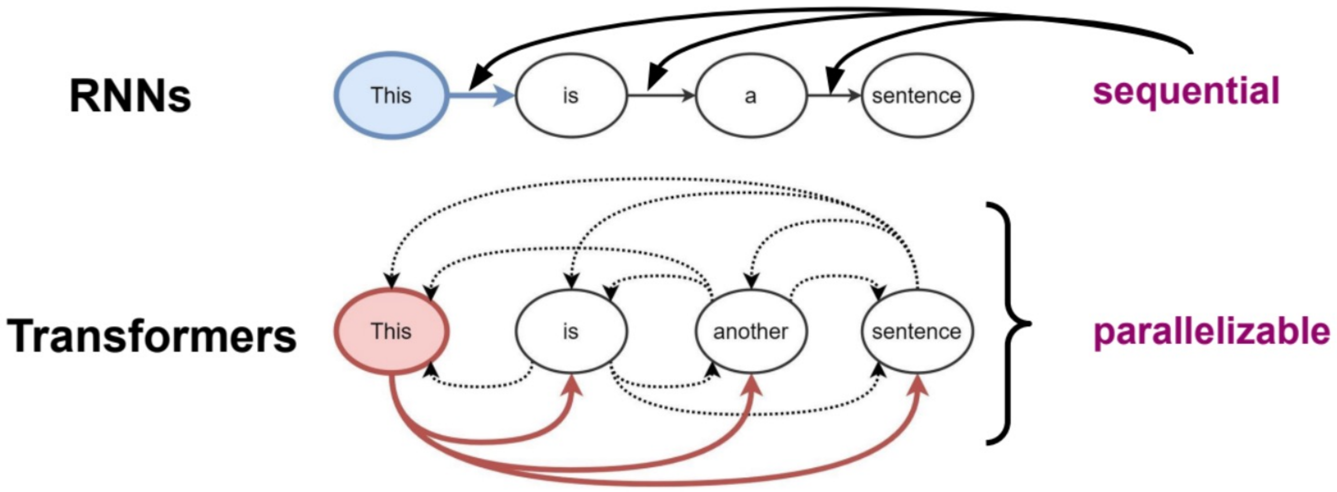
The animal didn't cross the street because it was too tired



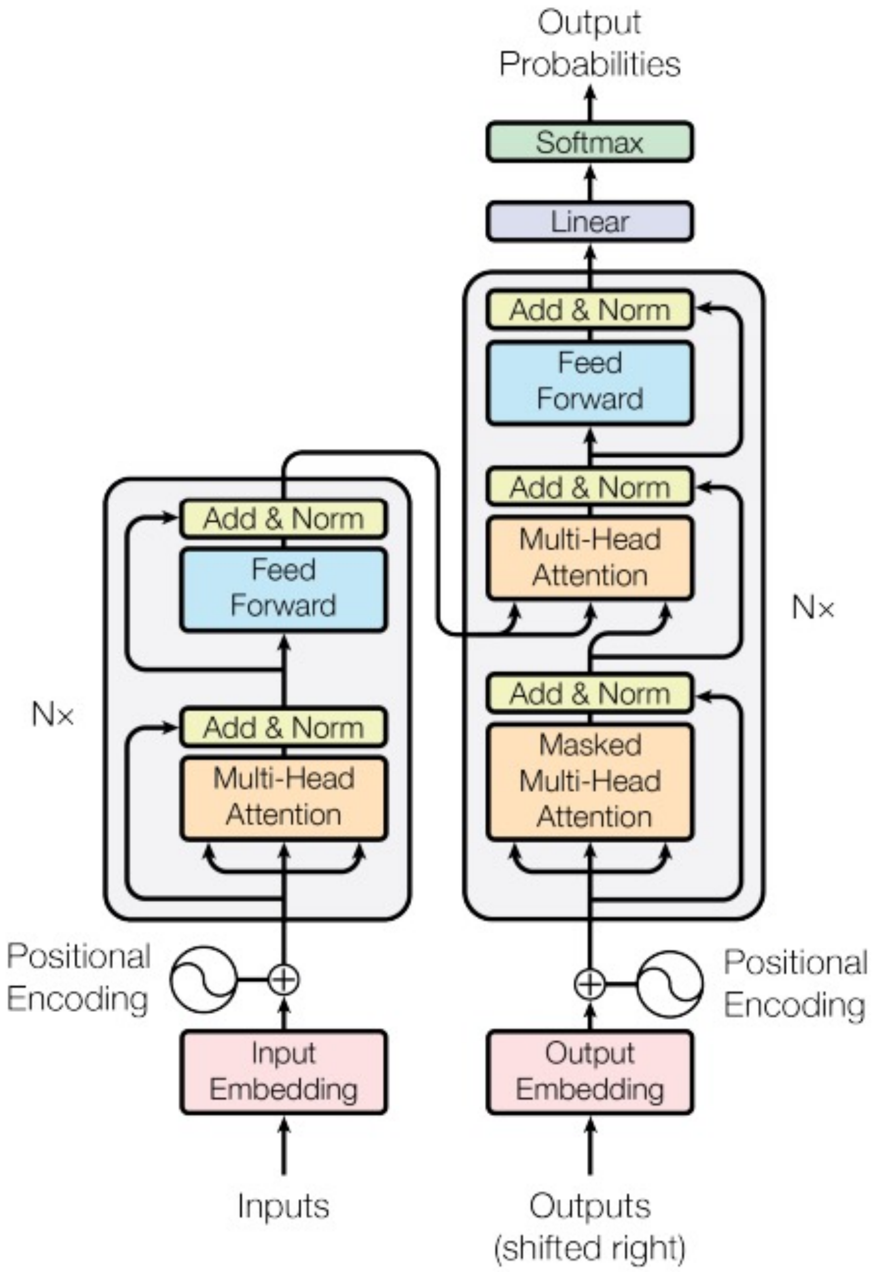
http://jalammar.github.io/illustrated-transformer/

Transformer Architecture

- **Scalable:** the input sequence can be calculated in parallel
- **Attention**: allows to focus on different parts of the input sequence (has long memory, can pay attention to future tokens)

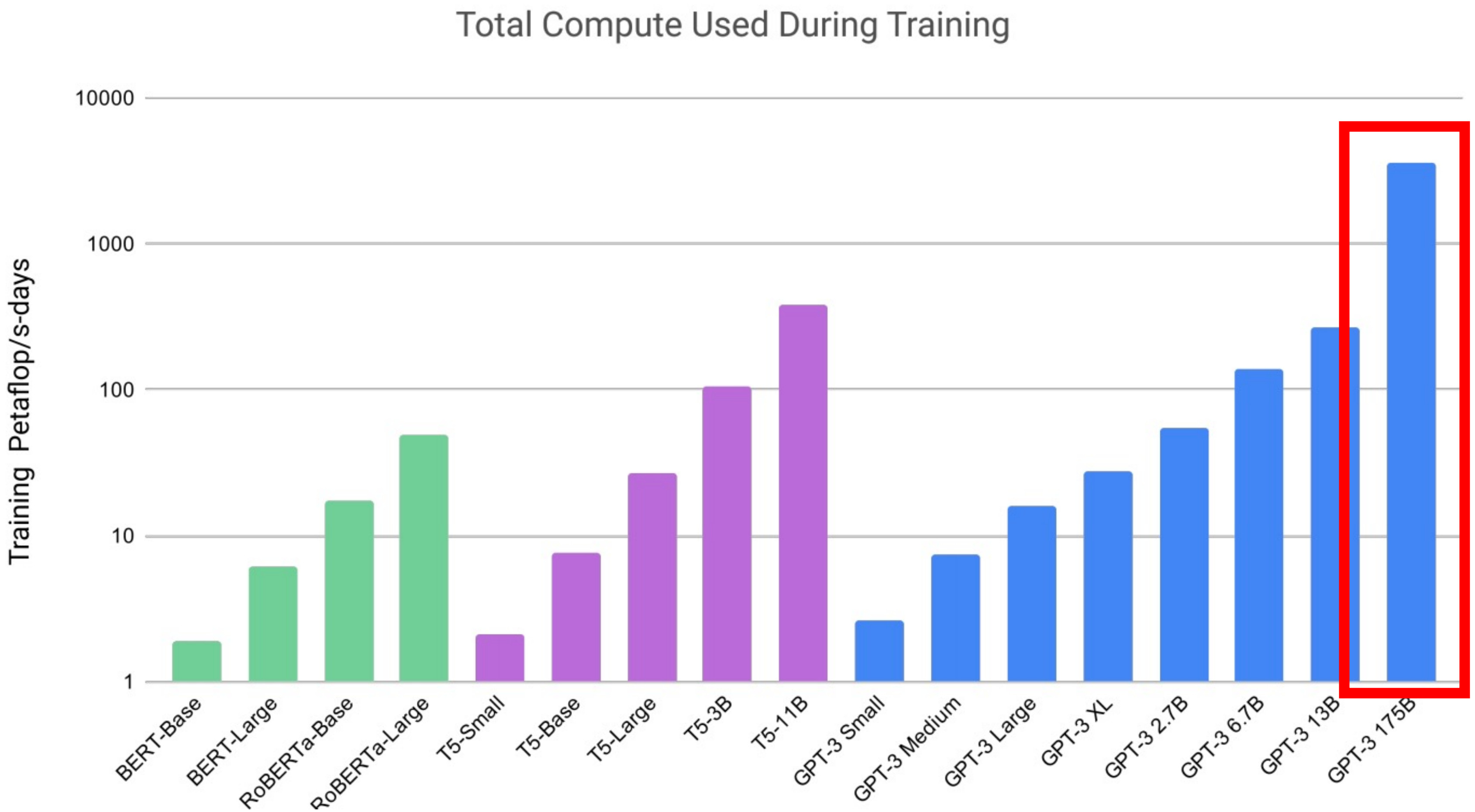


Natural Language Processing Lecture, ETH Zürich, Spring 2021

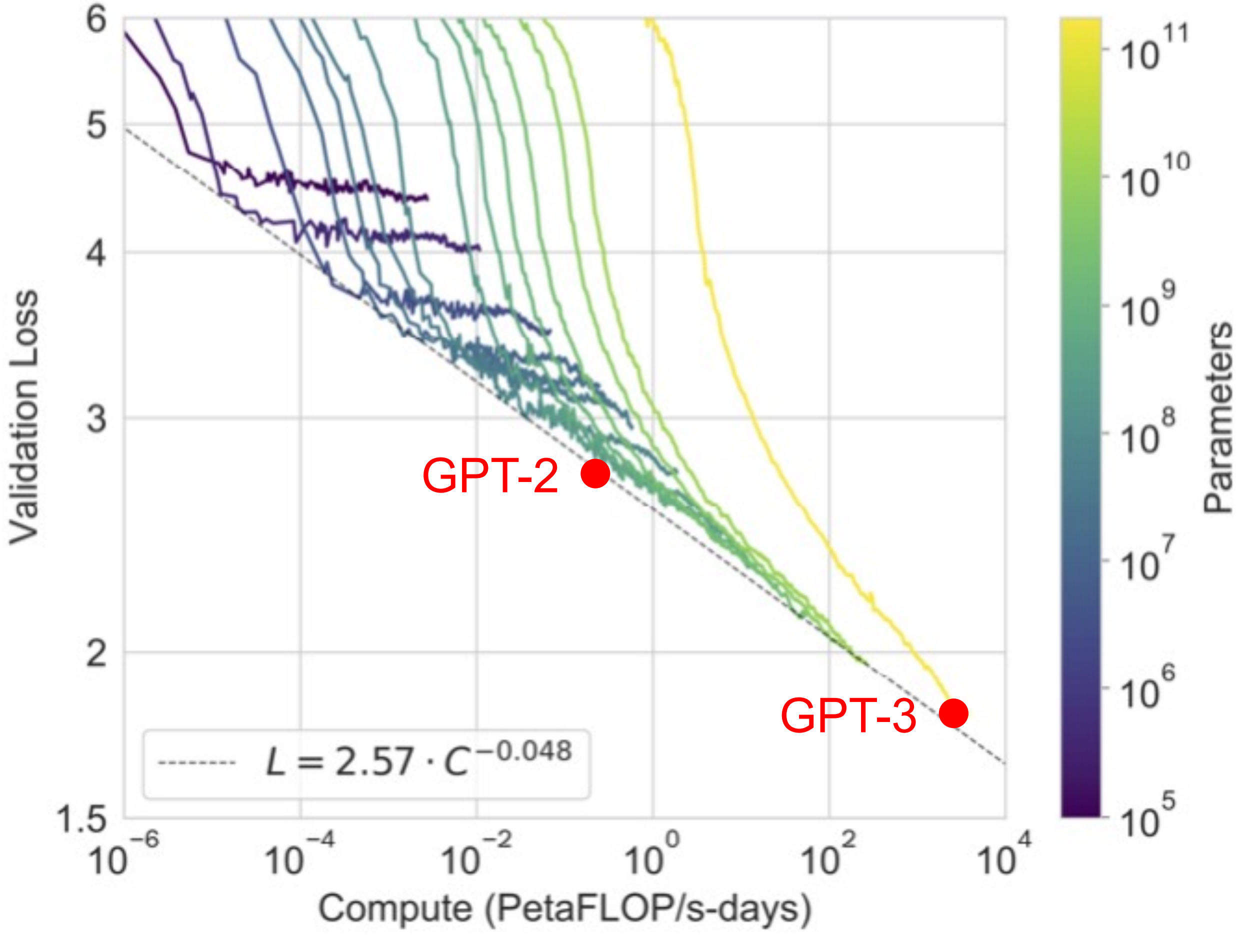


Attention Is All You Need, Vaswani et al. (arXiv:1706.03762)

GPT-3 is huge: Model and training set size

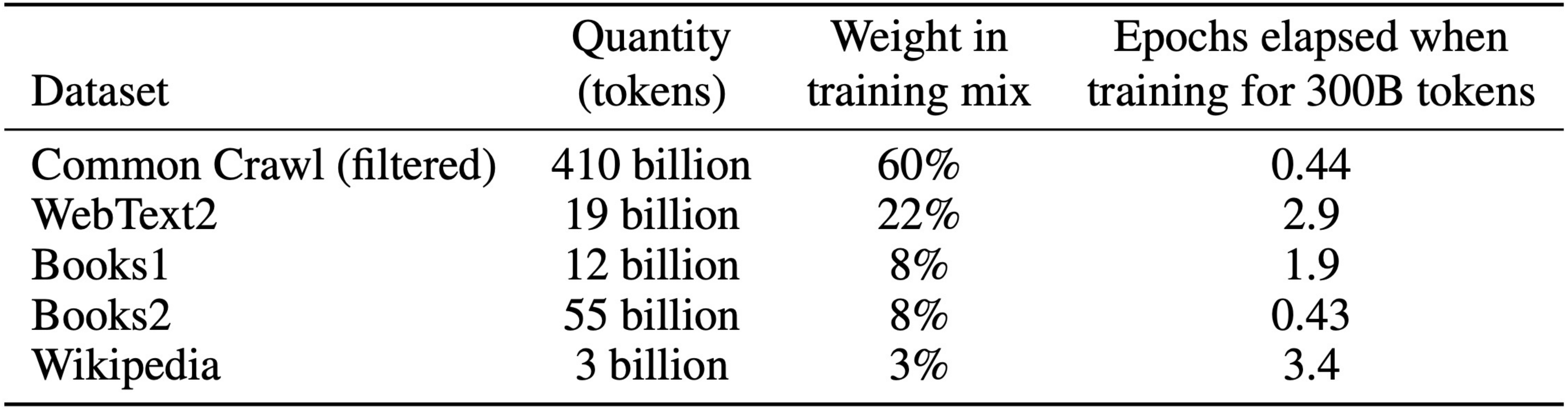


Motivation for scaling



Experiments

Dataset



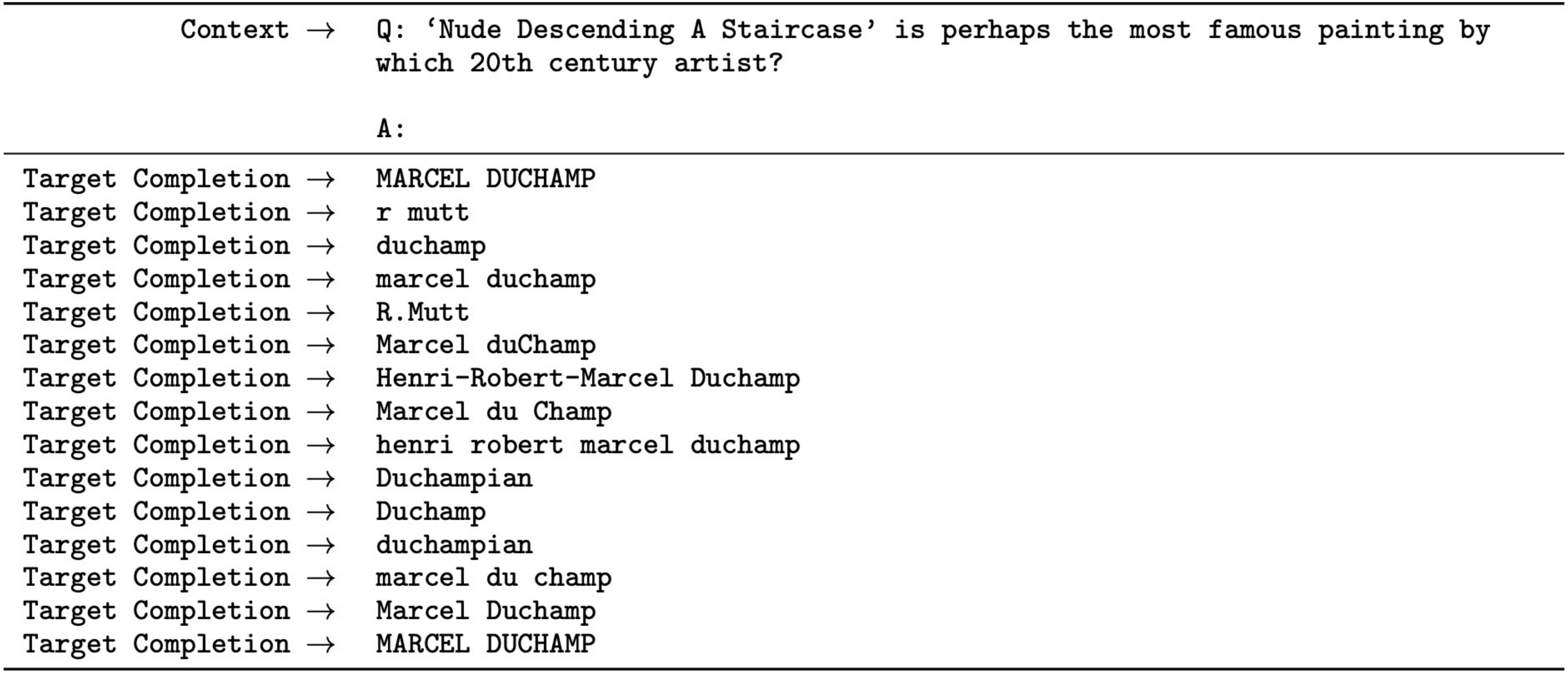
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Performance Overview

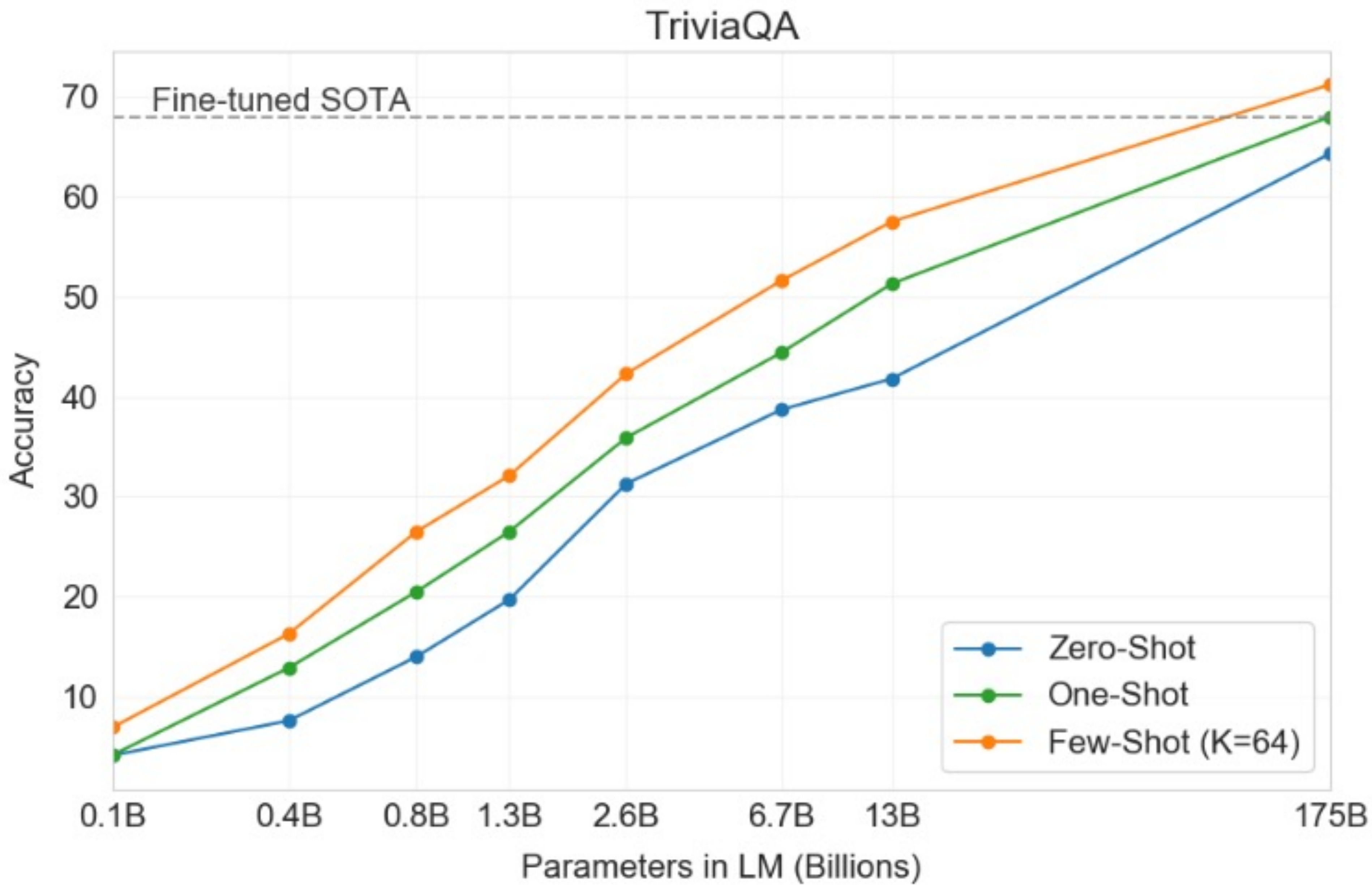
Task Class	Few-shot performance
Cloze, Completion and Language Modeling	Very Good
Question Answering / Knowledge Base	Very Good
Translation	Good
Winograd	Good
Common-Sense Reasoning	Mixed
Reading Comprehension	Mixed
NLI	Poor
Bias Issues	Poor

Table: Dario Amodei NeurIPS

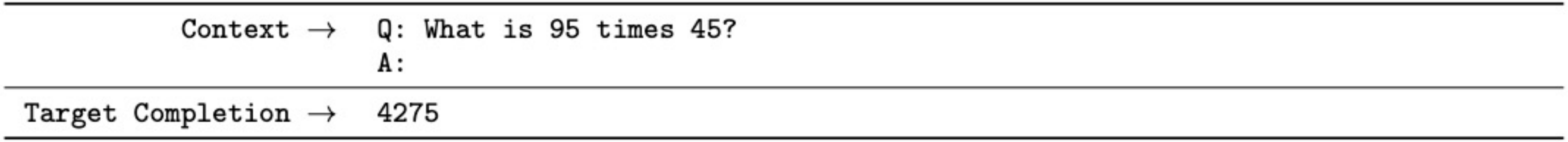
Trivia Q&A



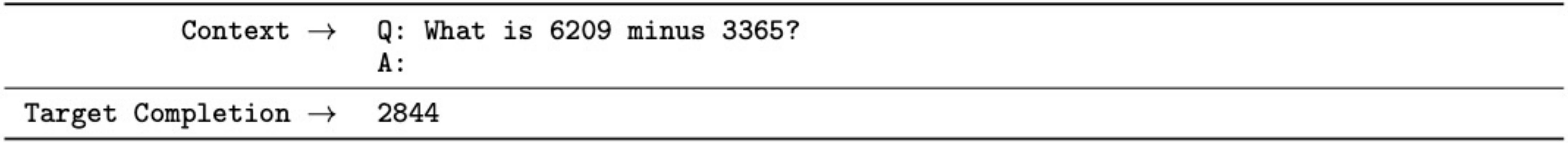
Context →	Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist? A:
Target Completion →	MARCEL DUCHAMP
Target Completion →	r mutt
Target Completion →	duchamp
Target Completion →	marcel duchamp
Target Completion →	R.Mutt
Target Completion →	Marcel duChamp
Target Completion →	Henri-Robert-Marcel Duchamp
Target Completion →	Marcel du Champ
Target Completion →	henri robert marcel duchamp
Target Completion →	Duchampian
Target Completion →	Duchamp
Target Completion →	duchampian
Target Completion →	marcel du champ
Target Completion →	Marcel Duchamp
Target Completion →	MARCEL DUCHAMP


TriviaQA
Fine-tuned SOTA
Accuracy
70
60
50
40
30
20
10
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)
Zero-Shot
One-Shot
Few-Shot (K=64)

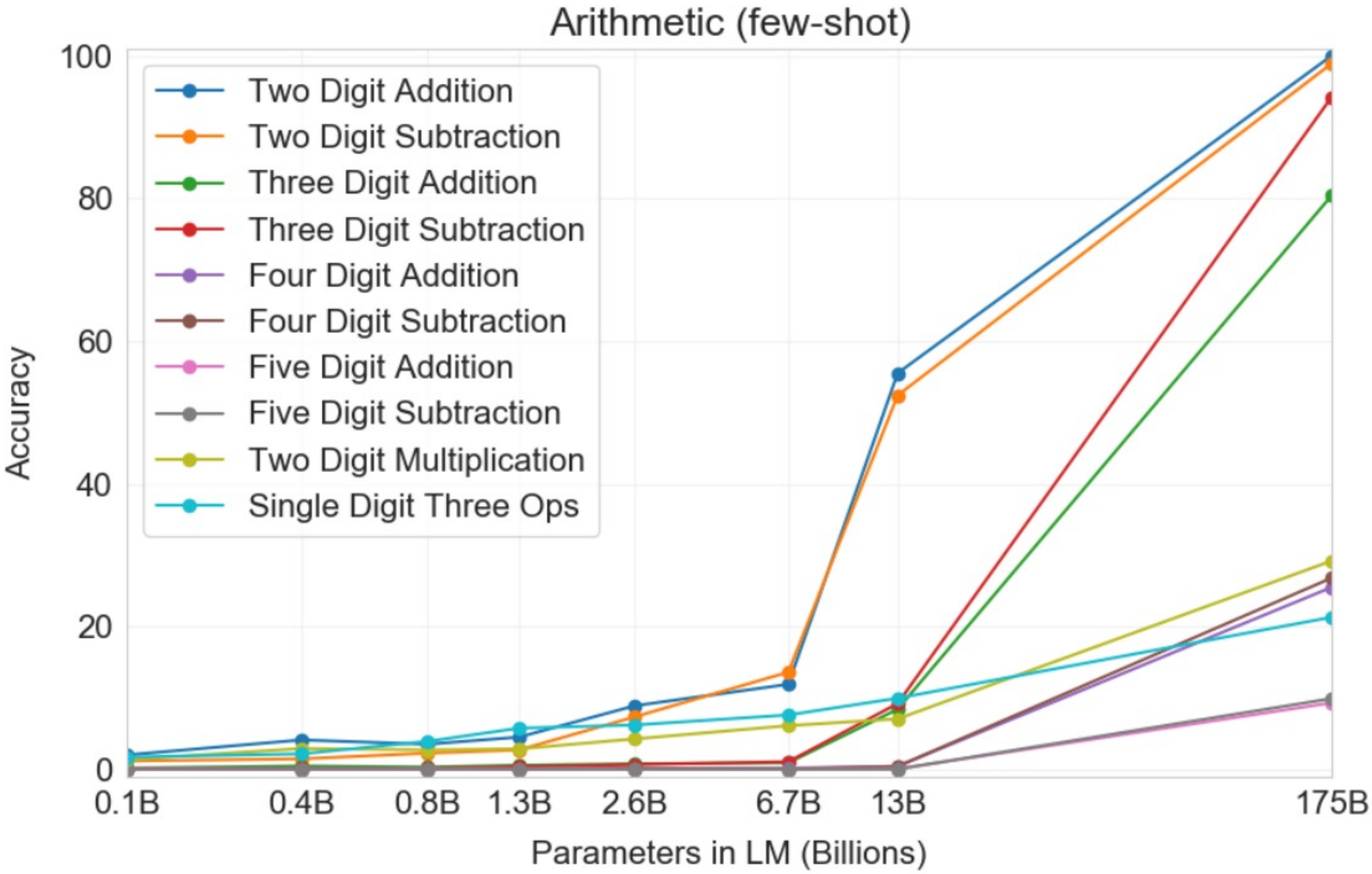
Arithmetic



Context →	Q: What is 95 times 45? A:
Target Completion →	4275



Context →	Q: What is 6209 minus 3365? A:
Target Completion →	2844


Arithmetic (few-shot)
Accuracy
100
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)
Two Digit Addition
Two Digit Subtraction
Three Digit Addition
Three Digit Subtraction
Four Digit Addition
Four Digit Subtraction
Five Digit Addition
Five Digit Subtraction
Two Digit Multiplication
Single Digit Three Ops

Decontamination

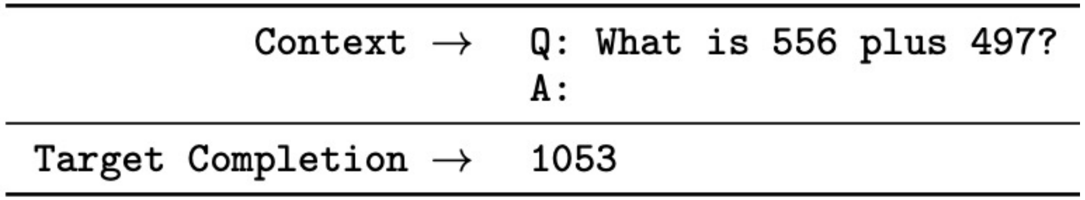
- Experiments used a contaminated datasets

> *“Unfortunately, a bug resulted in only partial removal of all detected overlaps from the training data. Due to the cost of training, it wasn’t feasible to retrain the model.”*

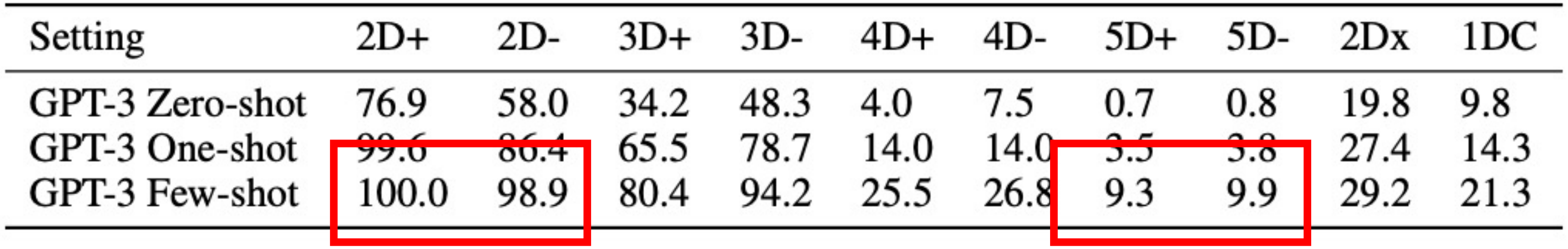
Discussion

Memorization

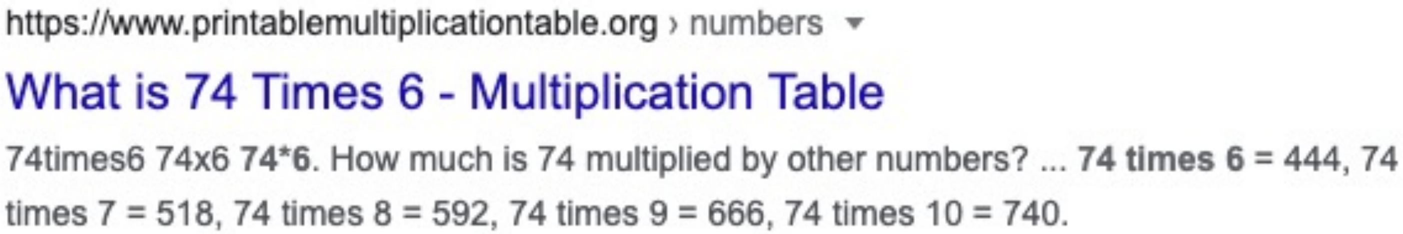
- Data from language models can be memorized
- Example: arithmetic experiment.



Context →	Q: What is 556 plus 497? A:
Target Completion →	1053



Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3



https://www.printablemultiplicationtable.org › numbers ▾

What is 74 Times 6 - Multiplication Table

74times6 74x6 **74*6**. How much is 74 multiplied by other numbers? ... **74 times 6** = 444, 74 times 7 = 518, 74 times 8 = 592, 74 times 9 = 666, 74 times 10 = 740.

Decontamination

<NUM1> * <NUM2> =

<NUM1> times <NUM2>

Is GPT-3 Reasoning or Pattern Matching?

Reasoning

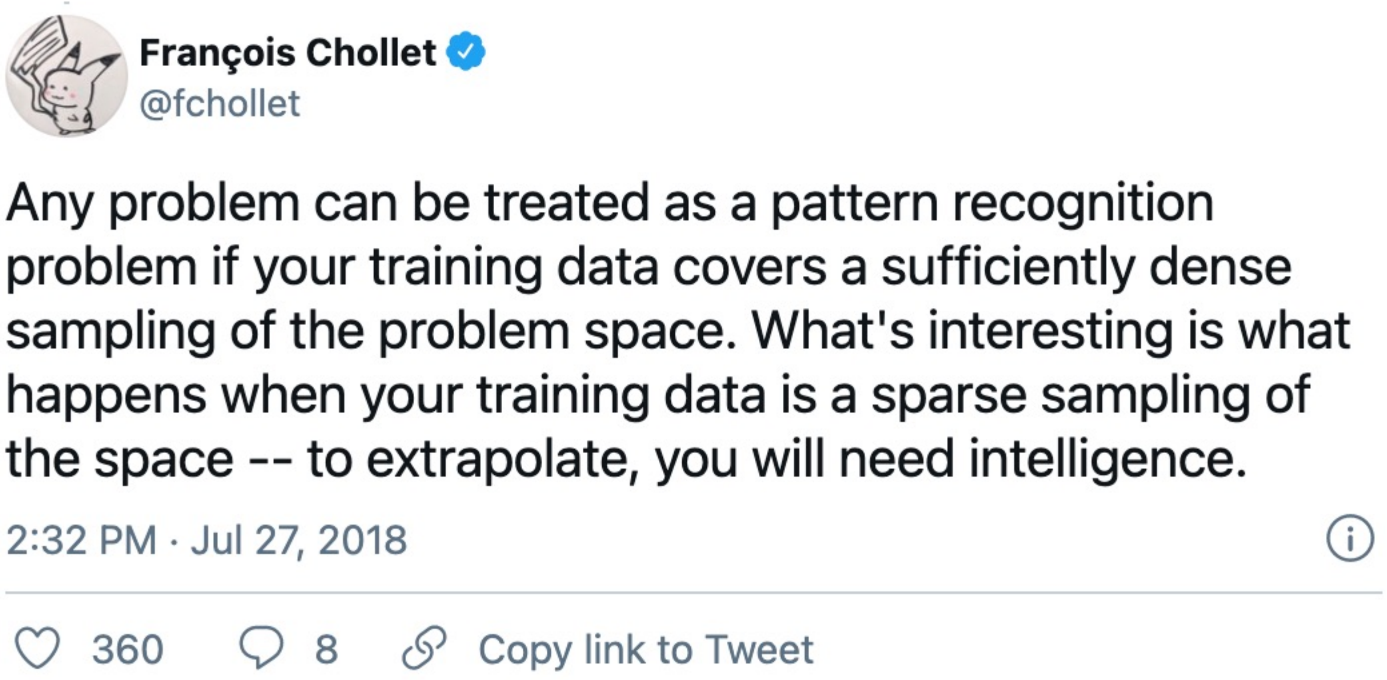
- The models learns the language in pre-training
- The model learned to reason, keep context, write coherent language etc.
- Answers the query using new skills

"understand and answer question"

Pattern Matching

- Model stores a large amount of text in its "database"
- Model uses the K examples as pattern matching (or "query")
- Model filters unsupervised text
- Model outputs interpolated result

"filter and interpolate training data"



François Chollet
@fchollet

Any problem can be treated as a pattern recognition problem if your training data covers a sufficiently dense sampling of the problem space. What's interesting is what happens when your training data is a sparse sampling of the space -- to extrapolate, you will need intelligence.

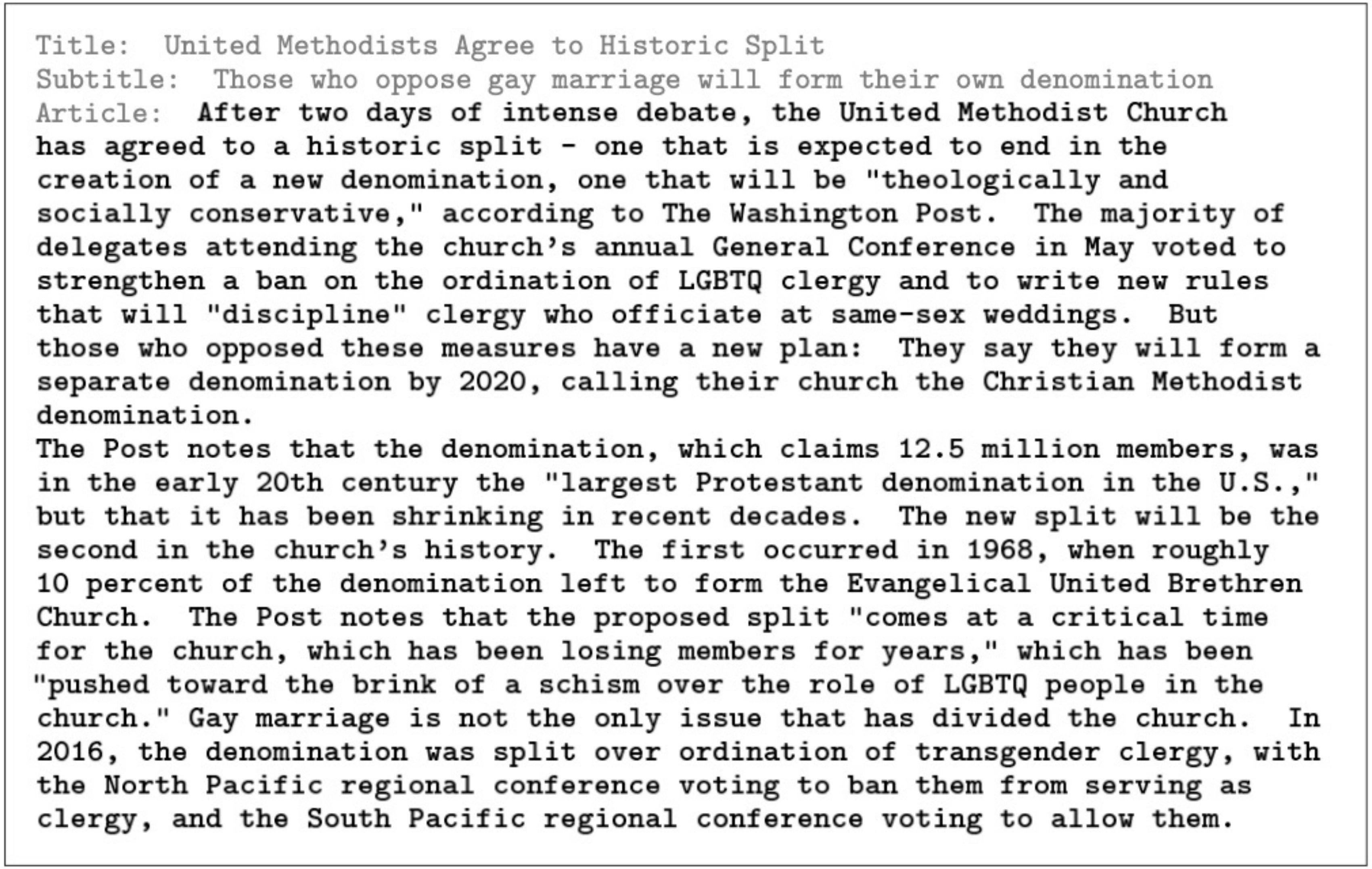
2:32 PM · Jul 27, 2018

Conclusion

- **Main scientific contribution**: Few-shot learning has larger effect for larger models
- Shows trend to make models larger more efficient
- Meta-learning (no fine-tuning) showed very good performance
- Unprecedented scalability of transformer-based architectures
- Memorization and pattern matching (no reasoning)
- Data contamination
- Bias

Appendix

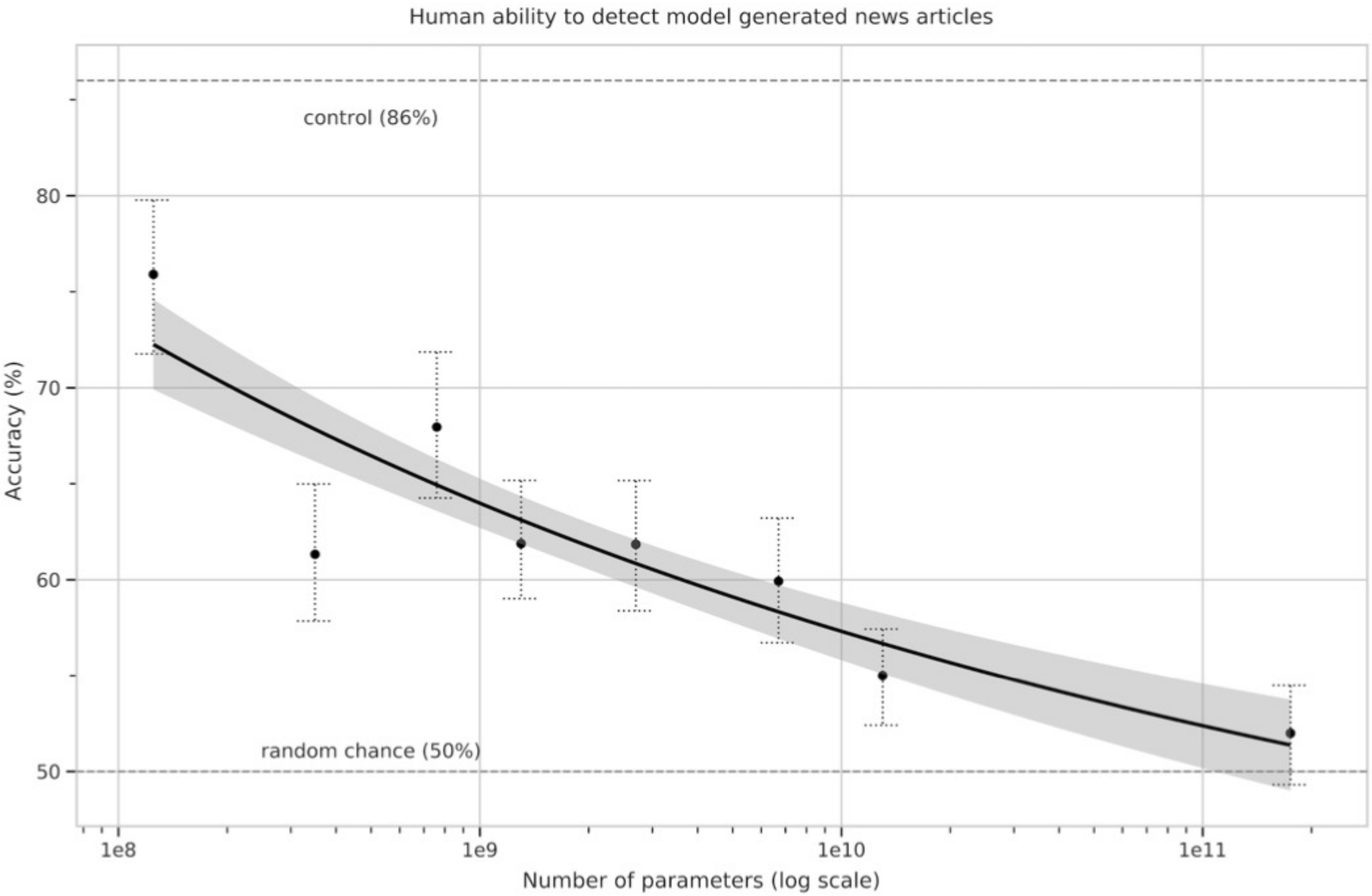
News article generation



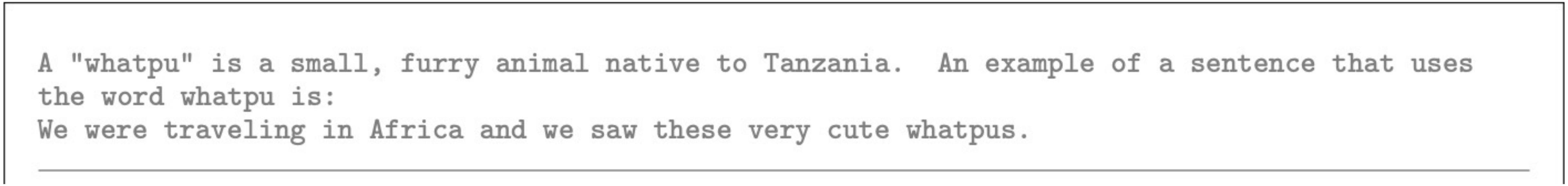
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Human ability to detect model generated news articles

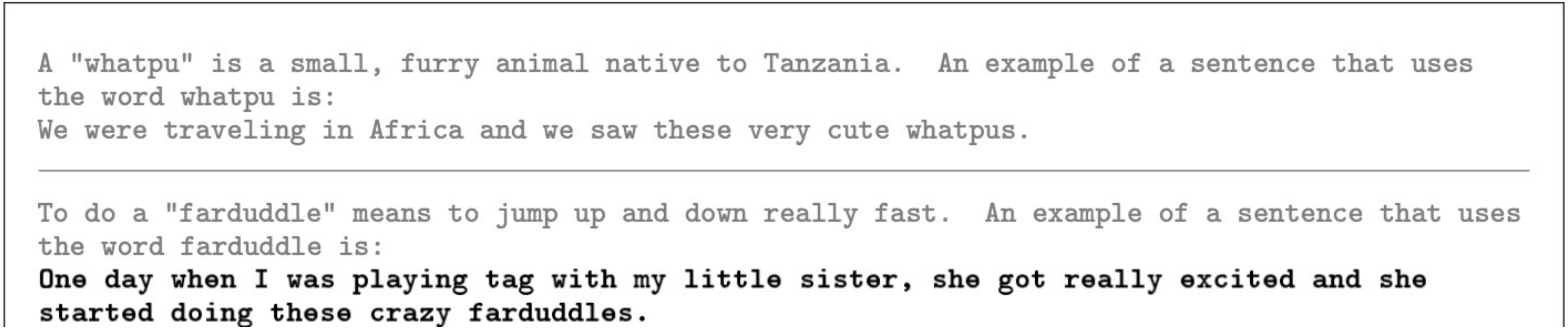


Learning and using novel words



A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

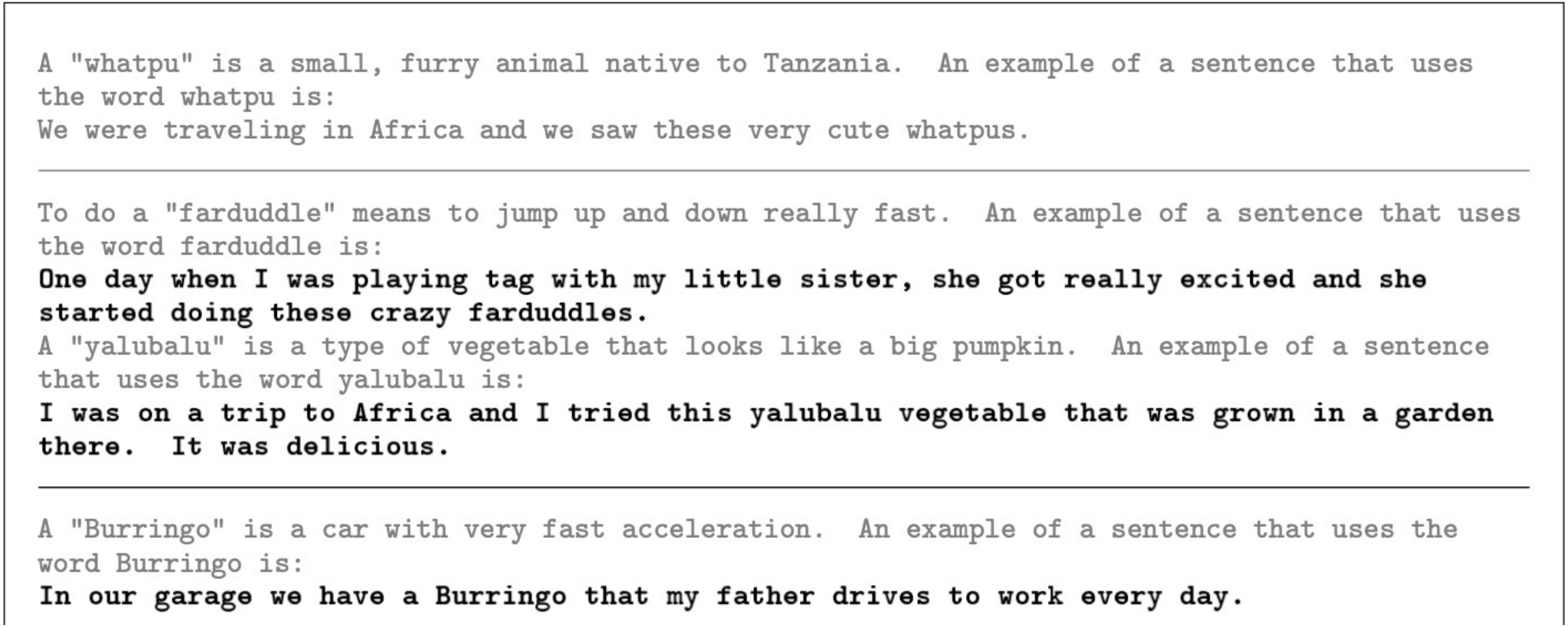
Learning and using novel words



A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

Learning and using novel words

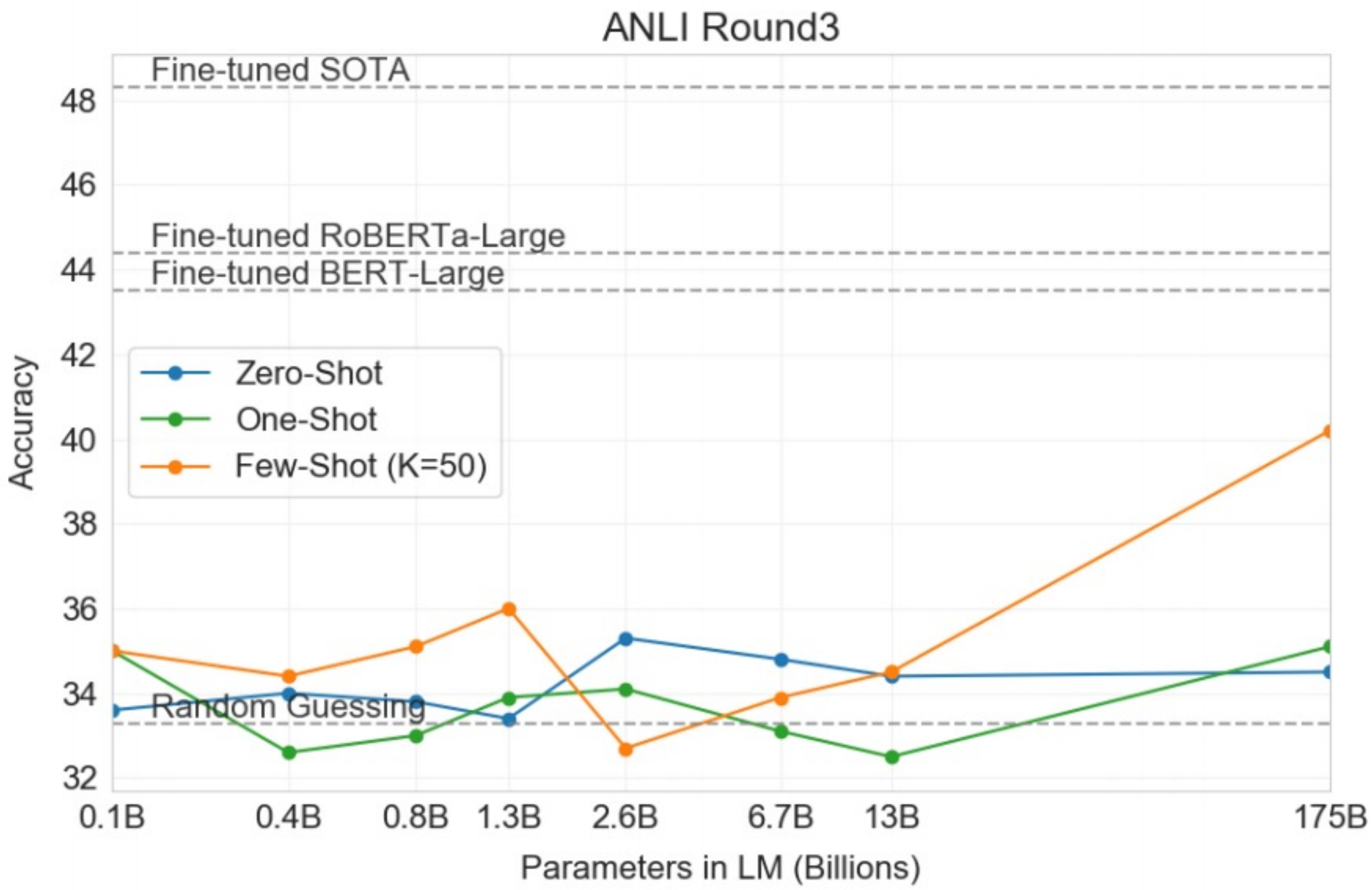


A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

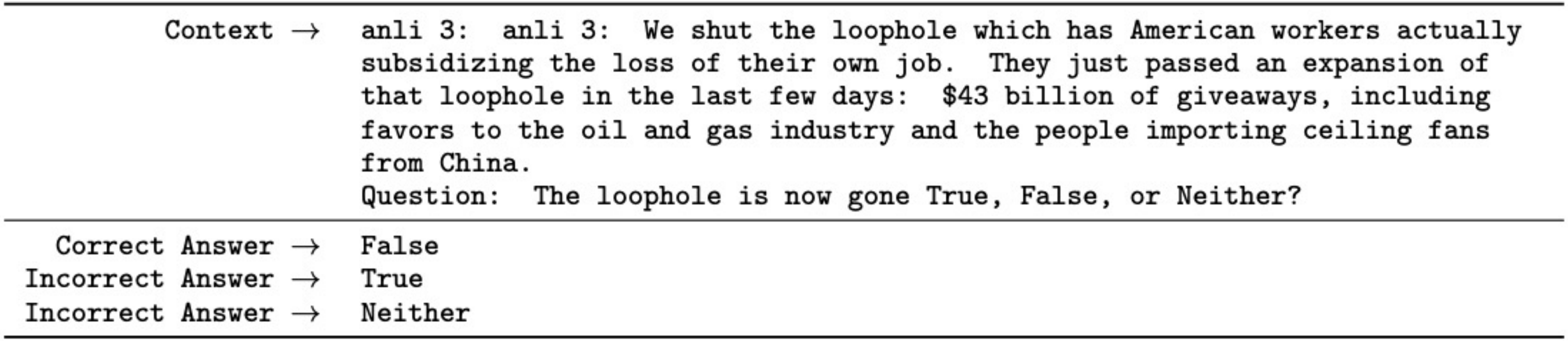
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

NLI performance



ANLI dataset example



Context →	anli 3: anli 3: We shut the loophole which has American workers actually subsidizing the loss of their own job. They just passed an expansion of that loophole in the last few days: $43 billion of giveaways, including favors to the oil and gas industry and the people importing ceiling fans from China. Question: The loophole is now gone True, False, or Neither?
Correct Answer →	False
Incorrect Answer →	True
Incorrect Answer →	Neither

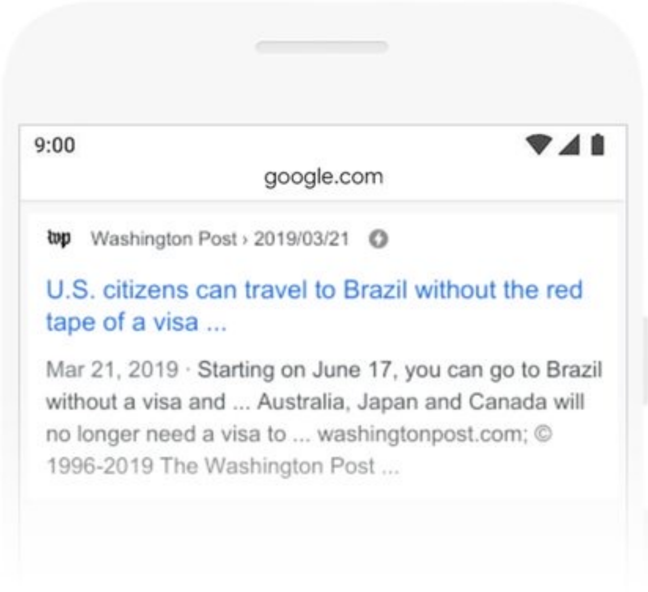
Figure G.10: Formatted dataset example for ANLI R3

Real World Applications

- Search: understanding search queries, finding out intent, improving autocomplete, conversational queries (BERT model)
- Translation
- Text generation: Chat, Customer Service
- Conversational Bots

2019 brazil traveler to usa need a visa

BEFORE

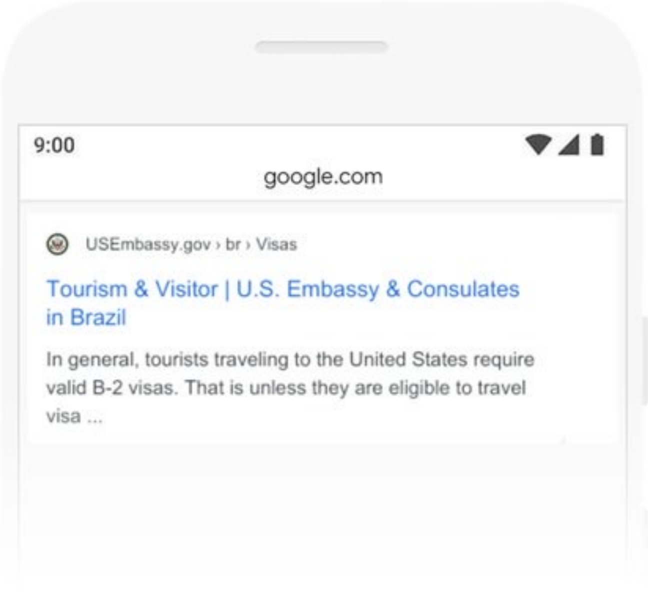


Washington Post › 2019/03/21

U.S. citizens can travel to Brazil without the red tape of a visa ...

Mar 21, 2019 · Starting on June 17, you can go to Brazil without a visa and ... Australia, Japan and Canada will no longer need a visa to ... washingtonpost.com; © 1996-2019 The Washington Post ...

AFTER



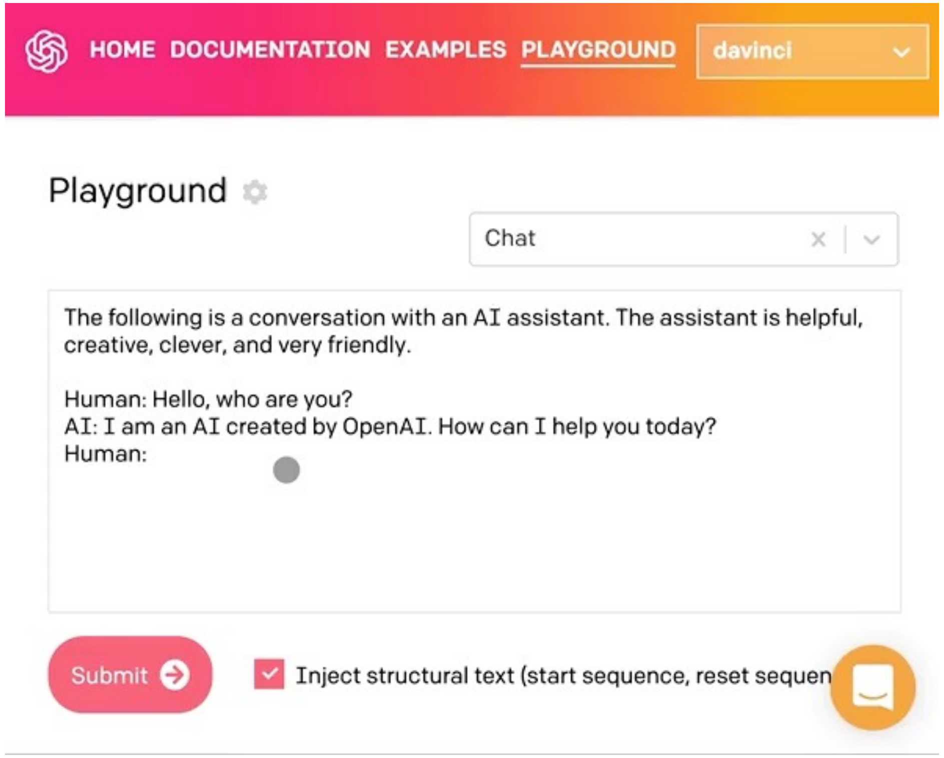
USEmbassy.gov › br › Visas

Tourism & Visitor | U.S. Embassy & Consulates in Brazil

In general, tourists traveling to the United States require valid B-2 visas. That is unless they are eligible to travel visa ...

Quizlet

“A popular use of Quizlet is to learn vocabulary faster. To enable a deeper understanding than rote memorization, Quizlet is building upon OpenAI’s powerful text generation capabilities to automatically generate examples of how each vocabulary word can be used in a sentence.”



Chat

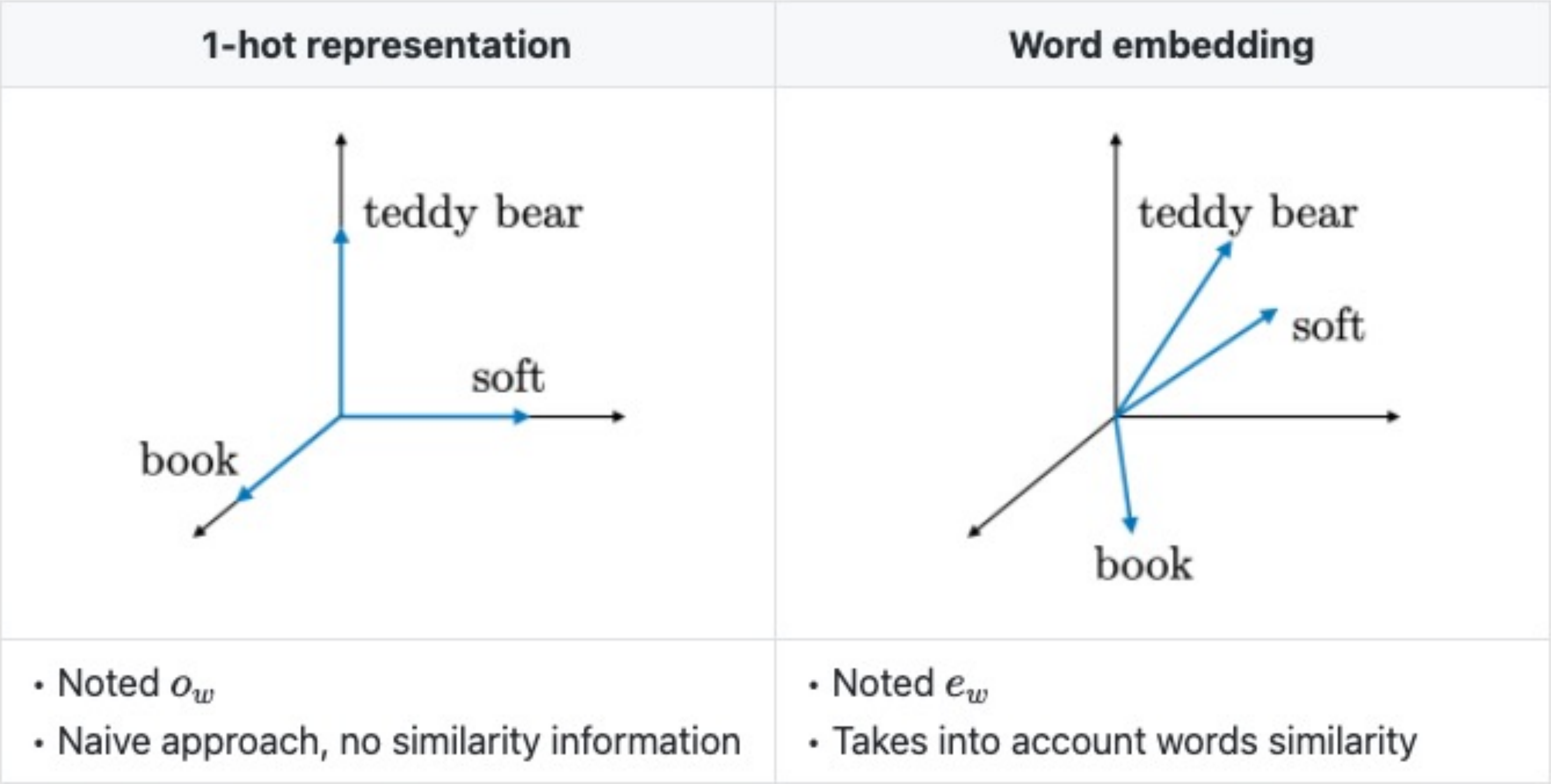
The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human:

☑ Inject structural text (start sequence, reset sequen

```
> python nlsh.py
```

Word representation



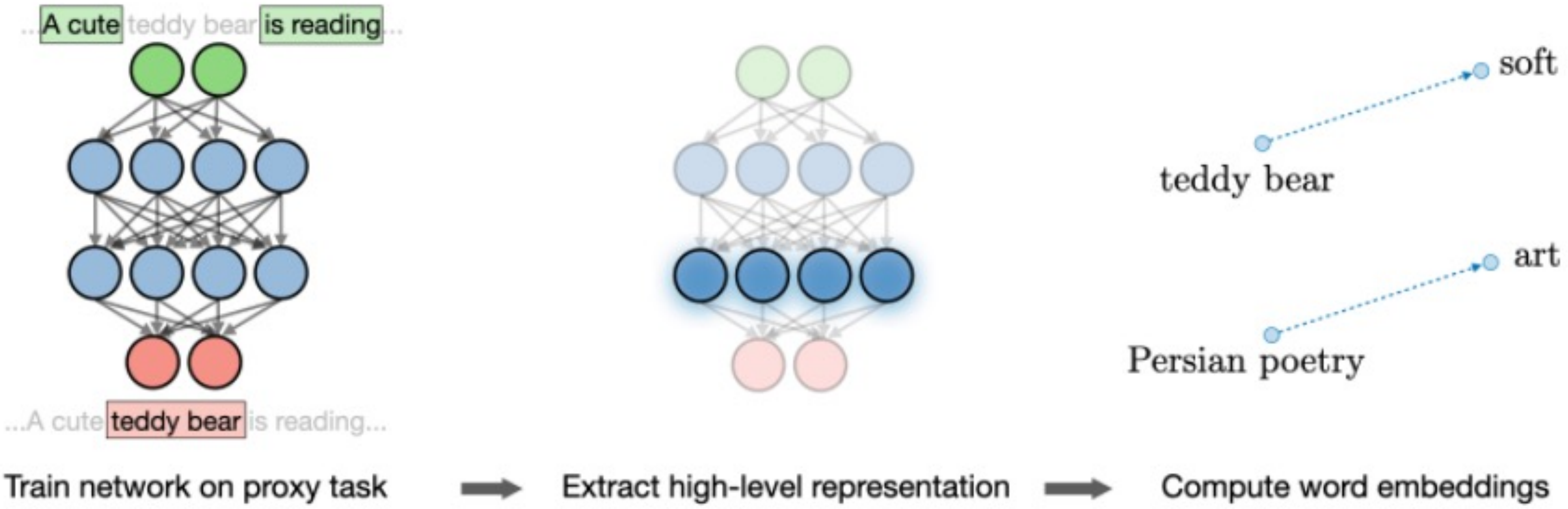
teddy bear
soft
book
teddy bear
soft
book

1-hot representation	Word embedding
• Noted o_w • Naive approach, no similarity information	• Noted e_w • Takes into account words similarity

https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

Word embeddings

❐ **Word2vec** — Word2vec is a framework aimed at learning word embeddings by estimating the likelihood that a given word is surrounded by other words. Popular models include skip-gram, negative sampling and CBOW.


...A cute teddy bear is reading...
...A cute teddy bear is reading...
soft
teddy bear
art
Persian poetry
Train network on proxy task
Extract high-level representation
Compute word embeddings