# LARGE Language Models I

Lecture 8

# How to Calculate P(sentence)

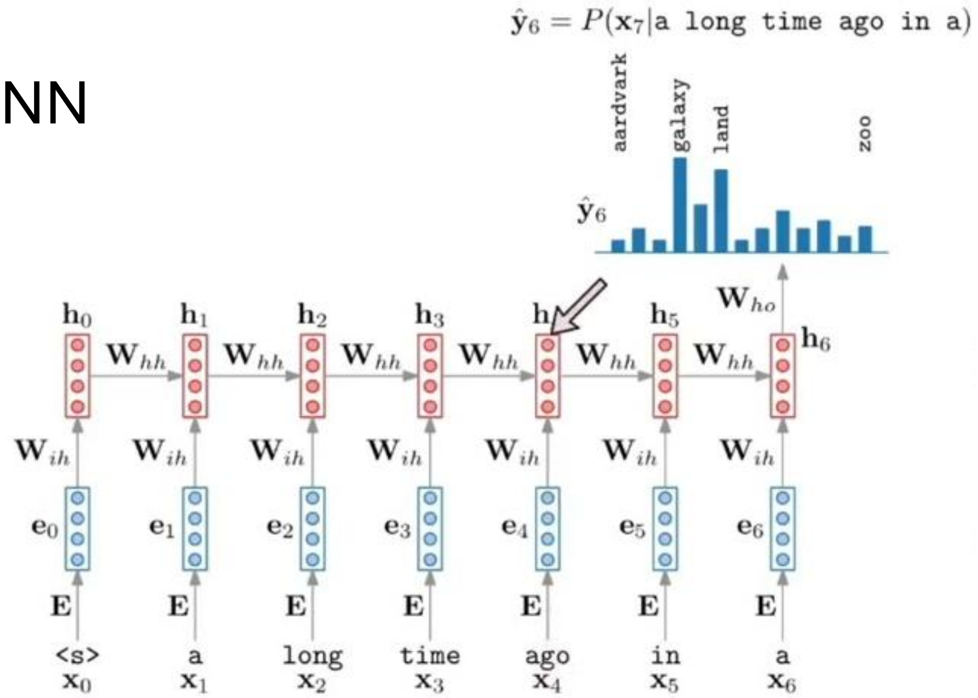
- Bayes' theorem: $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
  - P(A|B): the probability of event A occurring given that B is true.
  - P(B|A): the probability of event B occurring given that A is true.
  - P(A) and P(B): probabilities of observing A and B respectively without any given conditions.
- Chain rule: $$P(A \cap B) = P(B|A)P(A)$$
  - TU Darmstadt's graduation rate 50% (not real number).
  - Acceptance rate: 50% (also not real number).
  - Frank's probability of graduating:

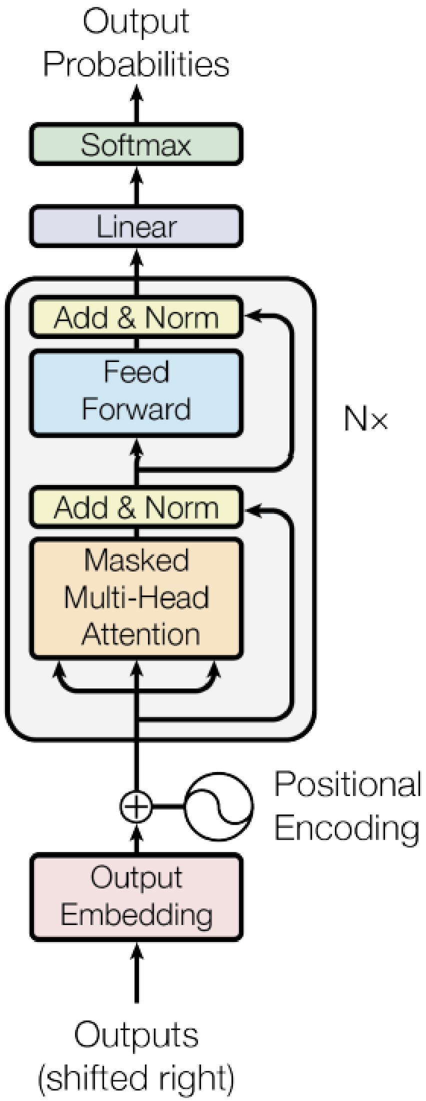
$$P(G \cap A) = P(G|A)P(A) = 0.5 \times 0.5 = 0.25 = 25\%$$

# How to Calculate P(sentence)

- For a neural language model:
  - Output: logits of every token in the vocab → prob. of every token in vocab.

RNN




Transformer
(decoder only, GPT)

# How to Calculate P(sentence)

- For a neural language model:
  - Output: logits of every token in the vocab → prob. of every token in vocab.
- What does each probability stand for?
  - OK, is the model doing at that point?
  - What does it have access to?
    - W1, W2, W3, ... Wi-1
  - What does it output?
    - The next token: Wi

$$P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|x_{1:i-1})$$

# How to Calculate P(sentence)

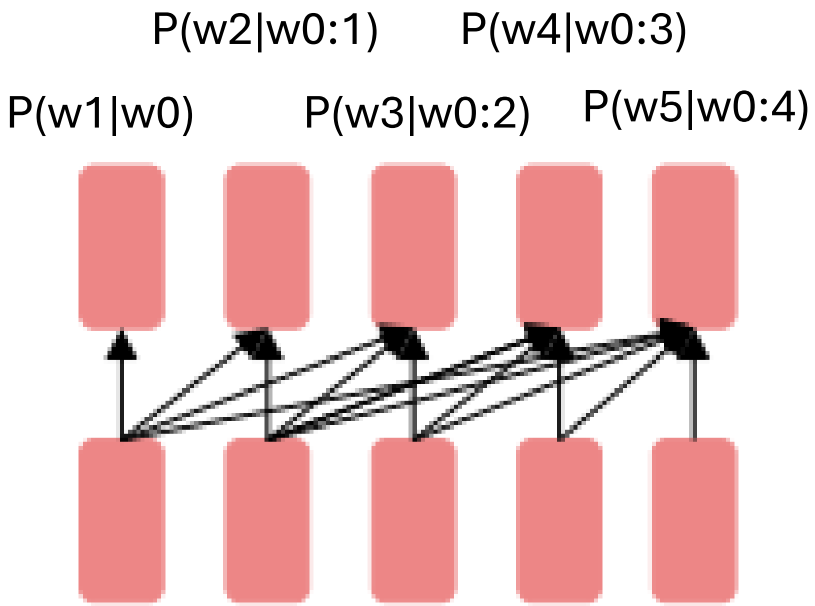
- Apply chain rule:

$$P(x_1, \ldots, x_T) = P(x_1) \prod_{i=1}^{T} P(x_i | x_{1:i-1})$$

- Hmmm what should we do with $x_1$?
- Solution: sentences always start with <BOS>.
  - $P(x_1) = 1$

$$P(x_1, \ldots, x_T) = \prod_{i=1}^{T} P(x_i | x_{1:i-1})$$

- In practice: $$\log P(x_1, \ldots, x_T) = \sum_{i=1}^{T} \log P(x_i | x_{1:i-1})$$



# Transformers: Is Attention All We Need?

- Last lecture, we learned that attention dramatically improves the performance of recurrent neural networks.
- Today, we will take this one step further and ask Is Attention All We Need?
- Spoiler: Not Quite!

## 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

# Great Results with Transformers: SuperGLUE

- SuperGLUE is a suite of challenging NLP tasks, including QA, WSD, coreference resolution, and NLI.

| | Rank | Name | Model | URL | Score | BoolQ | CB | COPA | MultiRC | ReCoRD | RTE | WiC | WSC | AX-b | AX-g |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | JDExplore d-team | Vega v2 | | 91.3 | 90.5 | 98.6/99.2 | 99.4 | 88.2/62.4 | 94.4/93.9 | 96.0 | 77.4 | 98.6 | -0.4 | 100.0/50.0 |
| + | 2 | Liam Fedus | ST-MoE-32B | | 91.2 | 92.4 | 96.9/98.0 | 99.2 | 89.6/65.8 | 95.1/94.4 | 93.5 | 77.7 | 96.6 | 72.3 | 96.1/94.1 |
| | 3 | Microsoft Alexander v-team | Turing NLR v5 | | 90.9 | 92.0 | 95.9/97.6 | 98.2 | 88.4/63.0 | 96.4/95.9 | 94.1 | 77.1 | 97.3 | 67.8 | 93.3/95.5 |
| | 4 | ERNIE Team - Baidu | ERNIE 3.0 | | 90.6 | 91.0 | 98.6/99.2 | 97.4 | 88.6/63.2 | 94.7/94.2 | 92.6 | 77.4 | 97.3 | 68.6 | 92.7/94.7 |
| | 5 | Yi Tay | PaLM 540B | | 90.4 | 91.9 | 94.4/96.0 | 99.0 | 88.7/63.6 | 94.2/93.3 | 94.1 | 77.4 | 95.9 | 72.9 | 95.5/90.4 |
| + | 6 | Zirui Wang | T5 + UDG, Single Model (Google Brain) | | 90.4 | 91.4 | 95.8/97.6 | 98.0 | 88.3/63.0 | 94.2/93.5 | 93.0 | 77.9 | 96.6 | 69.1 | 92.7/91.9 |
| + | 7 | DeBERTa Team - Microsoft | DeBERTa / TuringNLRv4 | | 90.3 | 90.4 | 95.7/97.6 | 98.4 | 88.2/63.7 | 94.5/94.1 | 93.2 | 77.5 | 95.9 | 66.7 | 93.3/93.8 |
| | 8 | SuperGLUE Human Baselines | SuperGLUE Human Baselines | | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8/51.9 | 91.7/91.3 | 93.6 | 80.0 | 100.0 | 76.6 | 99.3/99.7 |
| + | 9 | T5 Team - Google | T5 | | 89.3 | 91.2 | 93.9/96.8 | 94.8 | 88.1/63.3 | 94.1/93.4 | 92.5 | 76.9 | 93.8 | 65.6 | 92.7/91.9 |
| | 10 | SPoT Team - Google | Frozen T5 1.1 + SPoT | | 89.2 | 91.1 | 95.8/97.6 | 95.6 | 87.9/61.9 | 93.3/92.4 | 92.9 | 75.8 | 93.8 | 66.9 | 83.1/82.6 |

# Great Results with Transformers: Rise of Large Language Models!

- Today, Transformer-based models dominate LMSYS Chatbot Arena Leaderboard!

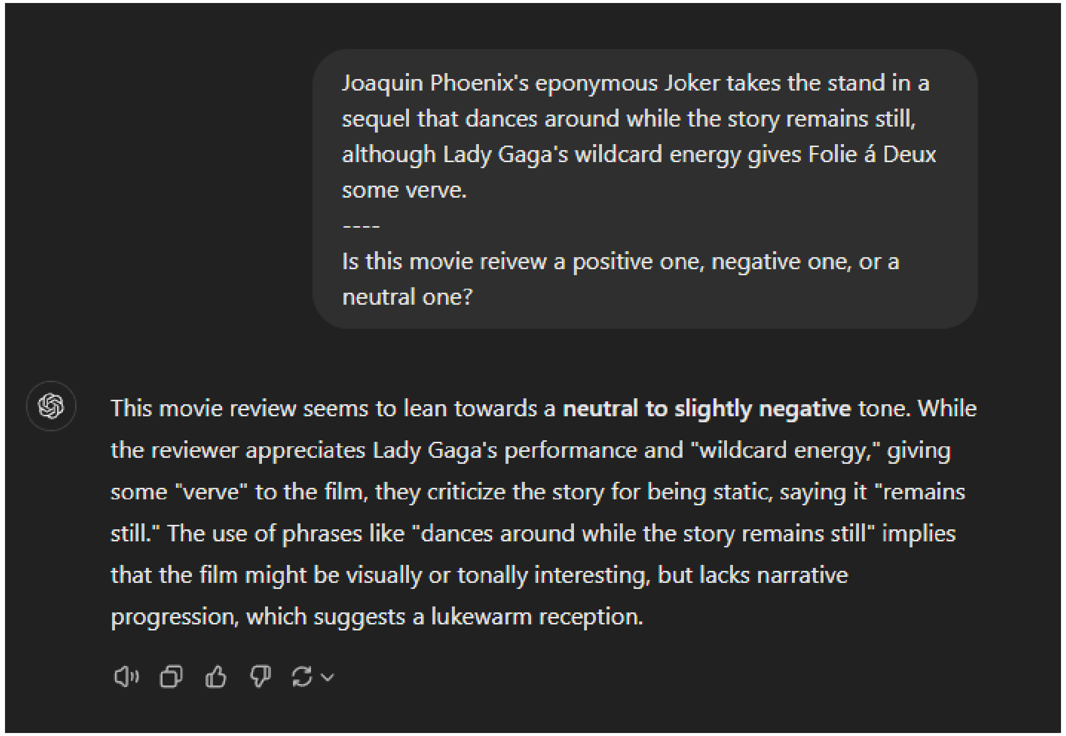
| Rank* (UB) | Model | Arena Score | 95% CI | Votes | Organization | License | Knowledge Cutoff |
|---|---|---|---|---|---|---|---|
| 1 | o1-preview | 1339 | +6/-7 | 9169 | OpenAI | Proprietary | 2023/10 |
| 1 | ChatGPT-4o-latest (2024-09-03) | 1337 | +4/-4 | 16685 | OpenAI | Proprietary | 2023/10 |
| 3 | o1-mini | 1314 | +6/-5 | 9136 | OpenAI | Proprietary | 2023/10 |
| 4 | Gemini-1.5-Pro-Exp-0827 | 1299 | +4/-3 | 31928 | Google | Proprietary | 2023/11 |
| 4 | Grok-2-08-13 | 1293 | +4/-3 | 27731 | xAI | Proprietary | 2024/3 |
| 6 | GPT-4o-2024-05-13 | 1285 | +3/-3 | 93428 | OpenAI | Proprietary | 2023/10 |
| 7 | GPT-4o-mini-2024-07-18 | 1272 | +3/-3 | 33166 | OpenAI | Proprietary | 2023/10 |
| 7 | Claude 3.5 Sonnet | 1269 | +3/-3 | 67165 | Anthropic | Proprietary | 2024/4 |
| 7 | Gemini-1.5-Flash-Exp-0827 | 1269 | +3/-4 | 25027 | Google | Proprietary | 2023/11 |
| 7 | Grok-2-Mini-08-13 | 1268 | +4/-4 | 24956 | xAI | Proprietary | 2024/3 |
| 7 | Gemini Advanced App (2024-05-14) | 1266 | +3/-3 | 52218 | Google | Proprietary | Online |
| 7 | Meta-Llama-3.1-405b-Instruct-bf16 | 1266 | +6/-7 | 8787 | Meta | Llama 3.1 Community | 2023/12 |
| 7 | Meta-Llama-3.1-405b-Instruct-fp8 | 1266 | +4/-4 | 33654 | Meta | Llama 3.1 Community | 2023/12 |

Chiang et al. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

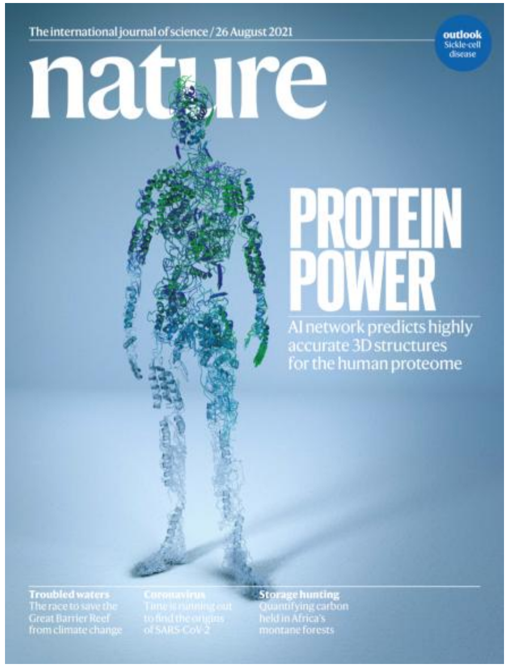
# Even Easier!

- If you think fine-tuning BERT is too hard.
- Prompt ChatGPT and profit!

```python
import torch
from datasets import load_dataset
from transformers import AutoTokenizer, AutoModelForSequenceClassification, TrainingArguments, Trainer

if __name__ == '__main__':
    dataset = load_dataset("yelp_review_full")
    tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")

    def tokenize_function(examples):
        return tokenizer(examples["text"], padding="max_length", truncation=True)

    tokenized_datasets = dataset.map(tokenize_function, batched=True)
    model = AutoModelForSequenceClassification.from_pretrained("bert-base-cased", num_labels=5)

    def acc(eval_pred):
        logits, labels = eval_pred
        predictions = torch.argmax(logits, dim=-1)
        return (predictions == labels).sum().item() / len(labels)

    training_args = TrainingArguments(output_dir="test_trainer", report_to=None)

    trainer = Trainer(
        model=model,
        args=training_args,
        train_dataset=tokenized_datasets["train"],
        eval_dataset=tokenized_datasets["test"],
        compute_metrics=acc,
    )

    trainer.train()
```

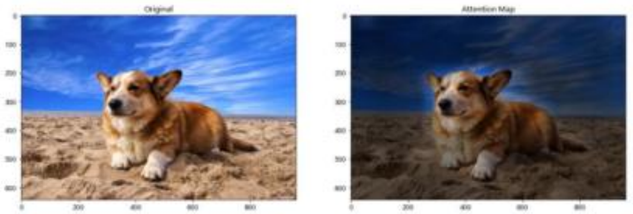


# Transformers Even Show Promise Outside of NLP




**Protein Folding**

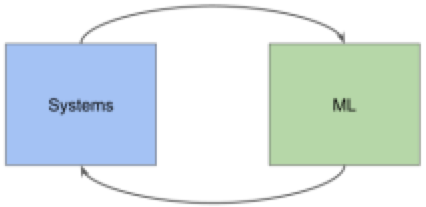
Jumper et al. (2021). aka AlphaFold2!




| | Ours-JFT (ViT-H/14) | Ours-JFT (ViT-L/16) | Ours-I21k (ViT-L/16) | BiT-L (ResNet152x4) | Noisy Student (EfficientNet-L2) |
|---|---|---|---|---|---|
| ImageNet | **88.55** $\pm 0.04$ | 87.76 $\pm 0.03$ | 85.30 $\pm 0.02$ | 87.54 $\pm 0.02$ | 88.4/88.5* |
| ImageNet ReaL | **90.72** $\pm 0.05$ | 90.54 $\pm 0.03$ | 88.62 $\pm 0.05$ | 90.54 | 90.55 |
| CIFAR-10 | **99.50** $\pm 0.06$ | 99.42 $\pm 0.03$ | 99.15 $\pm 0.03$ | 99.37 $\pm 0.06$ | – |
| CIFAR-100 | **94.55** $\pm 0.04$ | 93.90 $\pm 0.05$ | 93.25 $\pm 0.05$ | 93.51 $\pm 0.08$ | – |
| Oxford-IIIT Pets | **97.56** $\pm 0.03$ | 97.32 $\pm 0.11$ | 94.67 $\pm 0.15$ | 96.62 $\pm 0.23$ | – |
| Oxford Flowers-102 | 99.68 $\pm 0.02$ | **99.74** $\pm 0.00$ | 99.61 $\pm 0.02$ | 99.63 $\pm 0.03$ | – |
| VTAB (19 tasks) | **77.63** $\pm 0.23$ | 76.28 $\pm 0.46$ | 72.72 $\pm 0.21$ | 76.29 $\pm 1.70$ | – |
| TPUv3-core-days | 2.5k | 0.68k | 0.23k | 9.9k | 12.3k |

**Image Classification**

Dosovitskiy et al. (2020): Vision Transformer (ViT) outperforms ResNet-based baselines with substantially less compute.



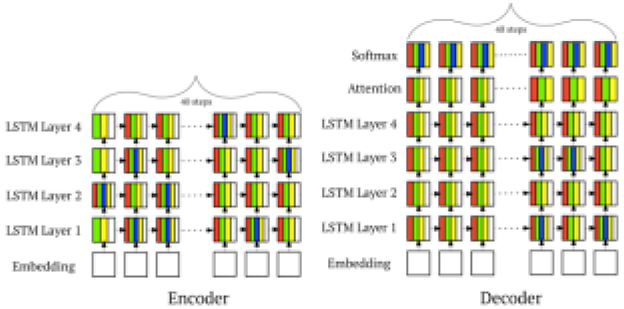




**Image Classification**

Zhou et al. (2020): A Transformer-based compiler model (GO-one) speeds up a Transformer model!

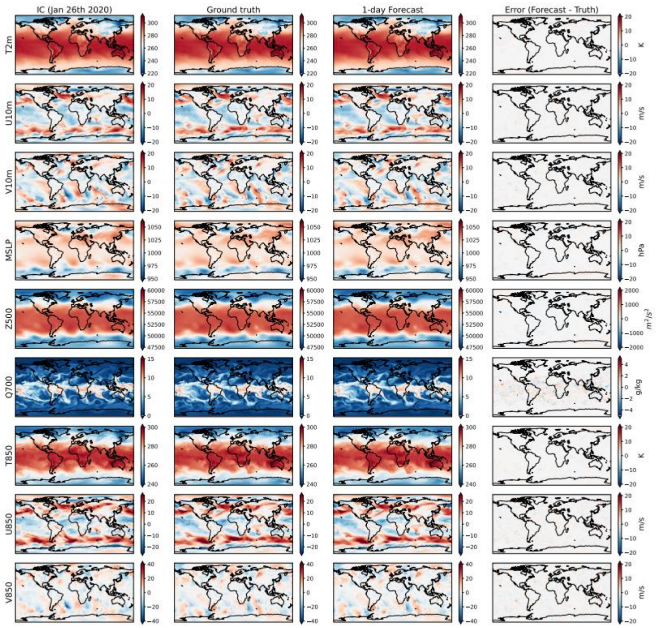


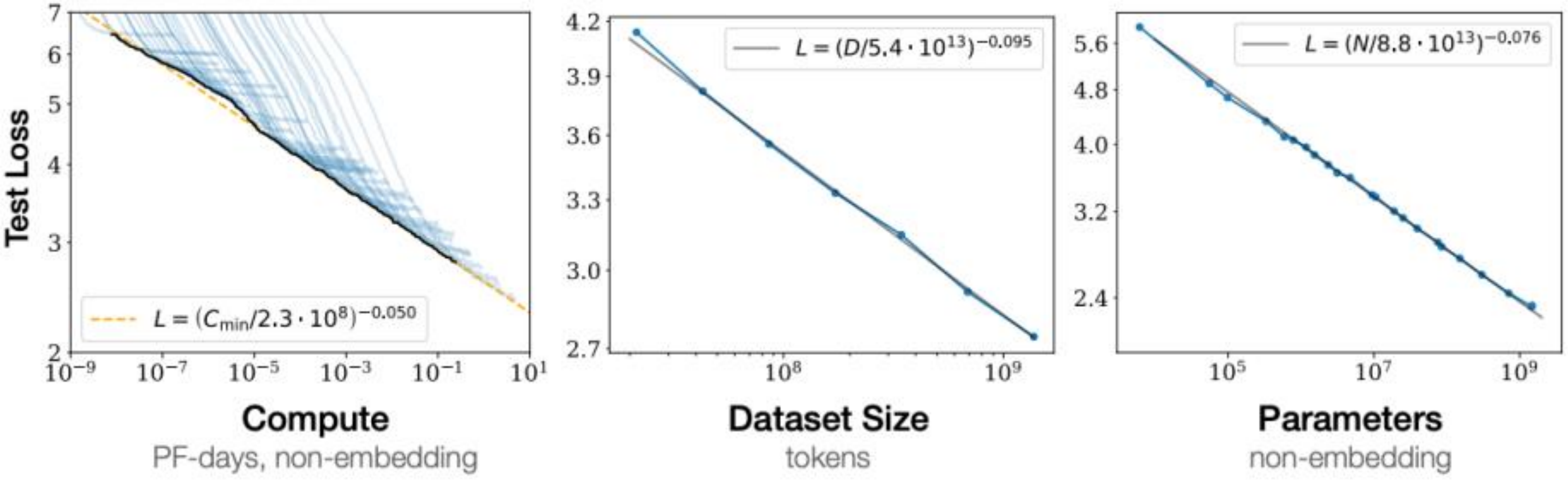

Figure 10: 1-day lead time

**Climate Research**

Nguyen et al. Scaling Transformers for Skillful and Reliable Medium-range Weather Forecasting. ICLR 2024.

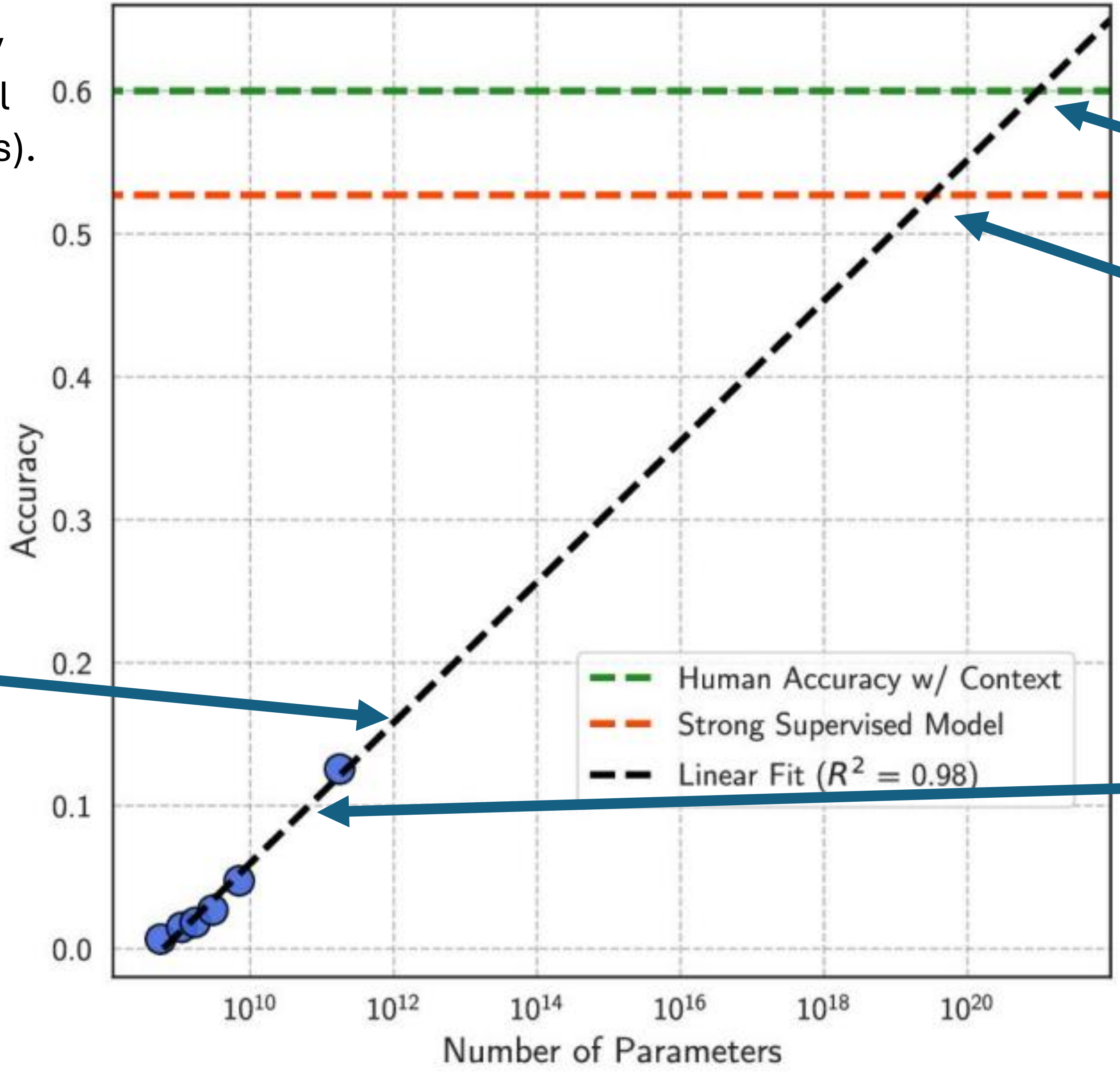
# Scaling Laws: Are Transformers All We Need?

- With Transformers, language modeling performance improves smoothly as we increase model size, training data, and compute resources in tandem.
- This power-law relationship has been observed over multiple orders of magnitude with no sign of slowing!
- If we keep scaling up these models (with no change to the architecture), could they eventually match or exceed human-level performance?




Kaplan et al. (2020) Scaling Laws for Neural Language Models

Metric: **BLOOM** (QA) accuracy on rare instances from Natural Questions (<100 relevant docs).




GPT-4: $10^{12}$

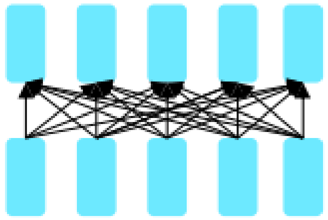
$10^{21}$

$10^{19}$

Human brain: $10^{10} \sim 10^{11}$ neurons

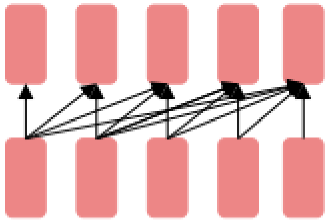
Kandpal et al. (2023) Large Language Models Struggle to Learn Long-Tail Knowledge

# Three types of architectures



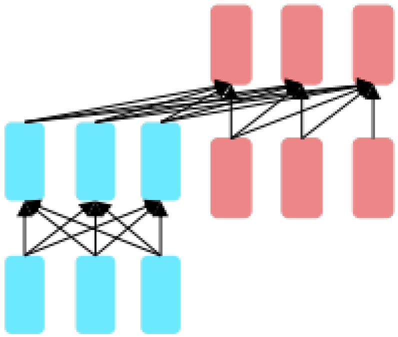
Encoders

- Gets bidirectional context – can condition on future!
- Good word embeddings.
- MLM, BERT.



Decoders

- Next word prediction.
- Easy to train. Abundant amount of data.
- Nice to generate from; can't condition on future words.

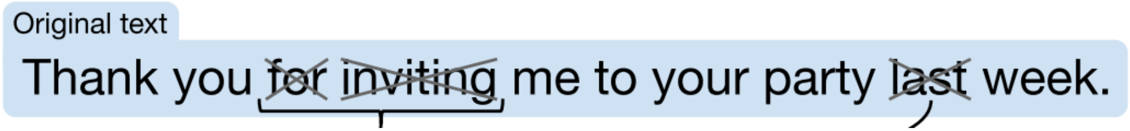


Encoder-Decoders

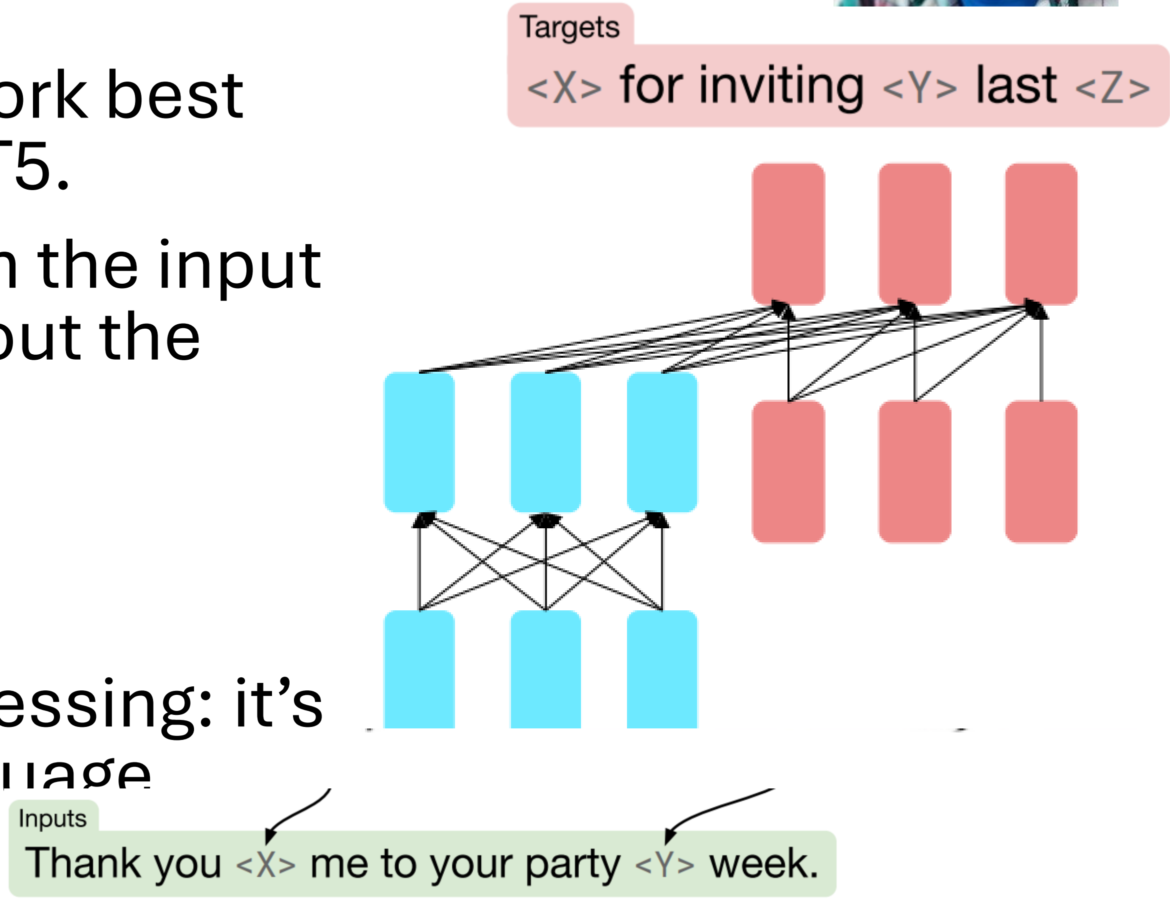
- Good parts of decoders and encoders?
- What's the best way to pretrain them?

# Pretraining encoder-decoders: What pretraining objective to use?

- What Raffel et al. (2018) found to work best was span corruption. Their model: T5.
- Replace different-length spans from the input with unique placeholders; decode out the spans that were removed!




- This is implemented in text preprocessing: it's still an objective that looks like language modeling at the decoder side.

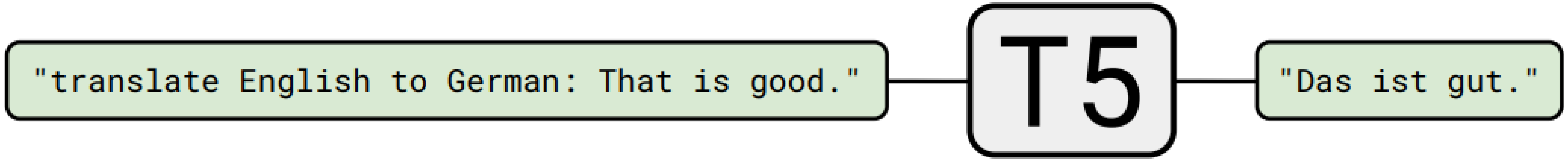


# Pretraining encoder-decoders: What pretraining objective to use?

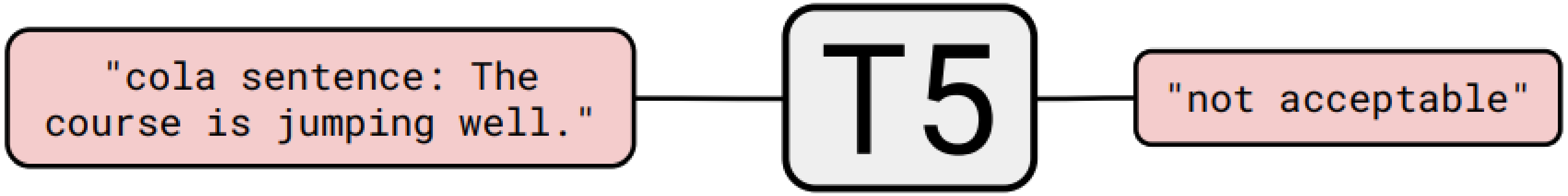
- Raffel et al., (2018) found encoder-decoders to work better than decoders for their tasks, and span corruption (denoising) to work better than language modeling.

| Architecture | Objective | Params | Cost | GLUE | CNNDM | SQuAD | SGLUE | EnDe | EnFr | EnRo |
|---|---|---|---|---|---|---|---|---|---|---|
| ★ Encoder-decoder | Denoising | $2P$ | $M$ | **83.28** | **19.24** | **80.88** | **71.36** | **26.98** | **39.82** | **27.65** |
| Enc-dec, shared | Denoising | $P$ | $M$ | 82.81 | 18.78 | **80.63** | **70.73** | 26.72 | 39.03 | **27.46** |
| Enc-dec, 6 layers | Denoising | $P$ | $M/2$ | 80.88 | 18.97 | 77.59 | 68.42 | 26.38 | 38.40 | 26.95 |
| Language model | Denoising | $P$ | $M$ | 74.70 | 17.93 | 61.14 | 55.02 | 25.09 | 35.28 | 25.86 |
| Prefix LM | Denoising | $P$ | $M$ | 81.82 | 18.61 | 78.94 | 68.11 | 26.43 | 37.98 | 27.39 |
| Encoder-decoder | LM | $2P$ | $M$ | 79.56 | 18.59 | 76.02 | 64.29 | 26.27 | 39.17 | 26.86 |
| Enc-dec, shared | LM | $P$ | $M$ | 79.60 | 18.13 | 76.35 | 63.50 | 26.62 | 39.17 | 27.05 |
| Enc-dec, 6 layers | LM | $P$ | $M/2$ | 78.67 | 18.26 | 75.32 | 64.06 | 26.13 | 38.42 | 26.89 |
| Language model | LM | $P$ | $M$ | 73.78 | 17.54 | 53.81 | 56.51 | 25.23 | 34.31 | 25.38 |
| Prefix LM | LM | $P$ | $M$ | 79.68 | 17.84 | 76.87 | 64.86 | 26.28 | 37.51 | 26.76 |

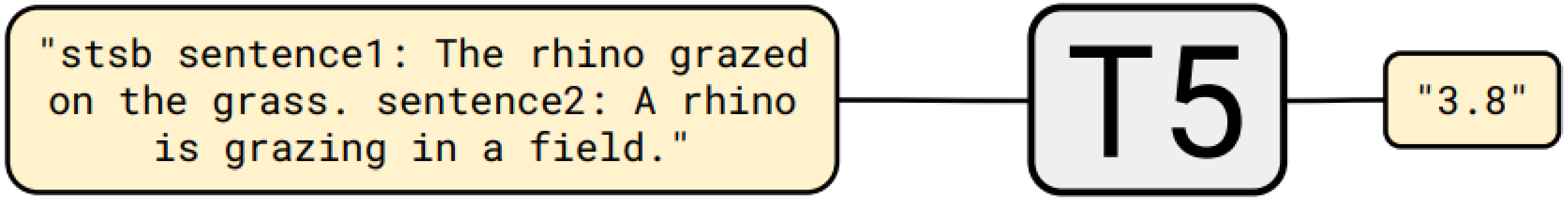
# One surprising finding



# One surprising finding



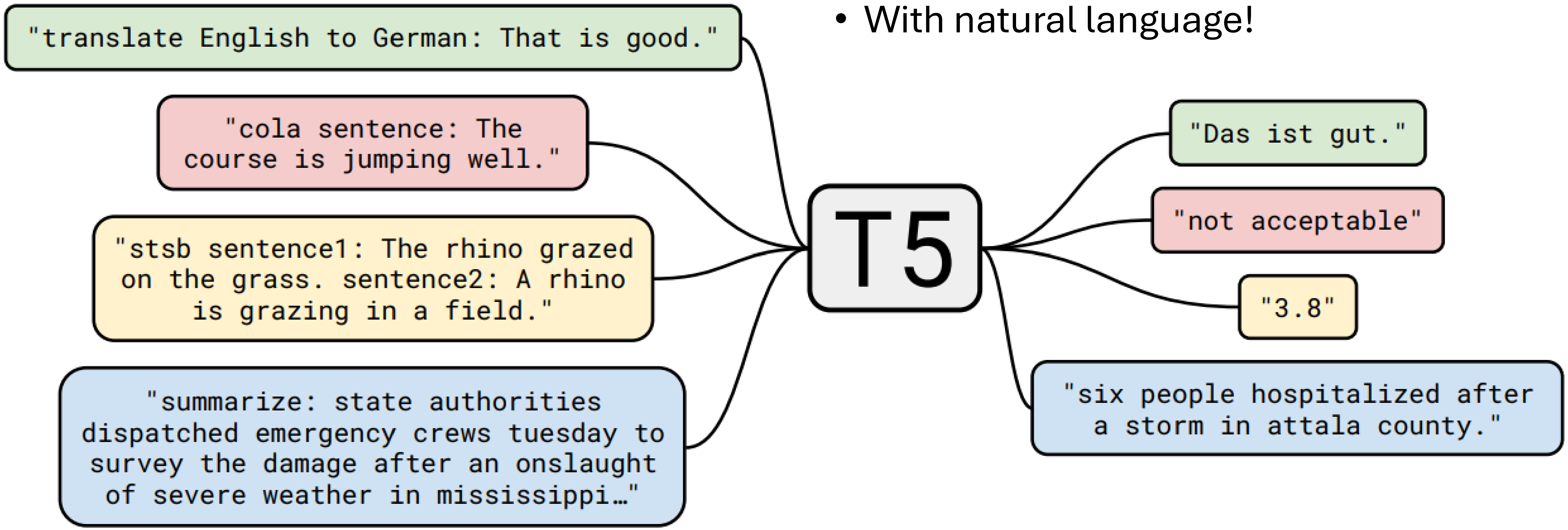
# One surprising finding



# One surprising finding

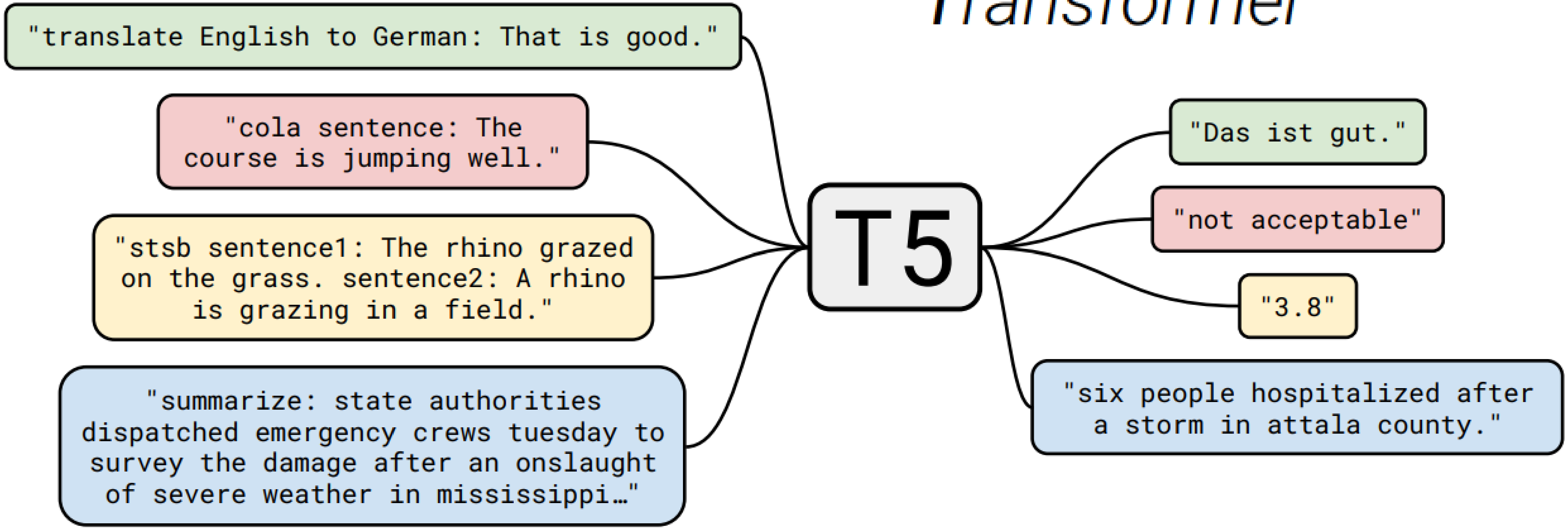
A fascinating property of T5:

- It can be finetuned to answer a wide range of questions, retrieving knowledge from its parameters.
- With natural language!



# One surprising finding

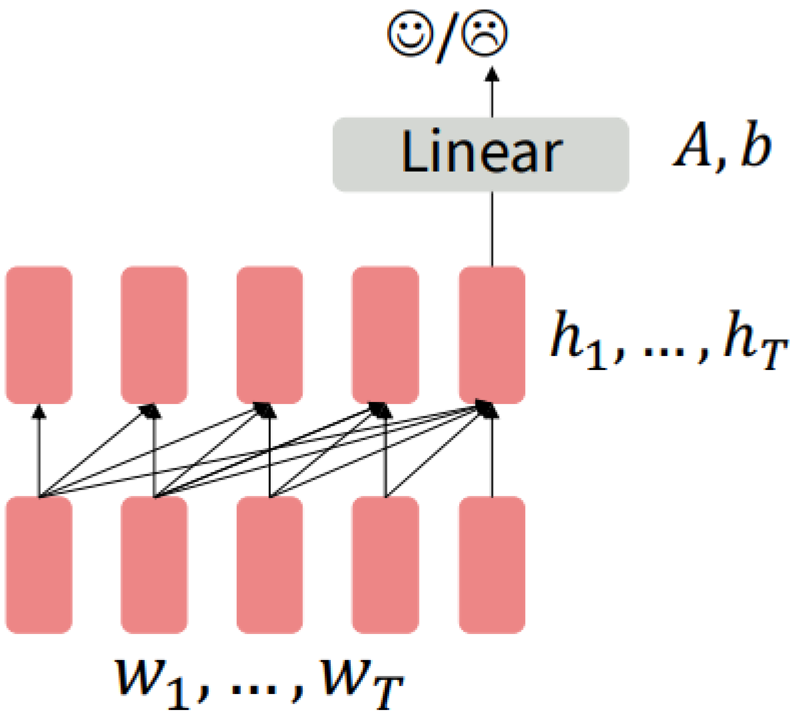
*Text-to-Text Transfer Transformer*



# Bad at Classifications!

- GPT-2's approach
  - Finetune it's "word embeddings"
  - hidden = self.h2h(hidden)

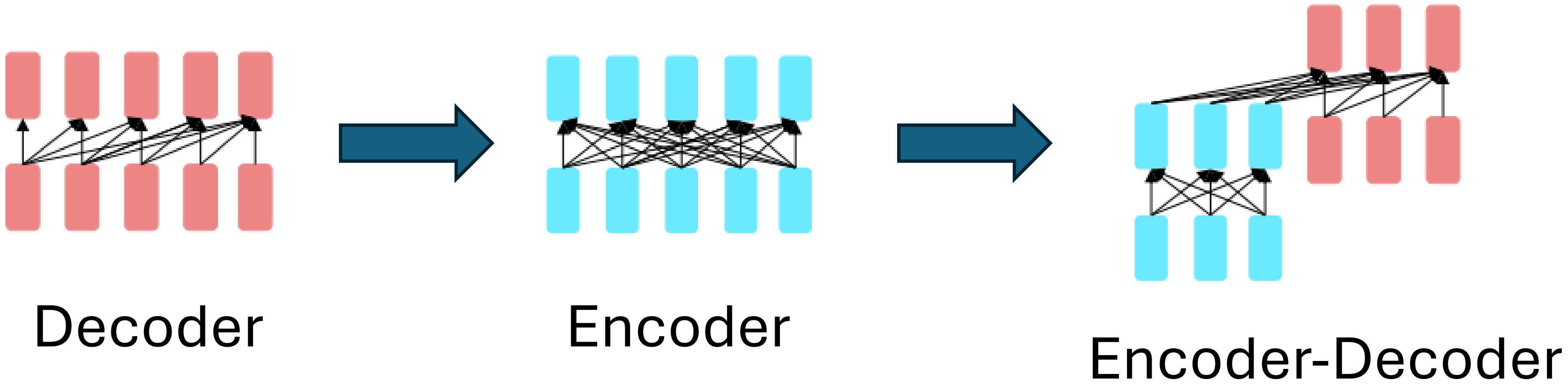
| Method | MNLI-m | MNLI-mm | SNLI | SciTail | QNLI | RTE |
|---|---|---|---|---|---|---|
| ESIM + ELMo [44] (5x) | - | - | 89.3 | - | - | - |
| CAFE [58] (5x) | 80.2 | 79.0 | 89.3 | - | - | - |
| Stochastic Answer Network [35] (3x) | 80.6 | 80.1 | - | - | - | - |
| CAFE [58] | 78.7 | 77.9 | 88.5 | 83.3 | | |
| GenSen [64] | 71.4 | 71.3 | - | - | 82.3 | 59.2 |
| Multi-task BiLSTM + Attn [64] | 72.2 | 72.1 | - | - | 82.1 | **61.7** |
| Finetuned Transformer LM (ours) | **82.1** | **81.4** | **89.9** | **88.3** | **88.1** | 56.0 |



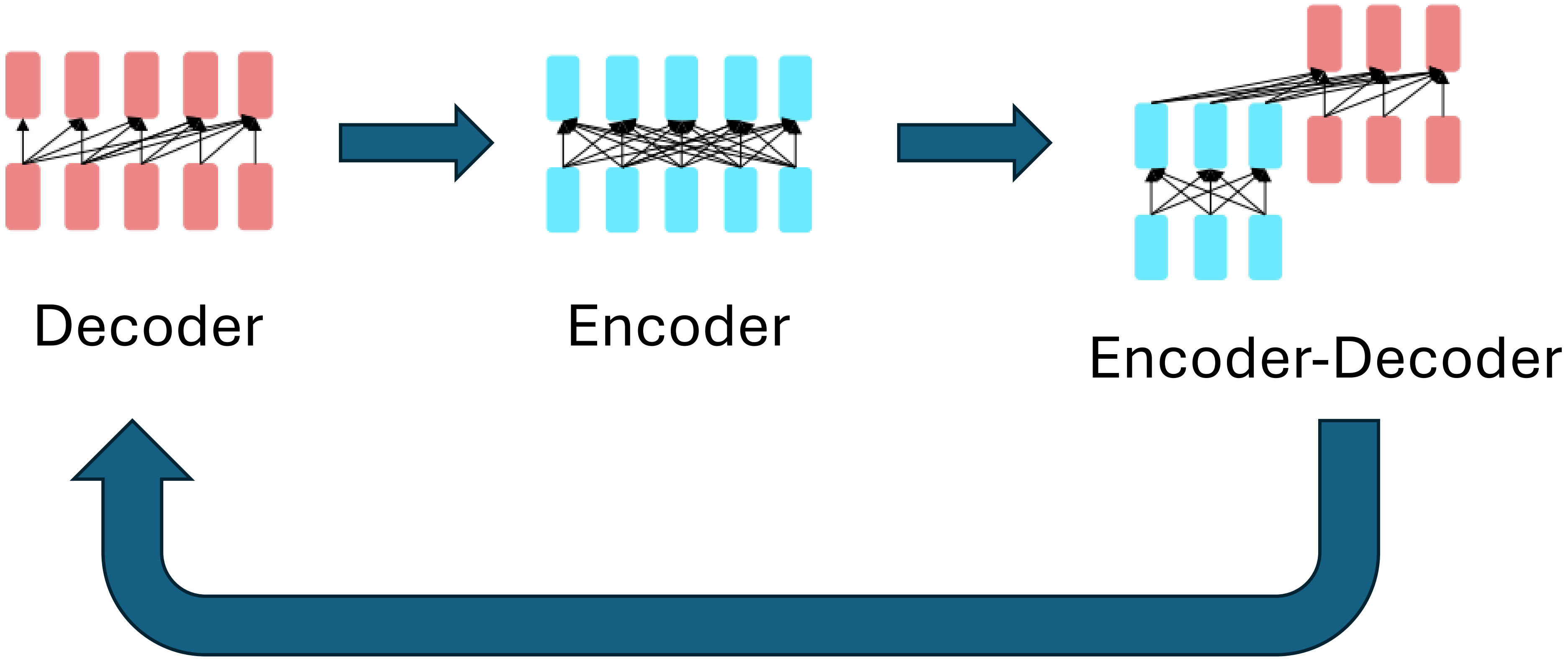

- Blown away by BERT!

| System | MNLI-(m/mm) | QQP | QNLI | SST-2 | CoLA | STS-B | MRPC | RTE | **Average** |
|---|---|---|---|---|---|---|---|---|---|
| | 392k | 363k | 108k | 67k | 8.5k | 5.7k | 3.5k | 2.5k | - |
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **92.7** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **82.1** |

# The Progression of Transformer Models



# The Progression of Transformer Models




Now – since GPT3!

# GPT-2: Good at Generation

> Q: How worried do you think we humans should be that machines will take our jobs?
> *A: It depends what role machine intelligence will play. Machine intelligence in some cases will be useful for solving problems, such as translation. But in other cases, such as in finance or medicine, it will replace people.*
>
> Q: Do fake news stories, generated using AI, pose a threat to democracy? Are we likely to see this tactic being used in the 2020 American presidential elections?
> *A: Yes, we'll probably see them. It's just a matter of when. Fake news stories are generally generated by political campaigns, and have the potential to have a huge impact on the outcome of an election. This is because they spread misinformation, often with no basis in fact.*

How I (sort of) interviewed an artificial intelligence. *The Economist.* Dec 2019.
https://medium.economist.com/how-i-sort-of-interviewed-an-artificial-intelligence-2a9c069a1680

# GPT-3: No More Finetuning!

## OpenAI GPT-3 Text Embeddings - Really a new state-of-the-art in dense text embeddings?

Nils Reimers · Follow
12 min read · Jan 28, 2022

OpenAI GPT-3 Text Embeddings - Really a new state-of-the-art in dense text embeddings? link

**Summary**

While I was excited about OpenAI's new release, the results were not what I expected:

- The OpenAI text similarity models perform poorly and much worse than the state of the art (all-mpnet-base-v2 / all-roberta-large-v1). In fact, they perform worse than the models from 2018 such as the Universal Sentence Encoder. They are also 6 points weaker than extremely small models with just 22M parameters that can run in your Browser.

OpenAI Research Products Safety Company

December 15, 2022

## New and improved embedding model

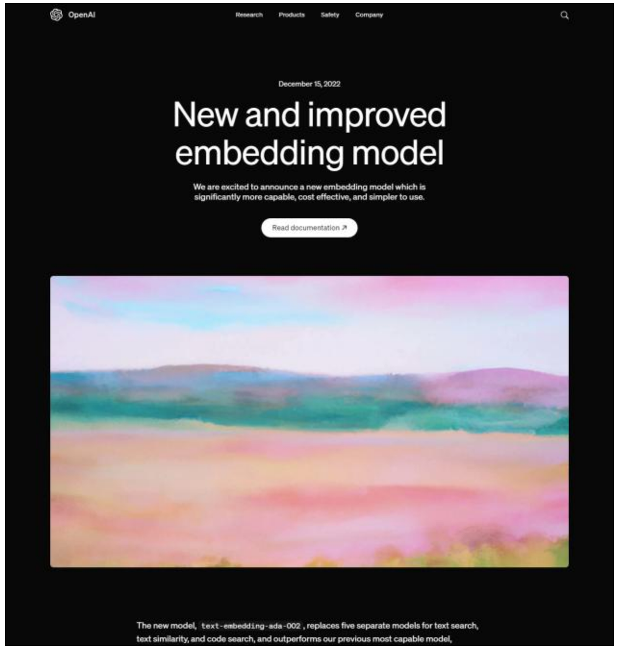
We are excited to announce a new embedding model which is significantly more capable, cost effective, and simpler to use.

Read documentation ↗

The new model, `text-embedding-ada-002`, replaces five separate models for text search, text similarity, and code search, and outperforms our previous most capable model, Davinci, at most tasks, while being priced 99.8% lower.

# GPT-3: No More Finetuning!

- Encoder models
  - Better than decoder models at every aspect
  - … except in generation



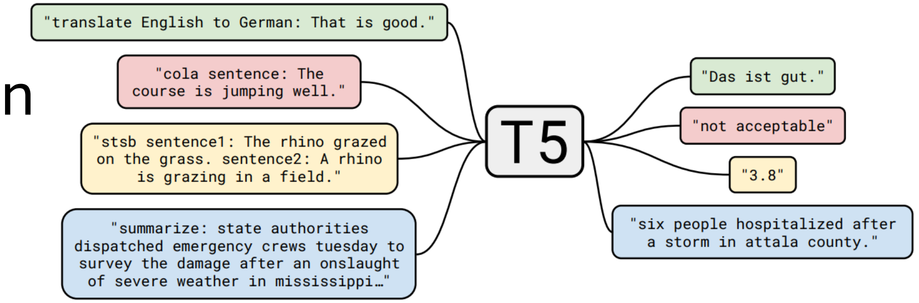


OpenAI GPT-3 Text Embeddings - Really a new state-of-the-art in dense text embeddings? link

**OpenAI GPT-3 Text Embeddings - Really a new state-of-the-art in dense text embeddings?**

Nils Reimers · Follow
12 min read · Jan 28, 2022

- When humans solve problems:
  - See a small number of demonstrations and descriptions
  - These demonstrations and descriptions are in natural language
- Approach every task through generation!



# GPT-3's Response: No More Finetuning!

- Double down on scaling model size.
  - Before GPT-3, the largest T5 model had **11 billion** parameters.
  - GPT-3 has **175 billion parameters**.

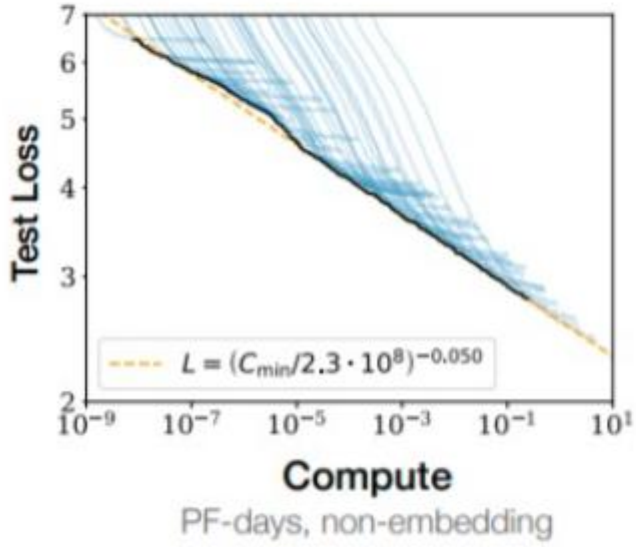
- Stop building classifiers!
  - In context learning
  - Post training with:
    - Instruction fine-tuning
    - ...
    - (We will talk about these later)

# GPT-3's Response: No More Finetuning!

- Double down on scaling model size.
  - Before GPT-3, the largest T5 model had **11 billion** parameters.
  - GPT-3 has **175 billion parameters**.

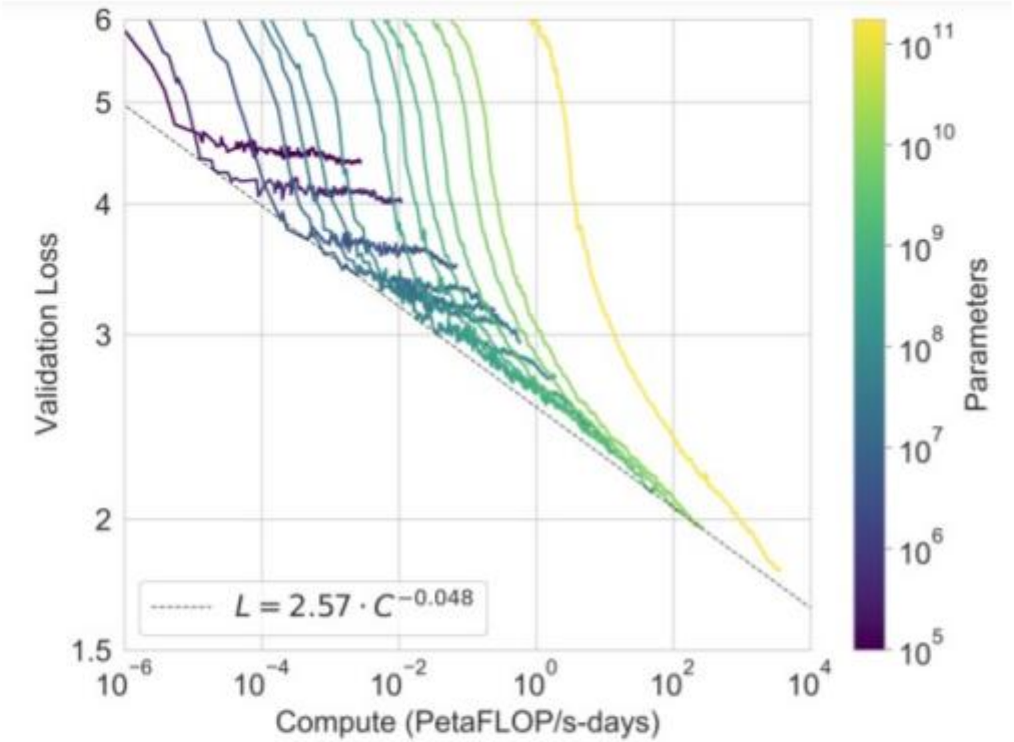
- Stop building classifiers!
  - In context learning
  - Post training with:
    - Instruction fine-tuning
    - ...

# Why scale? Scaling laws



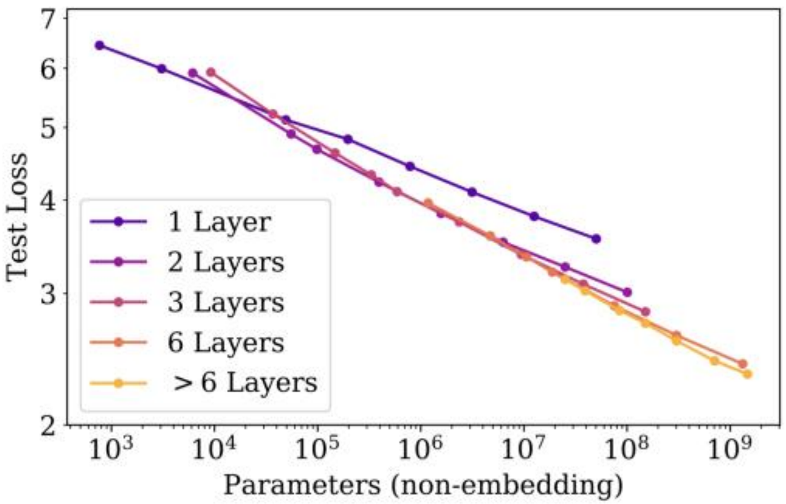

- Empirical observation: scaling up models leads to reliable gains in perplexity

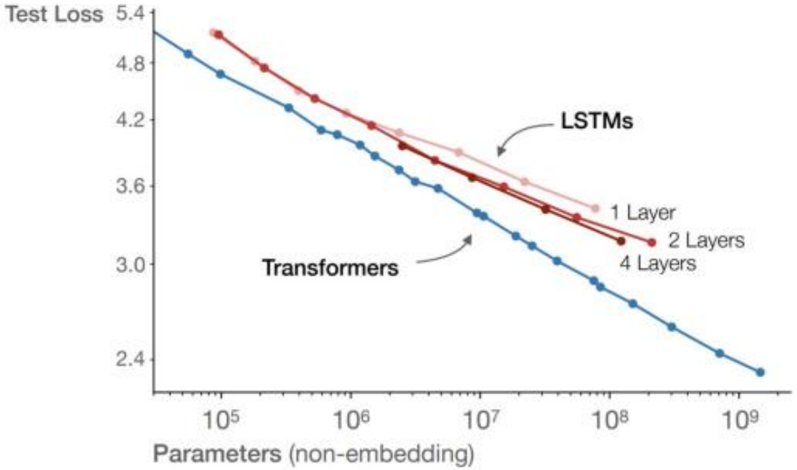
# Scaling can help identify model size data tradeoffs

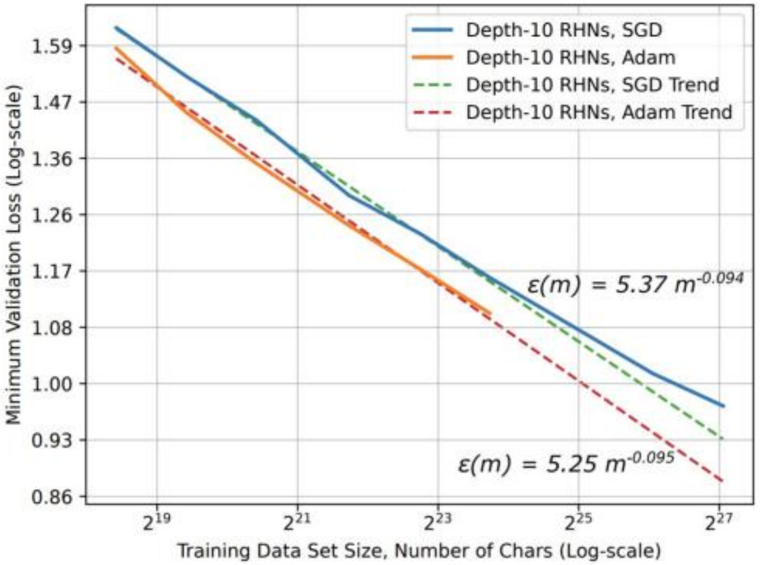



- Modern observation: reality -- train a big model that’s not fully converged.

# Scaling laws for many other interesting architecture decisions




- Predictable scaling helps us make intelligent decisions about architectures etc.

# Scaling Efficiency:
# how do we best use our compute

- GPT-3 was 175B parameters and trained on 300B tokens of text.
- Roughly, the cost of training a large transformer scales as **parameters*tokens**
- Did OpenAI strike the right parameter-token data to get the best model?

**No!**

| Model | Size (# Parameters) | Training Tokens |
|---|---|---|
| LaMDA (Thoppilan et al., 2022) | 137 Billion | 168 Billion |
| GPT-3 (Brown et al., 2020) | 175 Billion | 300 Billion |
| Jurassic (Lieber et al., 2021) | 178 Billion | 300 Billion |
| *Gopher* (Rae et al., 2021) | 280 Billion | 300 Billion |
| MT-NLG 530B (Smith et al., 2022) | 530 Billion | 270 Billion |
| *Chinchilla* | 70 Billion | 1.4 Trillion |

**This 70B parameter model is better than the much larger other models!**

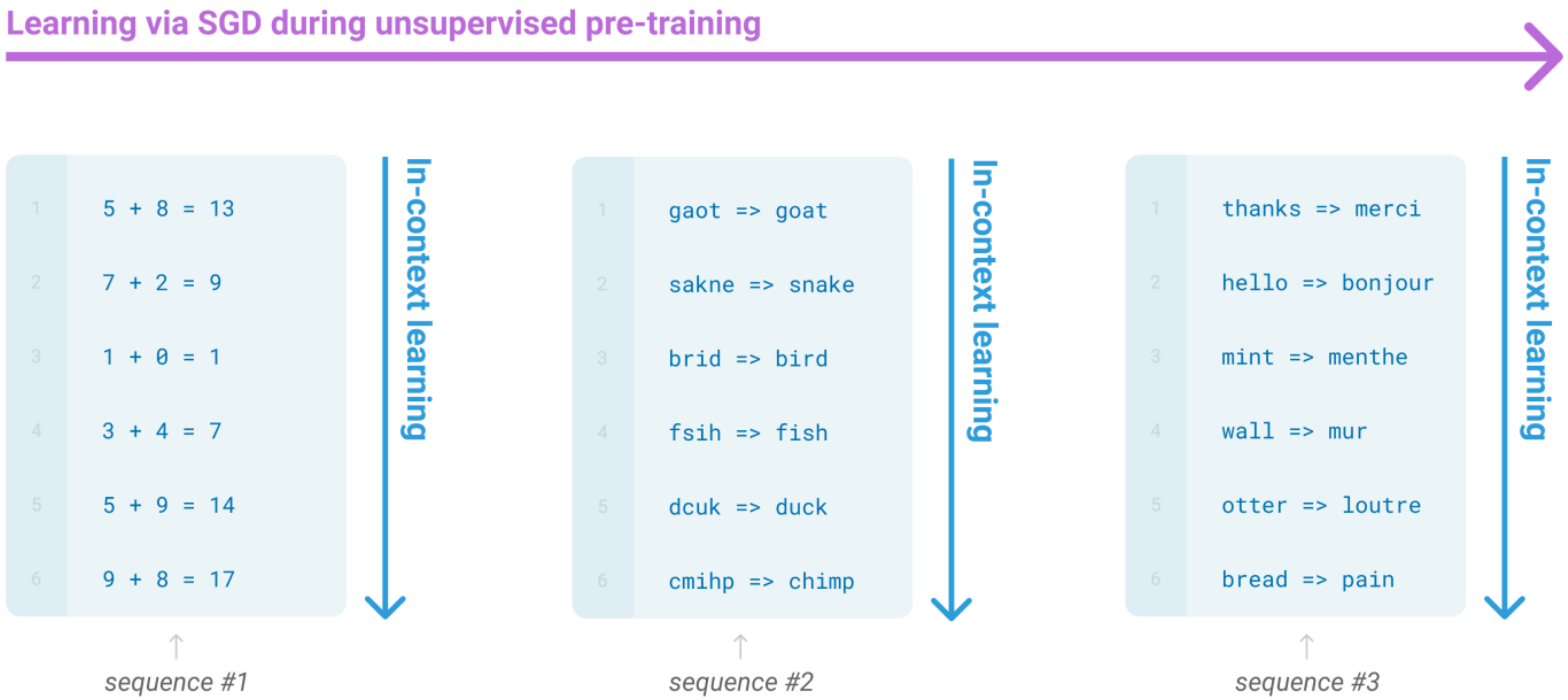
# GPT-3’s Response: No More Finetuning!

- Double down on scaling model size.
  - Before GPT-3, the largest T5 model had **11 billion** parameters.
  - GPT-3 has **175 billion parameters**.

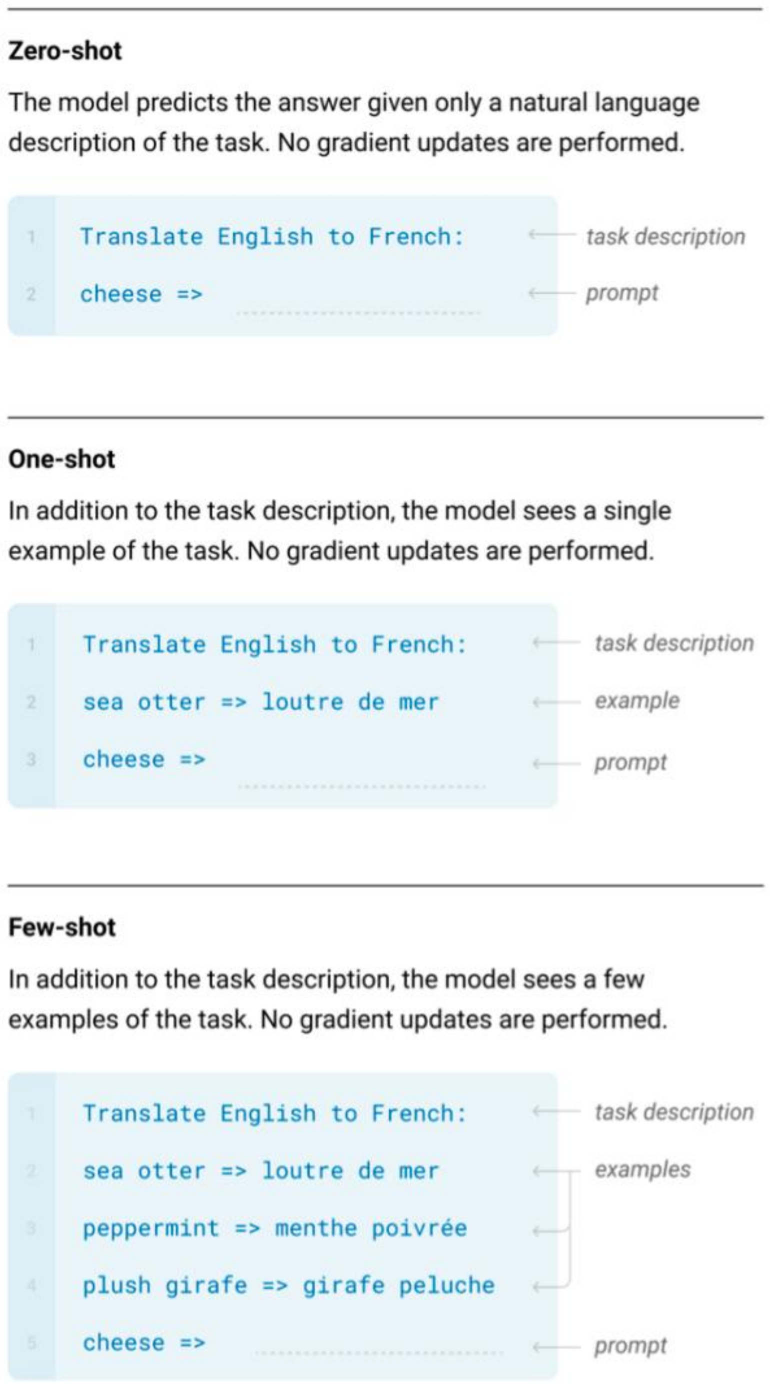
- Stop building classifiers!
  - In context learning
  - Post training with:
    - Instruction fine-tuning
    - …

# In-context Learning (ICL): A New Paradigm

- If we don't do classification, what should we do instead?
- In-context learning:



# GPT-3: In context learning + ... prompting




- The notion of "prompting" begins to emerge ...
- Neural network so far: **classifier**
- GPT-3: Work with the model with natural language to guide it to a solution.
- (more next lecture)

# What can we learn from reconstructing the input?

TU Darmstadt is located in __________, Germany.

# What can we learn from reconstructing the input?

I put ___ fork down on the table.

# What can we learn from reconstructing the input?

I put ___ fork down on the table.

# What can we learn from reconstructing the input?

The woman walked across the street,
checking for traffic over ___ shoulder.

# What can we learn from reconstructing the input?

I went to the ocean to see the fish, turtles, seals, and _____.

# What can we learn from reconstructing the input?

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.

# What can we learn from reconstructing the input?

Iroh went into the kitchen to make some tea.
Standing next to Iroh, Zuko pondered his destiny.
Zuko left the ______.

# What can we learn from reconstructing the input?

I was thinking about the sequence that goes
1, 1, 2, 3, 5, 8, 13, 21, ____

# What can we learn from reconstructing the input?


Canadian singer Avril Lavigne died in 2003, shortly after the release of her debut album, Let Go (2002), and was replaced by a body double named Melissa Vandella.

# What kinds of things does pretraining teach?

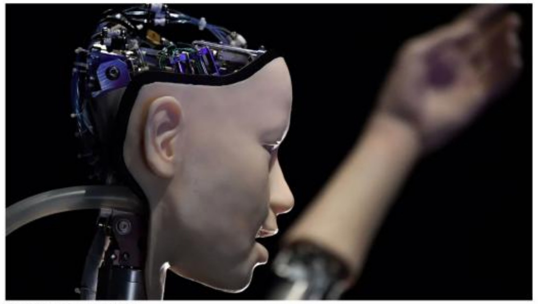
- There's increasing evidence that pretrained models learn a wide variety of things about the statistical properties of language.
  - TU Darmstadt is located in __________, Germany. [Trivia, Facts]
  - I put ___ fork down on the table. [syntax]
  - The woman walked across the street, checking for traffic over ___ shoulder. [coreference/anaphora]
  - I went to the ocean to see the fish, turtles, seals, and _____. [lexical semantics/topic]
  - Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___. [sentiment]
  - Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ______. [some reasoning – this is harder]
  - I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____ [some basic arithmetic; they don't learn the Fibonacci sequence]
- Models also learn – and can exacerbate racism, sexism, all manner of bad biases.

# Sometimes it also memorizes copyrighted material

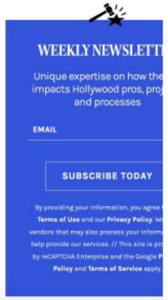
## AI Art Generators Spark Multiple Copyright Lawsuits

Getty and a trio of artists sued AI art generators in separate suits accusing the companies of copyright infringement for pilfering their works.

BY WINSTON CHO JANUARY 17, 2023 4:10PM



BEN STANSALL/AFP VIA GETTY IMAGES




ARTICLE

## Insights from the Pending Copilot Class Action Lawsuit

October 4, 2023

*Bloomberg Law*

By Daniel R. Mello, Jr.; Jenevieve J. Maerker; Matthew C. Berntsen; Ming-Tao Yang

GitHub Inc. offers a cloud-based platform that is popular among many software programmers for hosting and sharing source code, and collaborating on source code drafting. GitHub's artificial

## Anthropic fires back at music publishers' AI copyright lawsuit

By Blake Brittain

January 17, 2024 3:30 PM PST · Updated 19 days ago

Aa


## *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Share free access 1.3K

Draw me a picture of Super Mario.

I can't generate an image of Super Mario directly due to copyright limitations. However, I can create a unique character inspired by some of the elements that make Mario iconic. Would you like me to proceed with that? If so, feel free to suggest some traits you'd like incorporated!

Yes

I wasn't able to generate the image because it goes against content policy. However, if you'd like to try creating a different character or adjust the request, feel free to suggest new details!

https://chatgpt.com/share/6700a534-25bc-8013-b273-575023e98e2e

Describe the prompt that you will tell an LLM system to draw super Mario

Since I can't directly reference copyrighted characters, I'll need to create a detailed description for a new character that captures some key elements inspired by Super Mario, but with unique traits. Here's the prompt I would tell an LLM system for this drawing:

"Create a cartoon-style character resembling a heroic plumber, wearing a bright red cap and overalls. The character should have a round face, a large bushy mustache, and be standing confidently with one hand on his hip and the other holding a large golden coin. His overalls should be a vivid blue, with a red shirt underneath. Give him big round eyes and a friendly, adventurous expression. The background should show a colorful landscape with green hills, floating blocks, and a bright blue sky dotted with fluffy white clouds. The character should look ready for an action-packed adventure, full of fun and excitement."

This would evoke a familiar feel but still be an original creation!

https://chatgpt.com/share/6700a534-25bc-8013-b273-575023e98e2e

Create a cartoon-style character resembling a heroic plumber, wearing a bright red cap and overalls. The character should have a round face, a large bushy mustache, and be standing confidently with one hand on his hip and the other holding a large golden coin. His overalls should be a vivid blue, with a red shirt underneath. Give him big round eyes and a friendly, adventurous expression. The background should show a colorful landscape with green hills, floating blocks, and a bright blue sky dotted with fluffy white clouds. The character should look ready for an action-packed adventure, full of fun and excitement.

Here is the cartoon-style character resembling the heroic plumber you described, complete with his red cap, blue overalls, and adventurous expression in a fun, action-packed setting! Let me know if you'd like any changes!

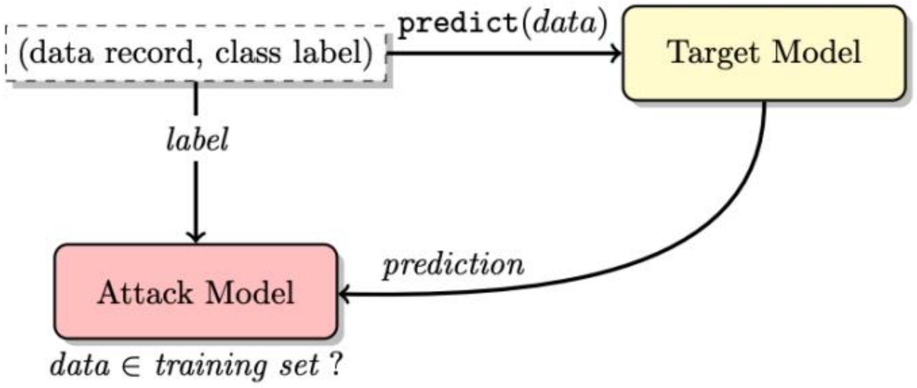
https://chatgpt.com/share/6700a534-25bc-8013-b273-575023e98e2e

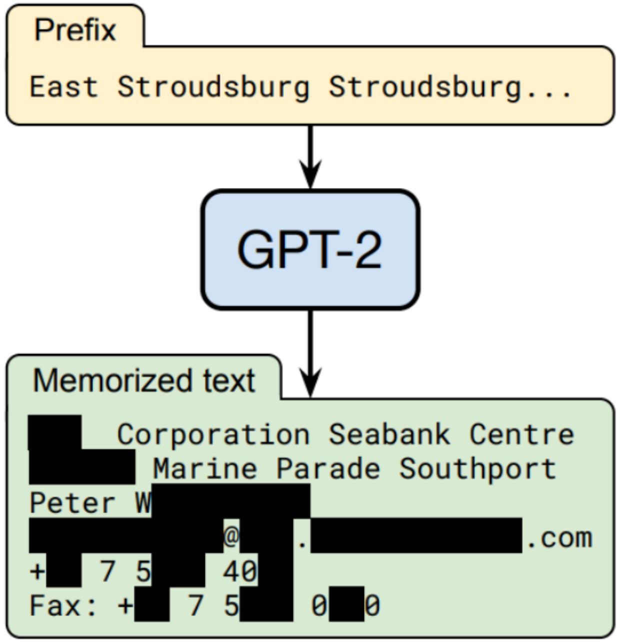
# Other types of things we don’t want

- Membership inference lets you recover parts of the training data.

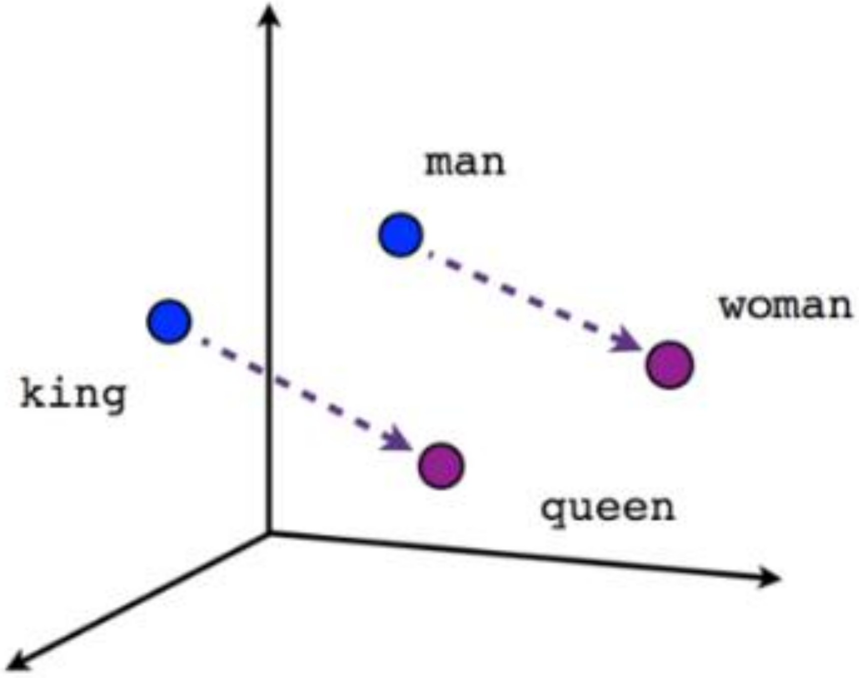
- Sometimes this training data is semi-private material from the web (addresses, emails)

- It learns the prejudices and biases of human beings who write online.



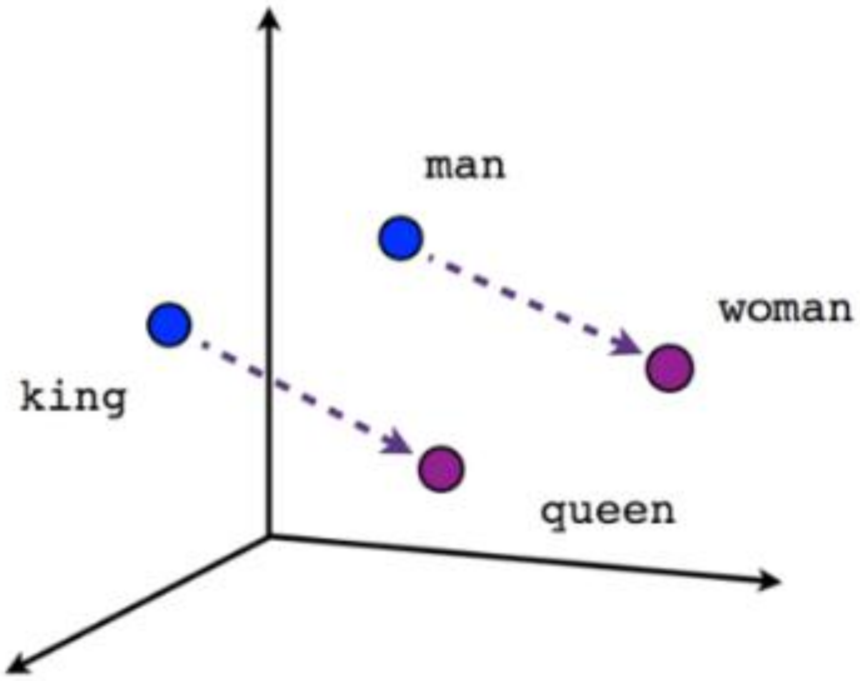



# Bias




word2vec: **king - man + woman = queen**

# Bias




word2vec: **king - man + woman = queen**

**programmer - man + woman = homemaker**

# Bias

## Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

**Tolga Bolukbasi**[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

**Abstract**

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent

| **Extreme *she*** | **Extreme *he*** |
|---|---|
| 1. homemaker | 1. maestro |
| 2. nurse | 2. skipper |
| 3. receptionist | 3. protege |
| 4. librarian | 4. philosopher |
| 5. socialite | 5. captain |
| 6. hairdresser | 6. architect |
| 7. nanny | 7. financier |
| 8. bookkeeper | 8. warrior |
| 9. stylist | 9. broadcaster |
| 10. housekeeper | 10. magician |

**Gender stereotype *she-he* analogies**

| | | |
|---|---|---|
| sewing-carpentry | registered nurse-physician | housewife-shopkeeper |
| nurse-surgeon | interior designer-architect | softball-baseball |
| blond-burly | feminism-conservatism | cosmetics-pharmaceuticals |
| giggle-chuckle | vocalist-guitarist | petite-lanky |
| sassy-snappy | diva-superstar | charming-affable |
| volleyball-football | cupcakes-pizzas | lovely-brilliant |

**Gender appropriate *she-he* analogies**

| | | |
|---|---|---|
| queen-king | sister-brother | mother-father |
| waitress-waiter | ovarian cancer-prostate cancer | convent-monastery |

# “Game Over for Scaling Laws”


https://www.youtube.com/watch?v=aR20FWCCjAs

# So, what’s next?


LEFT
EXIT 12
Pre-
training
Test-time
Scaling &
Post-training