# Learning to Make Documents Barrier-free and Accessible

## Language Models are Few-Shot Learners

Jack Heseltine, December 2023 - for: Machine Learning Institute
with: Institute for Application-oriented Knowledge Processing and Institute Integriert Studieren

JOHANNES KEPLER
UNIVERSITY LINZ

# Overview

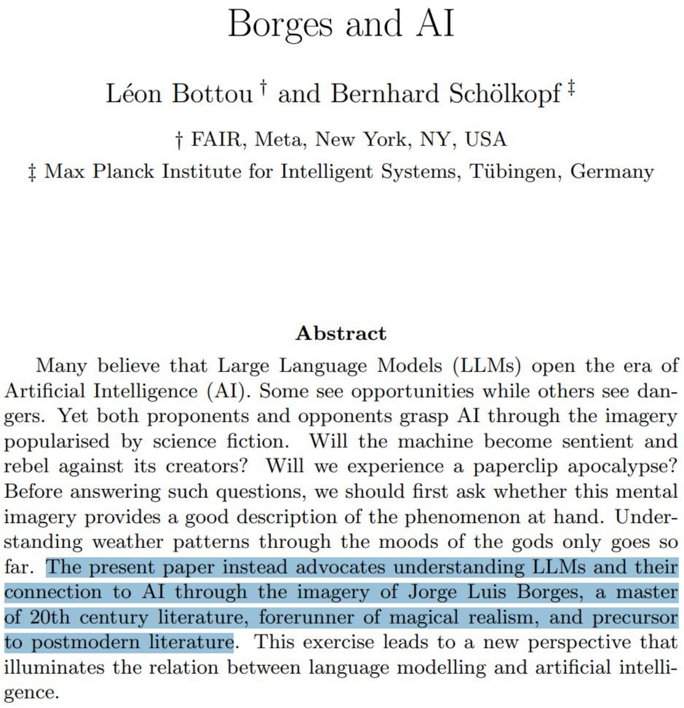
- Want to contrast the Brown et al. GPT-3 Open AI Paper **Language Models are Few-Shot Learners** (FSL in the slides/Brown et al. Open AI GPT-3 Paper, 2020)
    - with: **Language Models are Unsupervised Multitask Learners** (UML/Radford et al. Open AI GPT-2 Paper, 2019)
    - looking at model architecture, dataset sizes, tasks, approaches and concepts
    - => Idea is to uncover the persistent themes and modes of evaluating Language Model Few Shot (Multi-)task performance
    - While giving rough idea of these **huge papers**
- *Motivation: Barrier-free documents with Language Models, that is automating previously labor intensive task (PDF annotation) for direct benefits here at JKU — but the specifics of this task are not part of this presentation today*

# Introduction/Side: Borges Paper

*To open our thinking after a long day, let's take a moment to look at a paper that is a little more unusual*

- impulses from the arts lead to new ideas in how to apply these technologies, but in what genre?
- one idea as an example from magical realism* follows
- emotional aspect of LLM use today, examples
    - awe: Bubeck et al., Sparks of artificial general intelligence: Early experiments with GPT-4
    - fear: Bostrom, Superintelligence/Bengio et al., Pause giant AI Experiments: An open letter
    - greed: Metz et al., AI funding frenzy escalates/Microsoft bets big (New York Times)

*"a realistic view of the world while also adding magical elements": Magic realism - Wikipedia



Borges and AI

Léon Bottou [†] and Bernhard Schölkopf [‡]

† FAIR, Meta, New York, NY, USA
‡ Max Planck Institute for Intelligent Systems, Tübingen, Germany

**Abstract**

Many believe that Large Language Models (LLMs) open the era of Artificial Intelligence (AI). Some see opportunities while others see dangers. Yet both proponents and opponents grasp AI through the imagery popularised by science fiction. Will the machine become sentient and rebel against its creators? Will we experience a paperclip apocalypse? Before answering such questions, we should first ask whether this mental imagery provides a good description of the phenomenon at hand. Understanding weather patterns through the moods of the gods only goes so far. The present paper instead advocates understanding LLMs and their connection to AI through the imagery of Jorge Luis Borges, a master of 20th century literature, forerunner of magical realism, and precursor to postmodern literature. This exercise leads to a new perspective that illuminates the relation between language modelling and artificial intelligence.

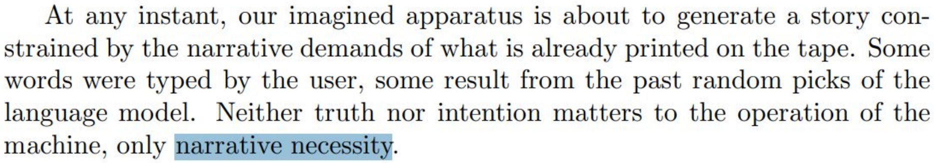
[2310.01425] Borges and AI (arxiv.org)

# Introduction/Side: Borges Paper

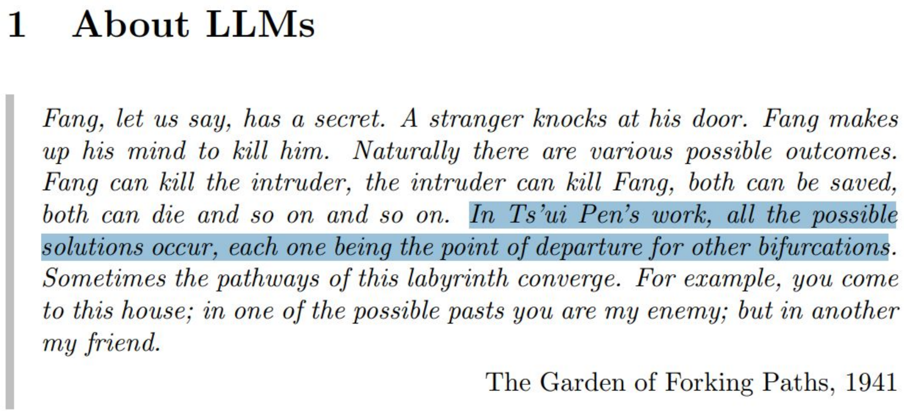
The Garden of Forking Paths, 1941 short story.

Ideas just from section 1 of “Borges and AI”:

- Collection of everything ever written and a long tape, “the perfect language model,” that is in dialogue with the collection
- Idea of a “garden” of **all plausible texts** that are approximated like a transcendental no. can the be arranged to the appropriate output according to **narrative necessity** (model and prompt)



At any instant, our imagined apparatus is about to generate a story constrained by the narrative demands of what is already printed on the tape. Some words were typed by the user, some result from the past random picks of the language model. Neither truth nor intention matters to the operation of the machine, only narrative necessity.

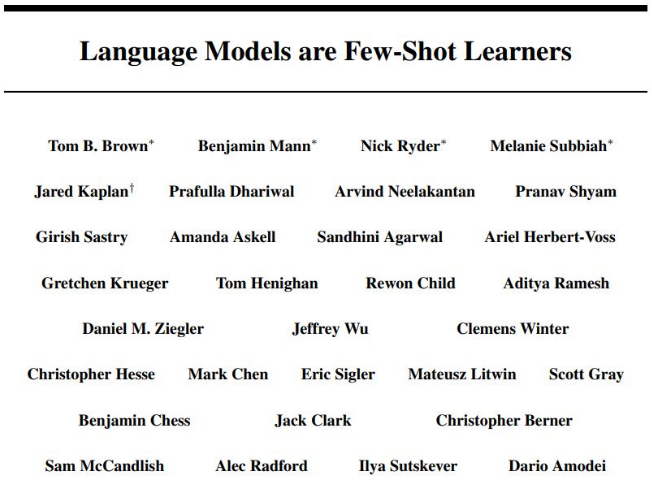


## 1 About LLMs

> *Fang, let us say, has a secret. A stranger knocks at his door. Fang makes up his mind to kill him. Naturally there are various possible outcomes. Fang can kill the intruder, the intruder can kill Fang, both can be saved, both can die and so on and so on. In Ts’ui Pen’s work, all the possible solutions occur, each one being the point of departure for other bifurcations. Sometimes the pathways of this labyrinth converge. For example, you come to this house; in one of the possible pasts you are my enemy; but in another my friend.*
>
> The Garden of Forking Paths, 1941

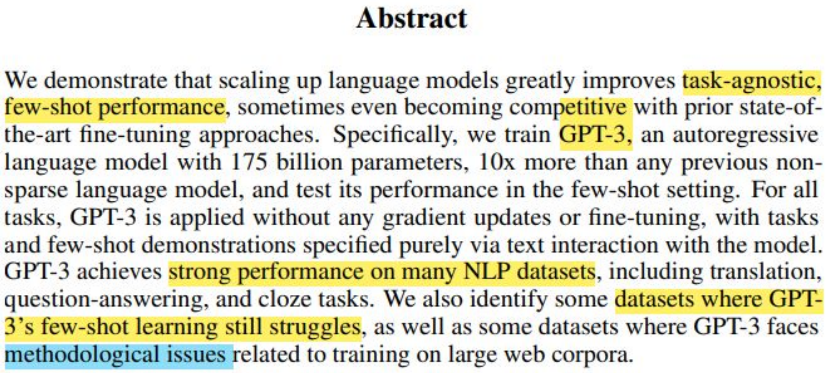
- For me also important the question of evaluating, insights from literature, esp. “murkier” or less linear tasks?

# >> Language Models are Few-Shot Learners: Overview



## Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam
Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss
Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh
Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei



### Abstract

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

- **few-shot instead of fine-tuning**
  with a look at the relevant GPT-2 paper
- **GPT-3 introduction to the model**
  + training datasets and results
- **NLP datasets/tasks where few-shot performance is good …**
- **… and where there are struggles, with a focus on Few-Shot***
  - methodological issues related to training? + limitations

[*1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (neurips.cc)](https://neurips.cc) (NeurIPS 2020, shorter) and [[2005.14165] Language Models are Few-Shot Learners (arxiv.org)](https://arxiv.org) (22 Jul 2020, v4)

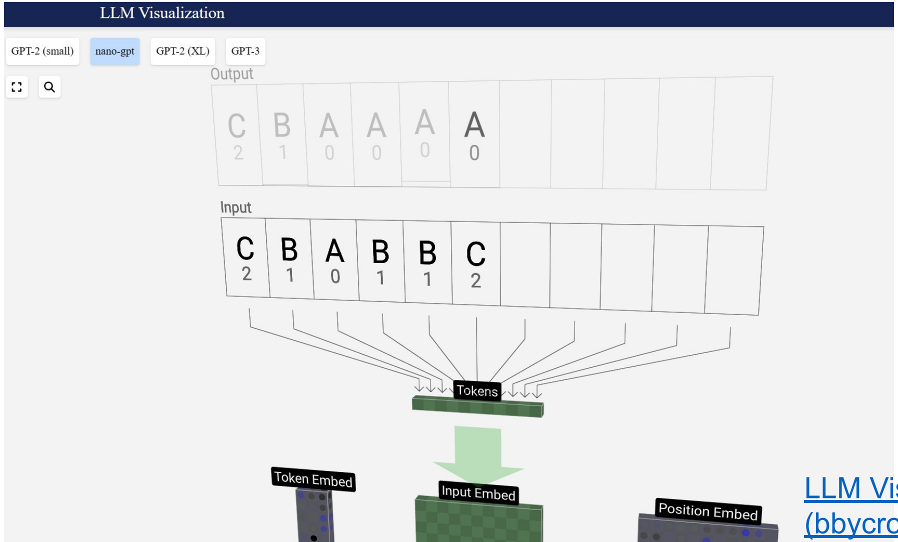
* "Showing Examples" - we will also talk about terminology

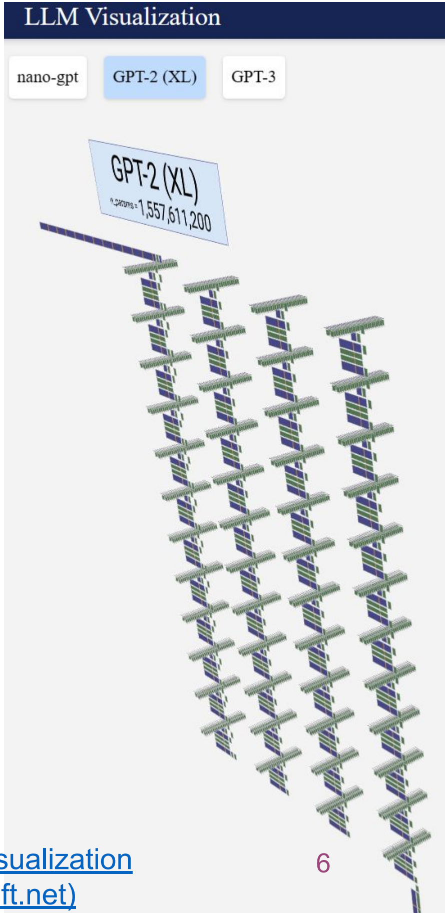
JOHANNES KEPLER
UNIVERSITY LINZ

# Language Models are Few-Shot Learners: Background

- **Autoregressive Language Models (GPT-2 and -3 are AR models)**
- **(Sparse) Transformers (used in GPT-3), Attention Mechanism**
- **Byte Pair Encoding and other encoding techniques**

… For an LLM intro:



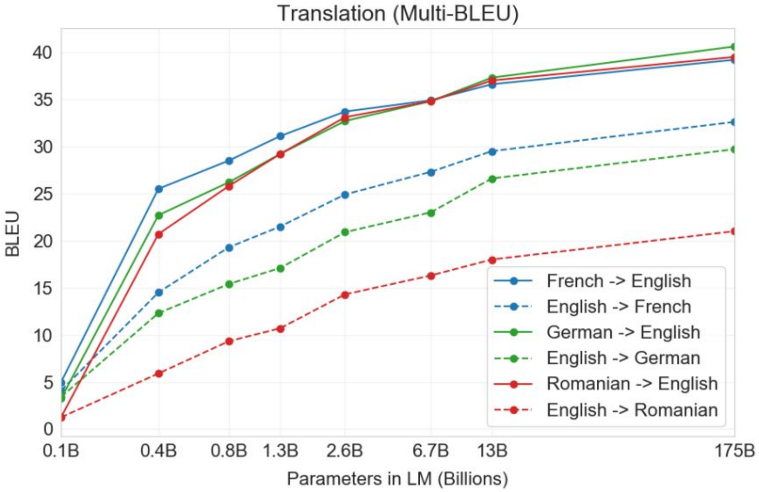


LLM Visualization (bbycroft.net)

# Language Models are Few-Shot Learners: Unsupervised Multitask Learners

The other major background to this paper is another Open AI Paper, but the core concept travels

- the paper: [Language Models Are Unsupervised Multitask Learners (UML/Radford et al.)]()
- the core concept as it is picked up in the Few-Shot Learners Paper however: per the example of **translation**
- from the first paper: “language models begin to learn these tasks without any explicit supervision when trained on a **new dataset of millions of webpages called WebText**” (from the abstract, **UML**)



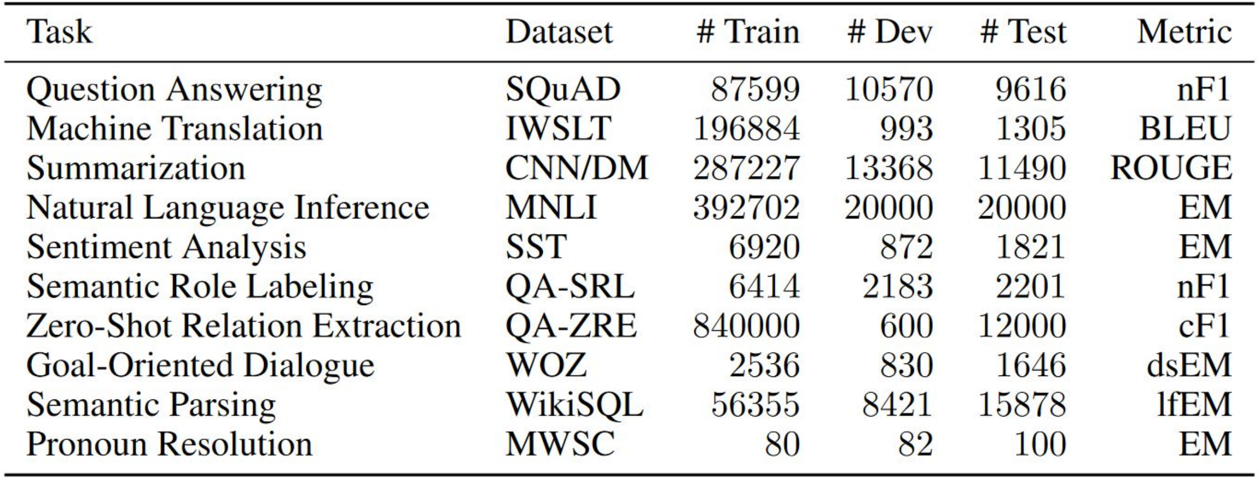
FSL, Fig. 3.4

**Figure 3.4:** Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and as well as tendency for translation into English to be

# Language Models are Few-Shot Learners: Unsupervised Multitask Learners

The other major background to this paper is another Open AI Paper, but the core concept travels

- from the first paper (UML): a translation problem might be specified as `(translate to french, english text, french text)` as a flexible way to specify tasks by language
- it was actually first shown by [McCann et al. (2018): The natural language decathlon: Multitask learning as question answering](#) that it is possible to train one model (MQAN 2018) to infer and perform many tasks on examples with this type of format, led to decaNLP
- we will pick up on some later



| Task | Dataset | # Train | # Dev | # Test | Metric |
|---|---|---|---|---|---|
| Question Answering | SQuAD | 87599 | 10570 | 9616 | nF1 |
| Machine Translation | IWSLT | 196884 | 993 | 1305 | BLEU |
| Summarization | CNN/DM | 287227 | 13368 | 11490 | ROUGE |
| Natural Language Inference | MNLI | 392702 | 20000 | 20000 | EM |
| Sentiment Analysis | SST | 6920 | 872 | 1821 | EM |
| Semantic Role Labeling | QA-SRL | 6414 | 2183 | 2201 | nF1 |
| Zero-Shot Relation Extraction | QA-ZRE | 840000 | 600 | 12000 | cF1 |
| Goal-Oriented Dialogue | WOZ | 2536 | 830 | 1646 | dsEM |
| Semantic Parsing | WikiSQL | 56355 | 8421 | 15878 | lfEM |
| Pronoun Resolution | MWSC | 80 | 82 | 100 | EM |

Table 1 in McCann et al. (2018)

# Language Models are Few-Shot Learners: Unsupervised Multitask Learners

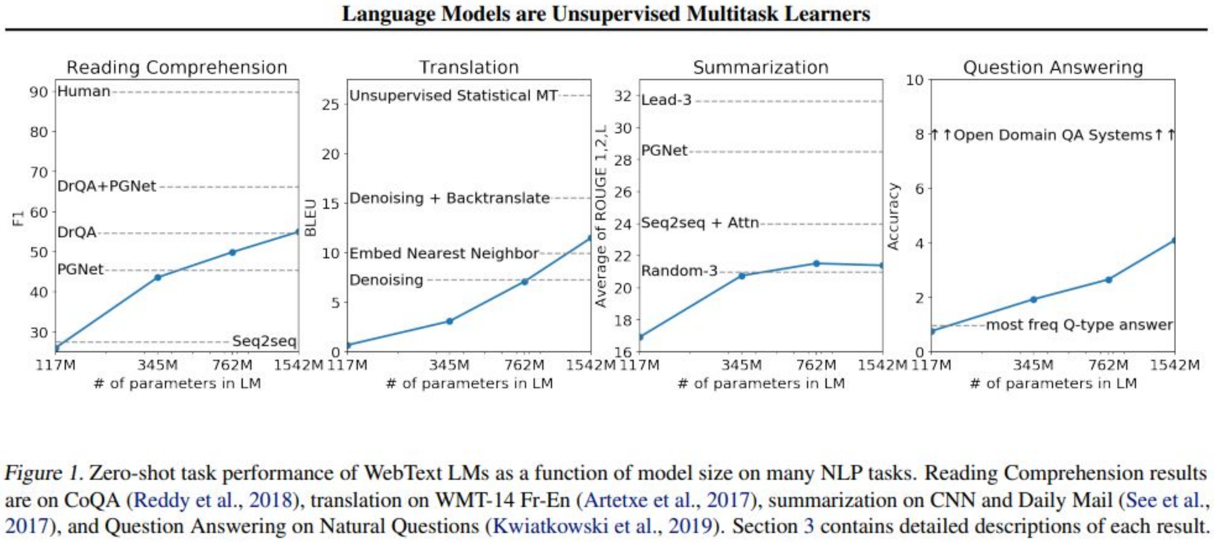
Aside from the McCann paper, this paper is cited as **inspirational** to the work in UML:

- Liu et al. (2018), Generating Wikipedia by Summarizing Long Sequences
    - From the abstract: “**generating English Wikipedia articles** can be approached as a multi-document summarization of source documents”
    - So while this paper implements task-specific models (baseline LSTM and Transformer-based ones), this result is remarkable, also to the Open AI researchers: “In manual inspection of model outputs, we noticed an unexpected side-effect: **models learn to translate** names from English into multiple languages [...] Although we did not do a systematic evaluation of the translations, we **found they are often correct, and often they are not found in the Wikipedia article itself**.” (p. 9 in the paper)

# Language Models are Few-Shot Learners: Unsupervised Multitask Learners

To conclude on the jist of the paper (for now):

- The capacity of the language model is essential to the performance of zero-shot task transfer
- will cover some common tasks



**Language Models are Unsupervised Multitask Learners**

*Figure 1.* Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

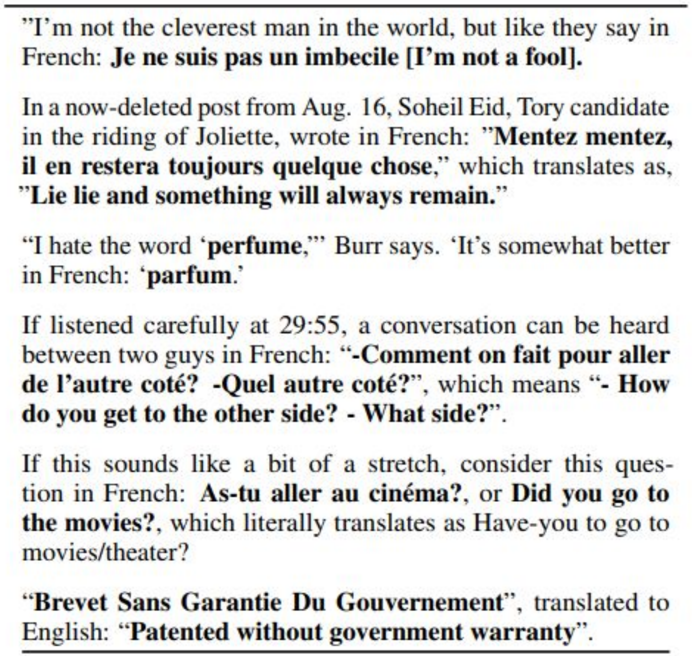
# Language Models are Few-Shot Learners:

# Unsupervised Multitask Learners

To conclude on the jist of the paper (for now):

WebText (text subset of 45 million web pages, 40 GB of text) major contribution from UML paper

- new web scrape to emphasize document quality, curated via Reddit: three upvote-minimum
- Byte Pair Encoding



”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

*Table 1.* Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

# Language Models are Few-Shot Learners: Unsupervised Multitask Learners

Note on Architecture:

- **Architecture** similar to GPT-1 approach
- relevant paper: [Radford et al., Improving Language Understanding by Generative Pre-Training (2018)]

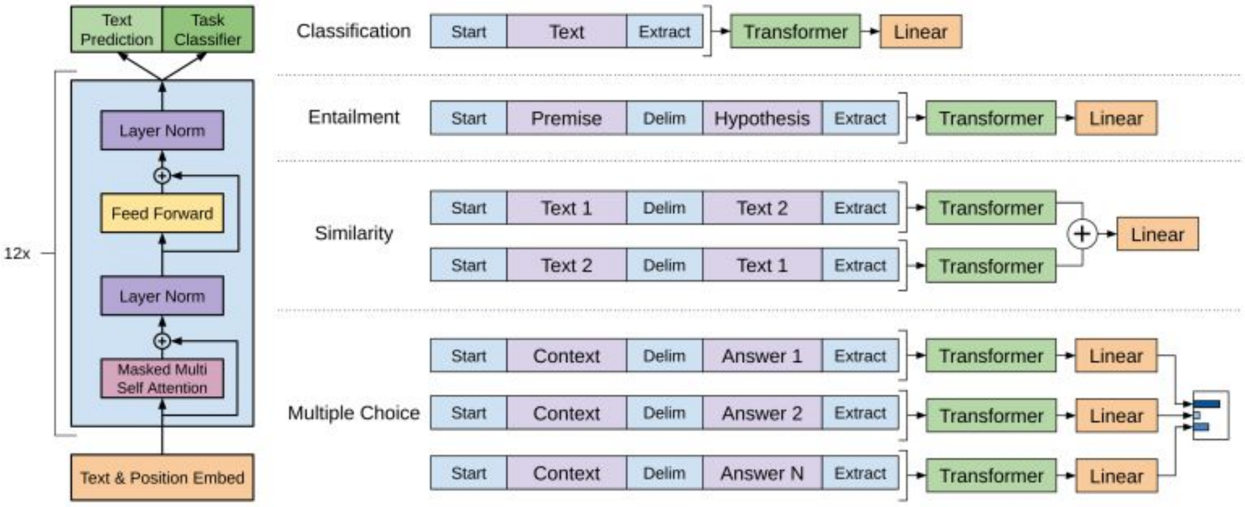


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

# Language Models are Few-Shot Learners:
# Historical Situation

Please refer to the paper for examples of these prior approaches, here a note of the transition and as presented in FSL paper:

- From **Task-Specific Architectures**
- to RNNs with **multiple layers of representations and contextual state**
- and more recently **pre-trained recurrent and transformer language models with direct fine-tuning**, no need for task-specific architectures
    - "However, a major limitation to this approach is that while the architecture is task-agnostic, there is still a need for task-specific datasets and task-specific fine-tuning: to achieve strong performance on a desired task typically **requires fine-tuning on a dataset of thousands to hundreds of thousands of examples specific to that task.**" (p. 3, from here on in the 2005.14165.pdf (arxiv.org) version (4) of the paper)

# Language Models are Few-Shot Learners: Historical Situation (GPT Focus)

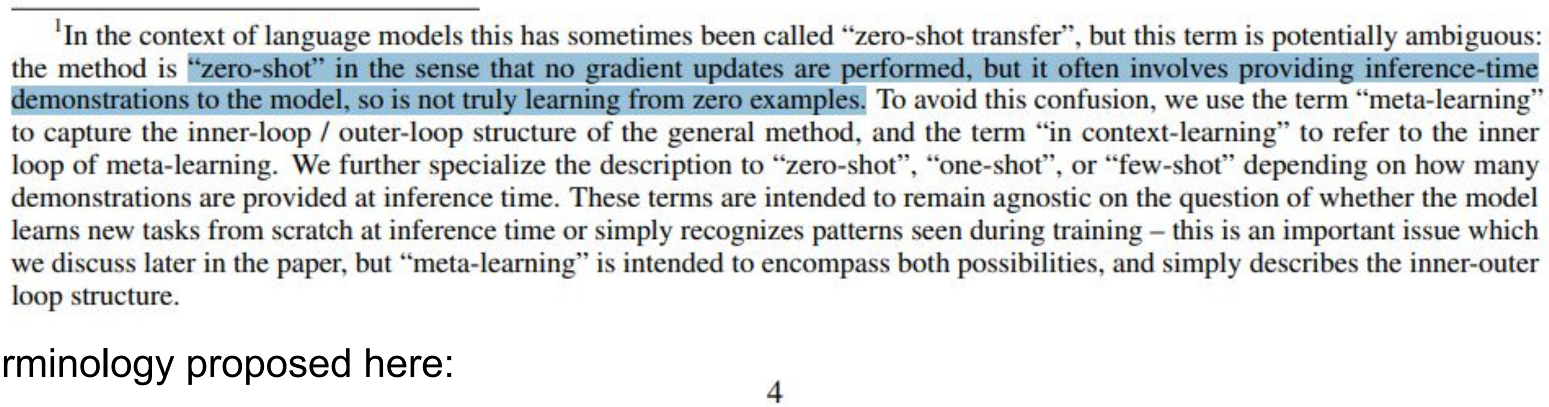
A note on GPT evolution and their introduction (relevant papers, already mentioned), without GPT-3.5/GPT-4 (current models) before we continue on the Need for Task Agnosticism

- **GPT (110 million parameters)**
  Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
  https://cdn.openai.com/research-covers/language-unsupervised/ language_understanding_paper.pdf
- **GPT-2 (1.5 billion parameters)**
  Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
  https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- **GPT-3 (175 billion parameters)**
  Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

# Language Models are Few-Shot Learners:
# Why Need for Task-Agnosticism?

- removes the **requirement of large, labeled dataset.** Certain types of task (correcting grammar, generating examples of an abstract concept, summaries!) might not have easily definable labels or it is difficult to collect them
- (when you do fine-tuning:) exploitation of **spurious correlations** in the training data grows with expressiveness of the model: Out-of-Distribution (OOD) Robustness, larger models do not necessarily generalize better out-of distribution but Transformer performance declines are better than other models' - Hendrycks, Dan, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. "Pretrained Transformers Improve Out-of-Distribution Robustness." arXiv, April 16, 2020. https://doi.org/10.48550/arXiv.2004.06100.
- **AGI perspective**: "humans do not require large supervised datasets to learn most language tasks" and,  "to be broadly useful, we would someday like our NLP systems to have this same fluidity and generality" (p. 4)

# Language Models are Few-Shot Learners: “” Metalearning/Zero-shot transfer (1)



---

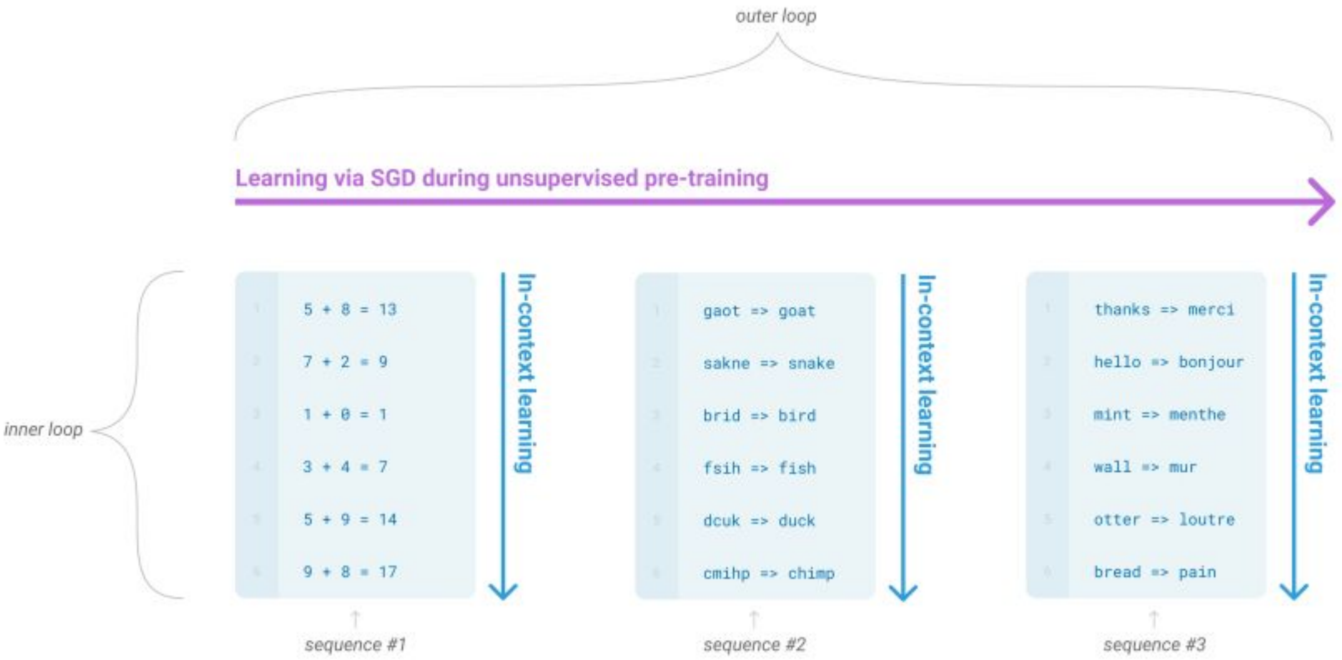
[1] In the context of language models this has sometimes been called “zero-shot transfer”, but this term is potentially ambiguous: the method is “zero-shot” in the sense that no gradient updates are performed, but it often involves providing inference-time demonstrations to the model, so is not truly learning from zero examples. To avoid this confusion, we use the term “meta-learning” to capture the inner-loop / outer-loop structure of the general method, and the term “in context-learning” to refer to the inner loop of meta-learning. We further specialize the description to “zero-shot”, “one-shot”, or “few-shot” depending on how many demonstrations are provided at inference time. These terms are intended to remain agnostic on the question of whether the model learns new tasks from scratch at inference time or simply recognizes patterns seen during training – this is an important issue which we discuss later in the paper, but “meta-learning” is intended to encompass both possibilities, and simply describes the inner-outer loop structure.

4

Terminology proposed here:

- **“Meta-learning”** for inner-outer loop structure and …
- … **“in context-learning”** for the inner loop: see the following slide re: loops
- **“zero-”/”one-”/”few-shot”** learning at inference time

# Language Models are Few-Shot Learners: Metalearning/Zero-shot transfer (2)

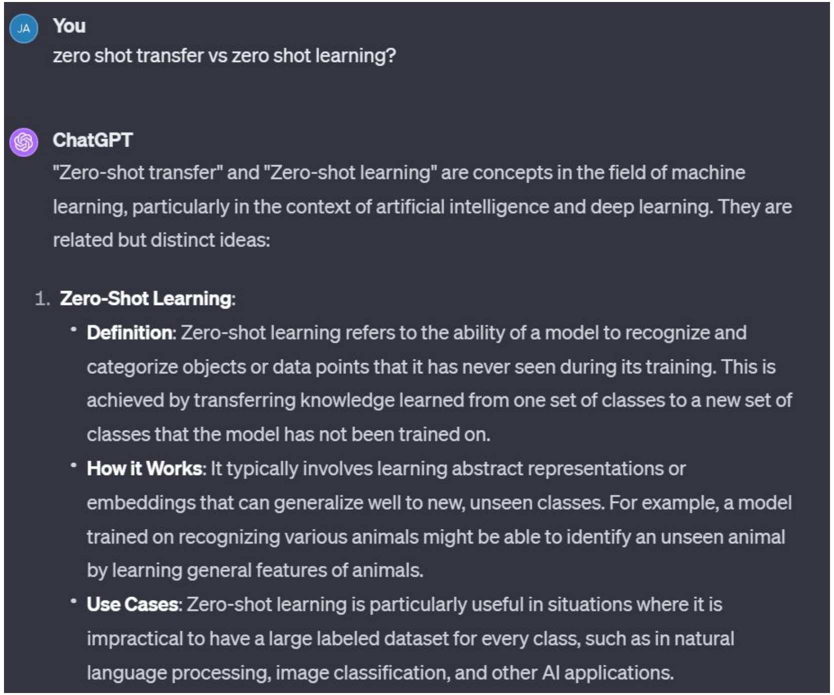


So again:

- metalearning overall loop structure, incl. Stochastic Gradient Descent (SGD)
- in-context pertains to the inner loops
- with “shots” per sequence provided

p. 3

# Language Models are Few-Shot Learners: Metalearning/Zero-shot transfer (3)

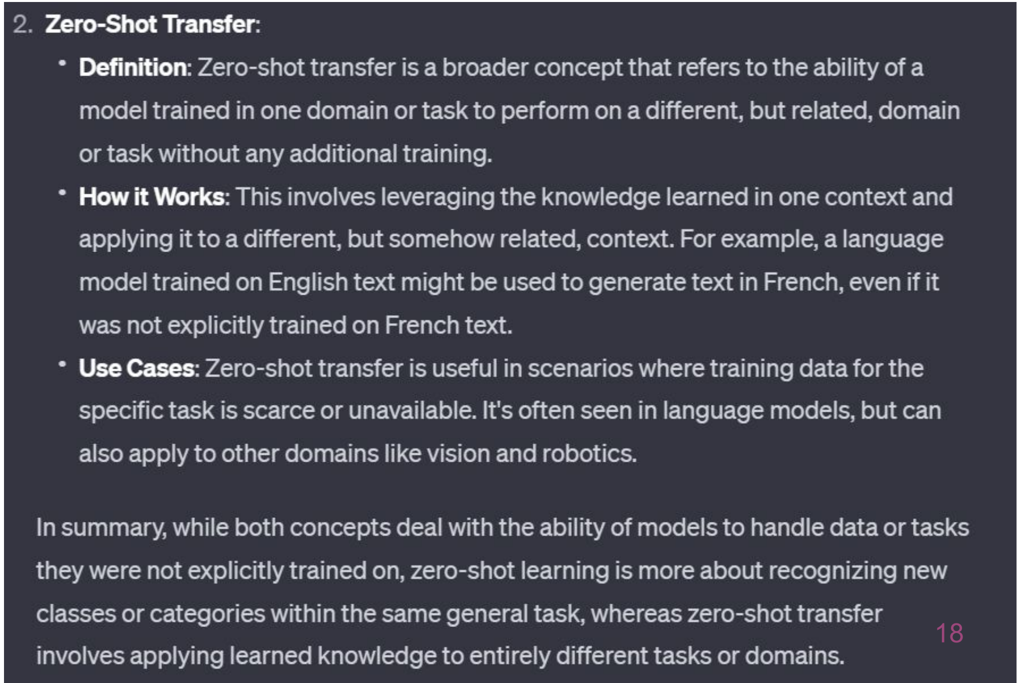


**You**
zero shot transfer vs zero shot learning?

**ChatGPT**
"Zero-shot transfer" and "Zero-shot learning" are concepts in the field of machine learning, particularly in the context of artificial intelligence and deep learning. They are related but distinct ideas:

1. **Zero-Shot Learning**:
    - **Definition**: Zero-shot learning refers to the ability of a model to recognize and categorize objects or data points that it has never seen during its training. This is achieved by transferring knowledge learned from one set of classes to a new set of classes that the model has not been trained on.
    - **How it Works**: It typically involves learning abstract representations or embeddings that can generalize well to new, unseen classes. For example, a model trained on recognizing various animals might be able to identify an unseen animal by learning general features of animals.
    - **Use Cases**: Zero-shot learning is particularly useful in situations where it is impractical to have a large labeled dataset for every class, such as in natural language processing, image classification, and other AI applications.

- Zero-shot transfer vs zero-shot learning?



2. **Zero-Shot Transfer**:
    - **Definition**: Zero-shot transfer is a broader concept that refers to the ability of a model trained in one domain or task to perform on a different, but related, domain or task without any additional training.
    - **How it Works**: This involves leveraging the knowledge learned in one context and applying it to a different, but somehow related, context. For example, a language model trained on English text might be used to generate text in French, even if it was not explicitly trained on French text.
    - **Use Cases**: Zero-shot transfer is useful in scenarios where training data for the specific task is scarce or unavailable. It's often seen in language models, but can also apply to other domains like vision and robotics.

In summary, while both concepts deal with the ability of models to handle data or tasks they were not explicitly trained on, zero-shot learning is more about recognizing new classes or categories within the same general task, whereas zero-shot transfer involves applying learned knowledge to entirely different tasks or domains.

# Language Models are Few-Shot Learners:

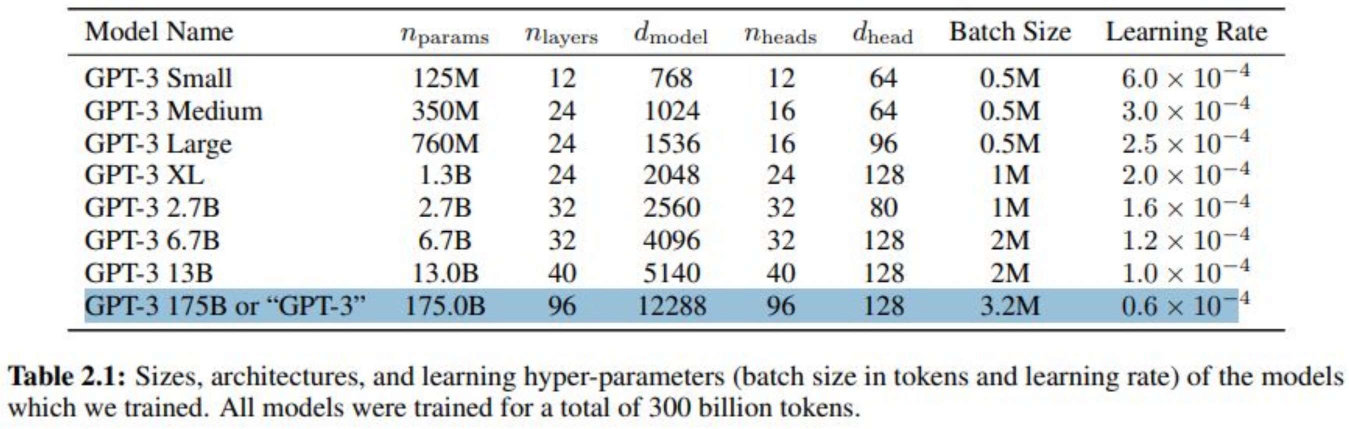
## GPT-3 Architecture

Since the model is introduced in this paper: We will end looking at datasets and tasks.

- Overall, similar to GPT-2, which was already similar to GPT-1
- 175 billion params instead of 1.5 billion (more layers etc.)
- Double context size (2048 instead of 1024 input tokens)
- Larger word embeddings (12.8k instead of 1.6k)
- Attention pattern from Sparse Transformer (sparse factorizations of attention matrix that grow with O(n sqrt(n)) instead of O(n^2))
  Child et al. (2019), Generating Long Sequences with Sparse Transformers

# Language Models are Few-Shot Learners:
# GPT-3 Architecture

Here a look at the GPT-3 versions in some more detail:



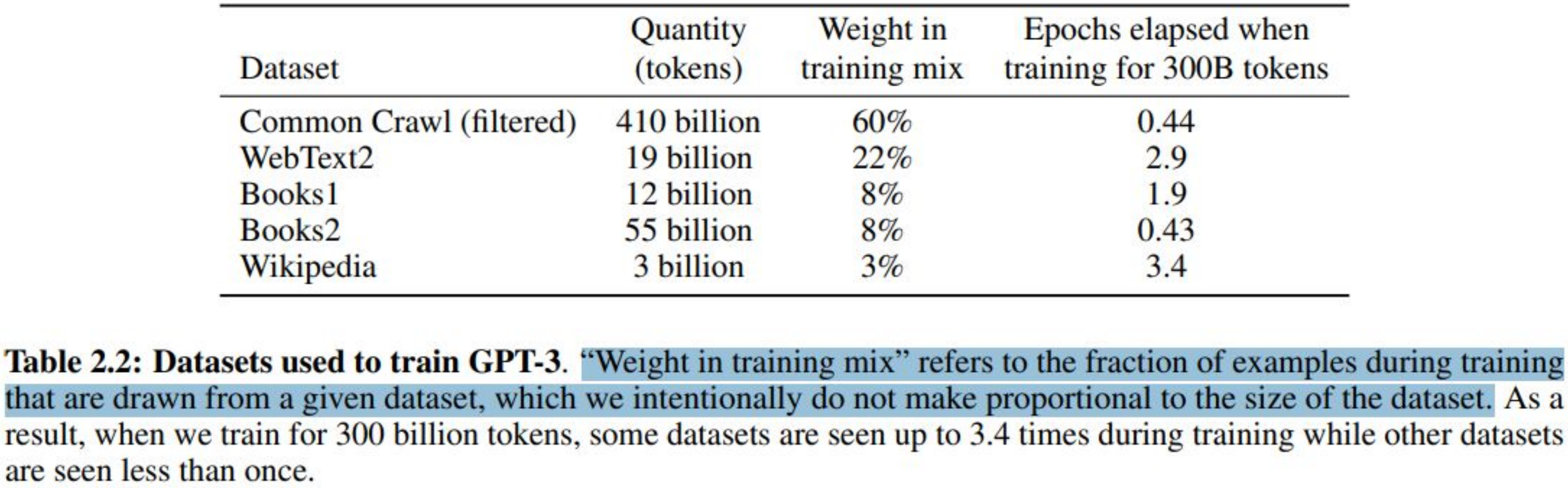
| Model Name | $n_{\text{params}}$ | $n_{\text{layers}}$ | $d_{\text{model}}$ | $n_{\text{heads}}$ | $d_{\text{head}}$ | Batch Size | Learning Rate |
|---|---|---|---|---|---|---|---|
| GPT-3 Small | 125M | 12 | 768 | 12 | 64 | 0.5M | $6.0 \times 10^{-4}$ |
| GPT-3 Medium | 350M | 24 | 1024 | 16 | 64 | 0.5M | $3.0 \times 10^{-4}$ |
| GPT-3 Large | 760M | 24 | 1536 | 16 | 96 | 0.5M | $2.5 \times 10^{-4}$ |
| GPT-3 XL | 1.3B | 24 | 2048 | 24 | 128 | 1M | $2.0 \times 10^{-4}$ |
| GPT-3 2.7B | 2.7B | 32 | 2560 | 32 | 80 | 1M | $1.6 \times 10^{-4}$ |
| GPT-3 6.7B | 6.7B | 32 | 4096 | 32 | 128 | 2M | $1.2 \times 10^{-4}$ |
| GPT-3 13B | 13.0B | 40 | 5140 | 40 | 128 | 2M | $1.0 \times 10^{-4}$ |
| GPT-3 175B or "GPT-3" | 175.0B | 96 | 12288 | 96 | 128 | 3.2M | $0.6 \times 10^{-4}$ |

**Table 2.1:** Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

- n_params: trainable params
- n_layers: total no of layers
- d_model: no of units in each bottleneck layer
- d_feedforward (not in the table): always 4 * model
- d_head: dimension of each attention head
- n_ctx (not in table): always 2048 tokens of context

# Language Models are Few-Shot Learners: Training Datasets

- We already talked about WebText (introduced with the GPT2 Paper, UML)
- That kind of high-quality dataset is then mixed with lower quality data, according to predetermined “weight” - in turn implying how often the data is seen in training epochs



| 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 |

**Table 2.2: Datasets used to train GPT-3.** “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

# Language Models are Few-Shot Learners: Training Results

- language modelling:
  as you go up with parameters, validation loss goes down
- as you scale up model size, data set size and compute size together, performance follows a power law
- (how far can this go? see the dashed line)

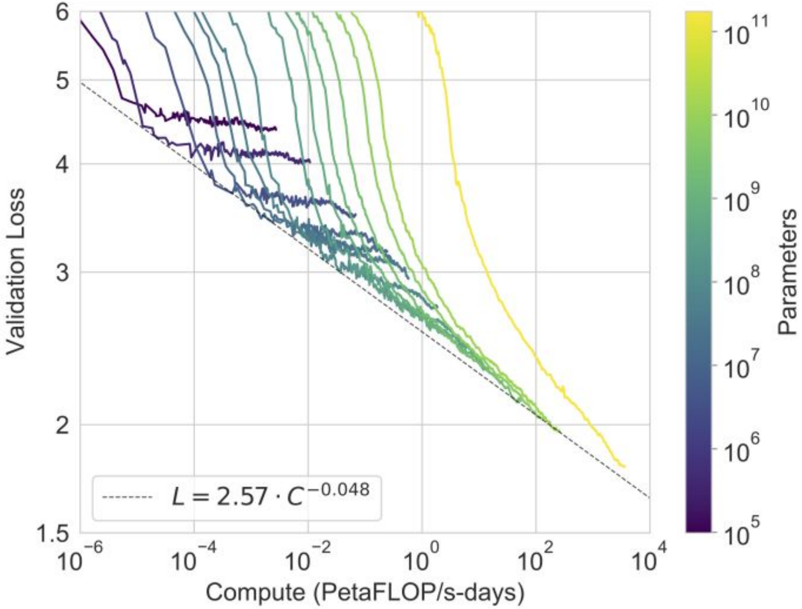


**Figure 3.1: Smooth scaling of performance with compute.** Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training. The power-law behavior observed in [KMH$^+$20] continues for an additional two orders of magnitude with only small deviations from the predicted curve. For this figure, we exclude embedding parameters from compute and parameter counts.

# Language Models are Few-Shot Learners: Showing Examples (vs Fine-Tuning)

**To recap**: BERT or similar style pretraining + finetuning gives way to “pure language modelling” approach. Is like direct use of pretrained model, so a pure language modelling way: Counting on the model having seen the structure of the task somewhere in the training data

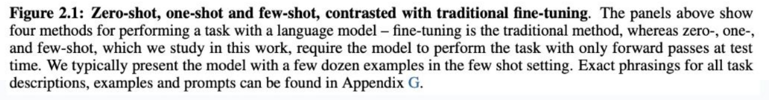
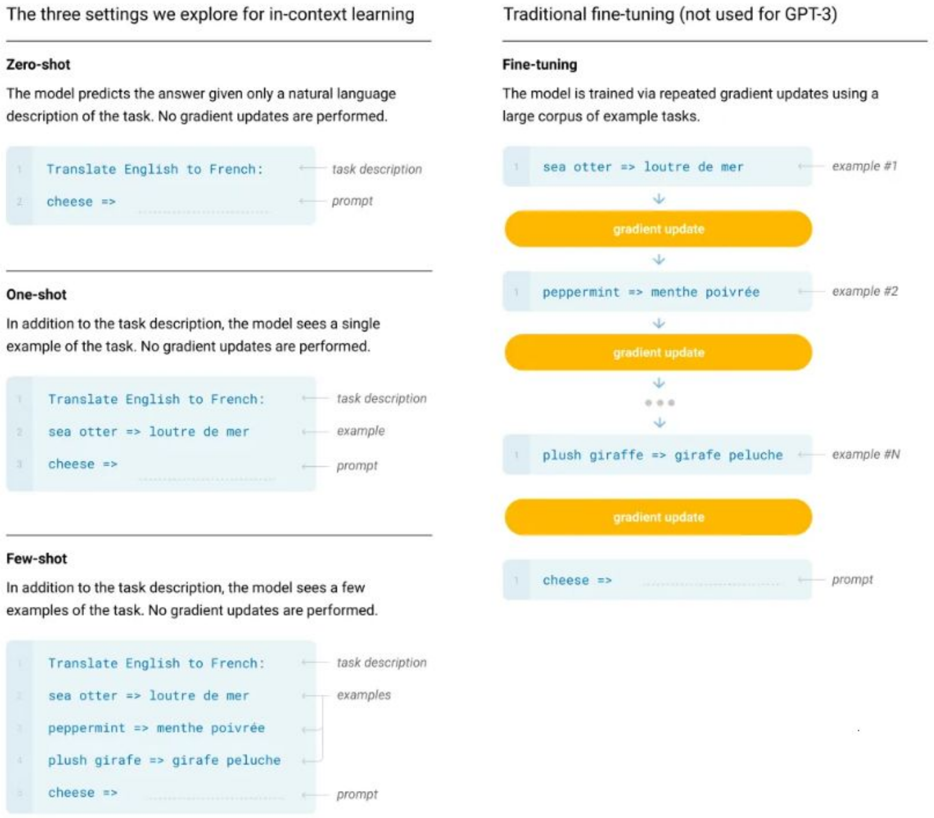


**Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning.** The panels above show four methods for performing a task with a language model – fine-tuning is the traditional method, whereas zero-, one-, and few-shot, which we study in this work, require the model to perform the task with only forward passes at test time. We typically present the model with a few dozen examples in the few shot setting. Exact phrasings for all task descriptions, examples and prompts can be found in Appendix G.



The three settings we explore for in-context learning

**Zero-shot**

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:     ← task description
cheese =>                        ← prompt
```

**One-shot**

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← example
cheese =>                        ← prompt
```

**Few-shot**

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

Traditional fine-tuning (not used for GPT-3)

**Fine-tuning**

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer          ← example #1
gradient update
peppermint => menthe poivrée        ← example #2
gradient update
• • •
plush giraffe => girafe peluche     ← example #N
gradient update
cheese =>                           ← prompt
```

JOHANNES KEPLER
UNIVERSITY LINZ

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

**How to evaluate?** In the fine-tuned/supervised regime this was more straightforward.

- draw $K$ random examples from the task's training set as conditioning, delimited by one or two newlines
- For **LAMBADA** and Storycloze: no supervised training set so conditioning examples are drawn from the development set and evaluated on the test set
    - we will look at LAMBADA in a moment
- for some tasks: natural language prompting (especially if $K = 0$)
- tasks with free-form completion: beam search with a beam of $\text{width} = 4$ and length penalty of $\text{alpha} = 0.6$

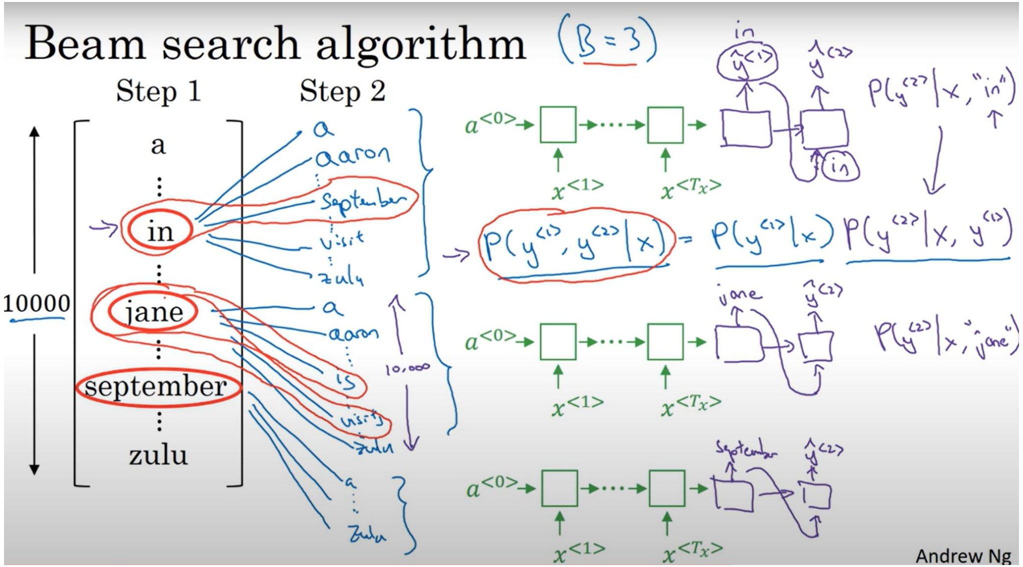
# Language Models are Few-Shot Learners: Showing Examples (Tasks)

**Beam Search*** for tasks with free-form completion

- Find top choices acc. to beam width
- compare with given completion?
- intuitive explanation (it’s not hard)

  [C5W3L03 Beam Search - YouTube](#)

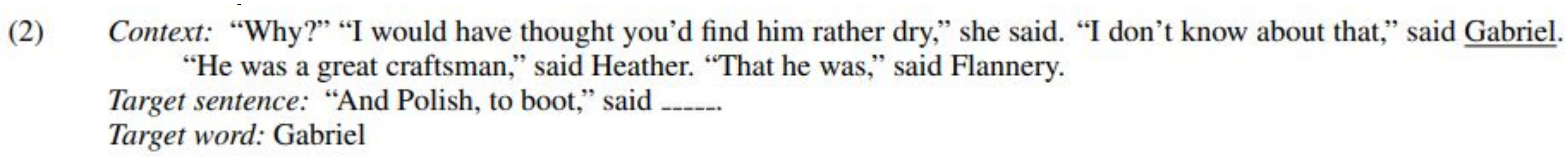
  (Andrew Ng)



# Language Models are Few-Shot Learners:
# Showing Examples (Tasks)

**LAMBADA** and Storycloze: already used for the Unsup. Multitask Learner paper as well

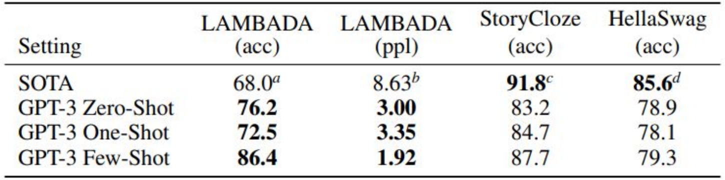
- LAMBADA (LAnguage Modeling Broadened to Account for Discourse Aspects)
- dataset introduced with Paperno et al. (2016), The LAMBADA dataset: Word prediction requiring a broad discourse context
- considered a difficult benchmark
- tests the modeling of long-range dependencies in text
- model is asked to predict the last world of sentences requiring a paragraph of context



(2) *Context:* "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
*Target sentence:* "And Polish, to boot," said ______.
*Target word:* Gabriel

# Language Models are Few-Shot Learners:
# Showing Examples (Tasks)

**LAMBADA** and Storycloze: GPT-3 beats -2 in the case of LAMBADA (perplexity score)



| Setting | LAMBADA (acc) | LAMBADA (ppl) | StoryCloze (acc) | HellaSwag (acc) |
|---|---|---|---|---|
| SOTA | 68.0[a] | 8.63[b] | **91.8**[c] | **85.6**[d] |
| GPT-3 Zero-Shot | **76.2** | **3.00** | 83.2 | 78.9 |
| GPT-3 One-Shot | **72.5** | **3.35** | 84.7 | 78.1 |
| GPT-3 Few-Shot | **86.4** | **1.92** | 87.7 | 79.3 |

**Table 3.2: Performance on cloze and completion tasks.** GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a][Tur20] [b][RWC+19] [c][LDL19] [d][LCH+20]

- [Tur20] Project Turing. Microsoft research blog, Feb 2020.
- [RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.
- [LDL19] Zhongyang Li, Xiao Ding, and Ting Liu. Story ending prediction by transferable bert.
- [LCH+20] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models.

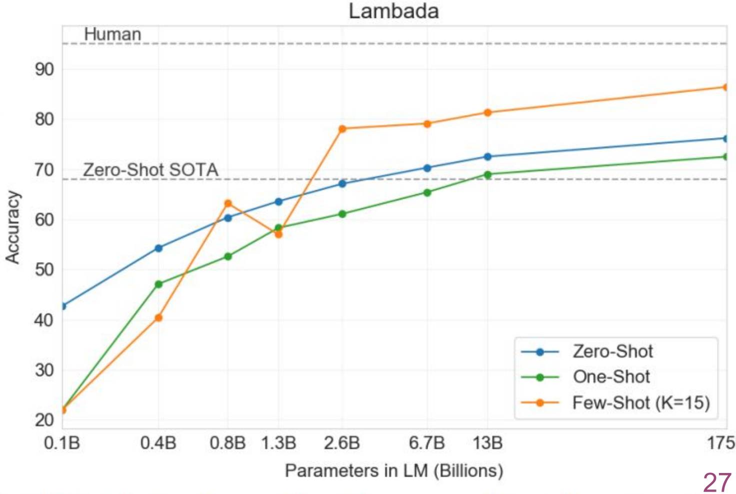


**Figure 3.2:** On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

There are *many* NLP tasks

- already mentioned decaNLP as an **aggregate task**/more in the paper (no time today)
  - The Natural Language Decathlon is a multitask challenge that spans ten tasks: **question answering** (SQuAD), machine **translation** (IWSLT), **summarization** (CNN/DM), natural language **inference** (MNLI), **sentiment analysis** (SST), **semantic role labeling**(QA-SRL), zero-shot **relation extraction** (QA-ZRE), **goal-oriented dialogue** (WOZ), **semantic parsing** (WikiSQL), and **commonsense reasoning** (MWSC). Each task is cast as question answering, which makes it possible to use our new Multitask Question Answering Network (MQAN). This model jointly learns all tasks in decaNLP without any task-specific modules or parameters in the multitask setting

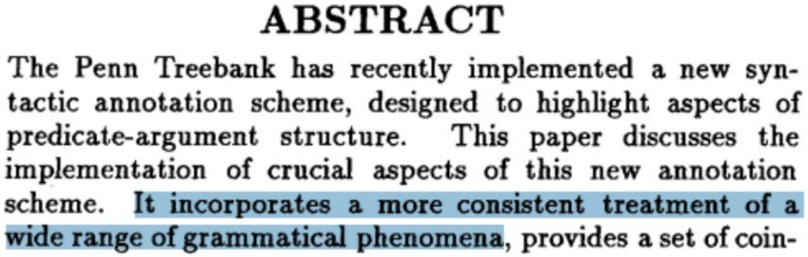
decaNLP

GitHub - salesforce/decaNLP: The Natural Language Decathlon: A Multitask Challenge for NLP

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

One more **interesting** dataset: **The Penn Treebank**/leads to test set **contamination question, these are the methodological issues the paper discusses**

- Marcus et al. (**1994**): [The Penn Treebank: Annotating Predicate Argument Structure]
- **predates the modern internet and so feasibly was not contained in the training data,** which made the dataset interesting for model development



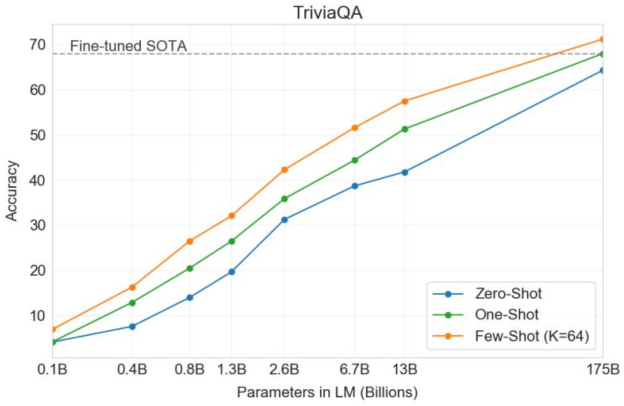
### ABSTRACT

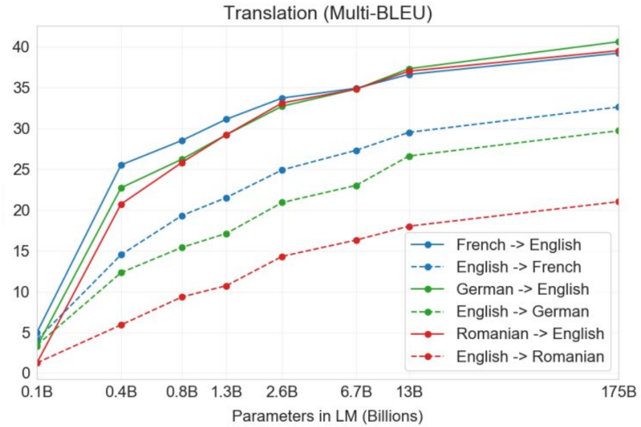
The Penn Treebank has recently implemented a new syntactic annotation scheme, designed to highlight aspects of predicate-argument structure. This paper discusses the implementation of crucial aspects of this new annotation scheme. It incorporates a more consistent treatment of a wide range of grammatical phenomena, provides a set of coin-

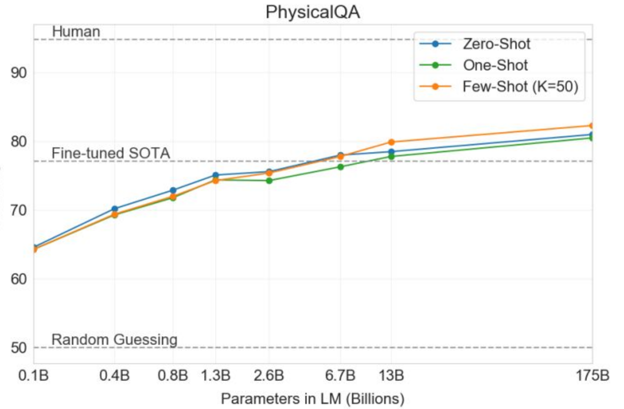
- **Memorization**: “Since our training dataset is sourced from the internet, it is **possible that our model was trained on some of our benchmark test sets.** Accurately detecting test contamination from internet-scale datasets is a new area of research without established best practices.” (FSL p. 29) - contamination is accepted to a degree (partial removal of detected overlaps only “due to a bug,” (p. 31) as it happens)

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

The core question in the Few-Shot Leaning Paper is how Few-Shot evaluates on NLP tasks, here some generally **successful** examples (without details on the datasets for time)







(better translation to English than from)

pp. 14-17 in FSL paper

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

And datasets/tasks where the models **struggle or could be better** (without dataset details)

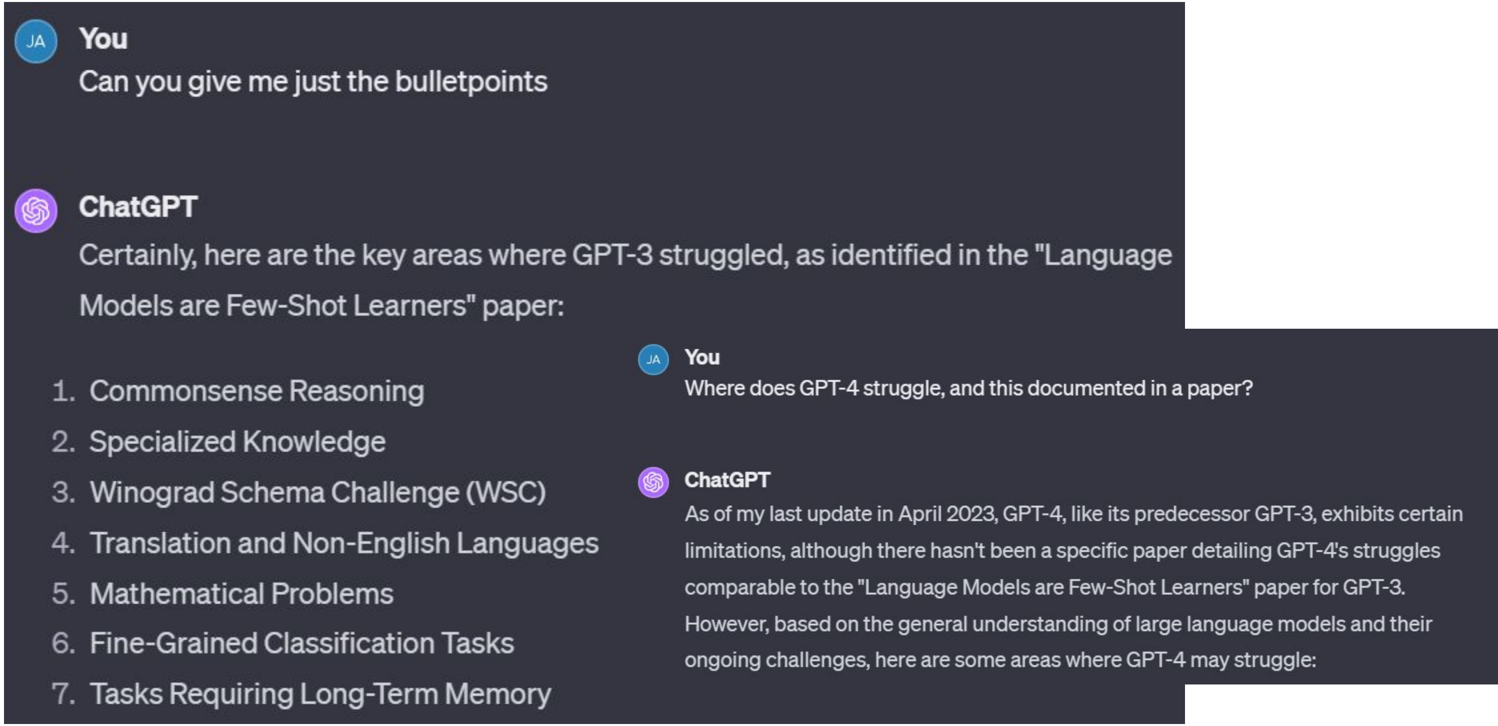
- To pick one: Commonsense Reasoning
  (Choice of Plausible Alternatives (COPA) + **HellaSwag**)

  HellaSwag involves picking the best ending to a story, intended to be difficult for LLMs

  - 95.6 % acc humans/78.1% acc GPT-3 (one-shot) and 79.3% few-shot/SOTA acc 85.6%
- Note on PIQA (Physical QA): “PIQA shows relatively shallow scaling with model size and is still over 10% worse than human performance, but GPT-3’s few-shot and even zero-shot result outperform the current state-of-the-art. Our analysis flagged PIQA for a potential data contamination issue (despite hidden test labels), and we therefore conservatively mark the result with an asterisk” (p. 18) as a note for how Open AI deals with reporting memorization

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

The full list,
of tasks where
GPT-3 struggles,
as per GPT-4
(for reference)



**You**

Can you give me just the bulletpoints

**ChatGPT**

Certainly, here are the key areas where GPT-3 struggled, as identified in the "Language Models are Few-Shot Learners" paper:

1. Commonsense Reasoning
2. Specialized Knowledge
3. Winograd Schema Challenge (WSC)
4. Translation and Non-English Languages
5. Mathematical Problems
6. Fine-Grained Classification Tasks
7. Tasks Requiring Long-Term Memory

**You**

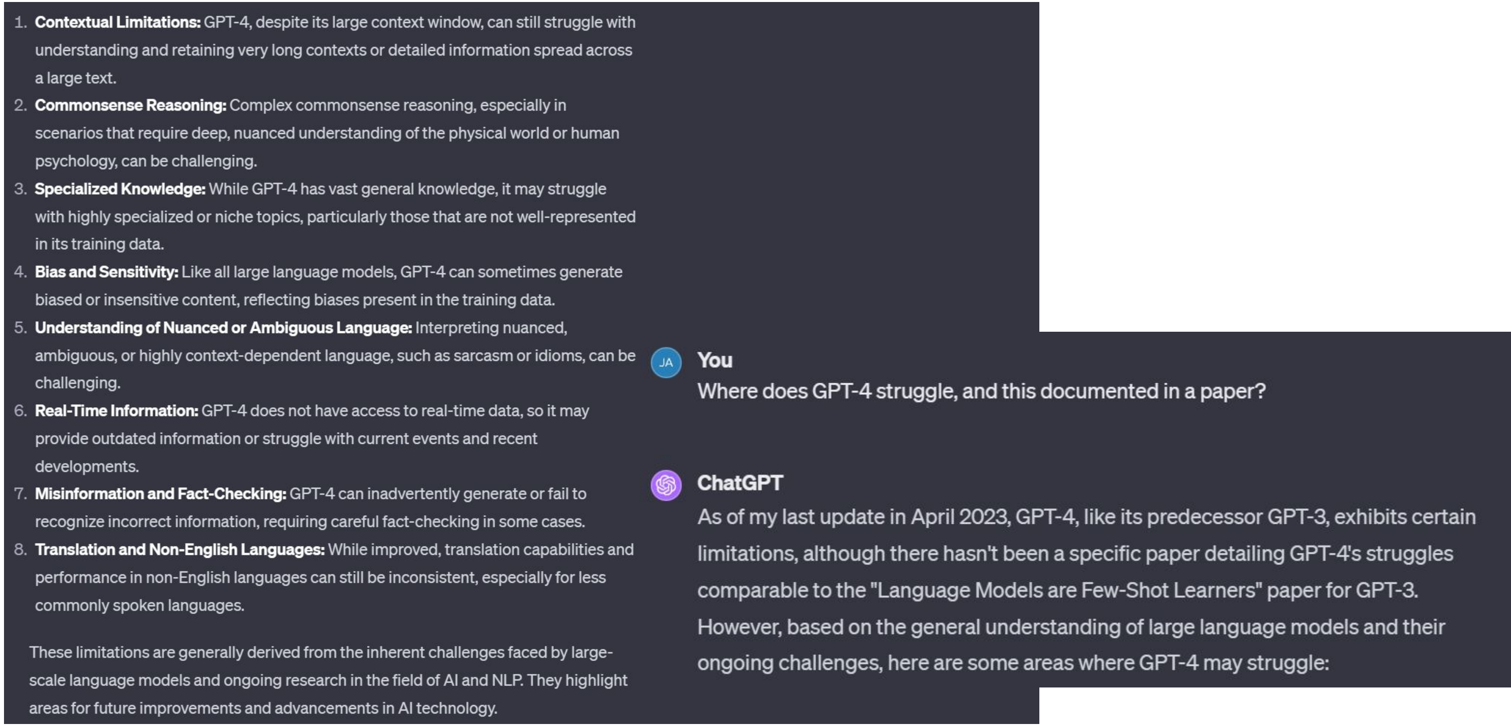
Where does GPT-4 struggle, and this documented in a paper?

**ChatGPT**

As of my last update in April 2023, GPT-4, like its predecessor GPT-3, exhibits certain limitations, although there hasn't been a specific paper detailing GPT-4's struggles comparable to the "Language Models are Few-Shot Learners" paper for GPT-3. However, based on the general understanding of large language models and their ongoing challenges, here are some areas where GPT-4 may struggle:

# Language Models are Few-Shot Learners: Showing Examples (Tasks)

GPT-4 BTW:



**You**

Where does GPT-4 struggle, and this documented in a paper?

**ChatGPT**

As of my last update in April 2023, GPT-4, like its predecessor GPT-3, exhibits certain limitations, although there hasn't been a specific paper detailing GPT-4's struggles comparable to the "Language Models are Few-Shot Learners" paper for GPT-3. However, based on the general understanding of large language models and their ongoing challenges, here are some areas where GPT-4 may struggle:

1. **Contextual Limitations:** GPT-4, despite its large context window, can still struggle with understanding and retaining very long contexts or detailed information spread across a large text.
2. **Commonsense Reasoning:** Complex commonsense reasoning, especially in scenarios that require deep, nuanced understanding of the physical world or human psychology, can be challenging.
3. **Specialized Knowledge:** While GPT-4 has vast general knowledge, it may struggle with highly specialized or niche topics, particularly those that are not well-represented in its training data.
4. **Bias and Sensitivity:** Like all large language models, GPT-4 can sometimes generate biased or insensitive content, reflecting biases present in the training data.
5. **Understanding of Nuanced or Ambiguous Language:** Interpreting nuanced, ambiguous, or highly context-dependent language, such as sarcasm or idioms, can be challenging.
6. **Real-Time Information:** GPT-4 does not have access to real-time data, so it may provide outdated information or struggle with current events and recent developments.
7. **Misinformation and Fact-Checking:** GPT-4 can inadvertently generate or fail to recognize incorrect information, requiring careful fact-checking in some cases.
8. **Translation and Non-English Languages:** While improved, translation capabilities and performance in non-English languages can still be inconsistent, especially for less commonly spoken languages.

These limitations are generally derived from the inherent challenges faced by large-scale language models and ongoing research in the field of AI and NLP. They highlight areas for future improvements and advancements in AI technology.

# Language Models are Few-Shot Learners: Limitations

Other limitations addressed with references to later releases of samples, qualitatively:

- **text synthesis**: “although the overall quality is high, GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs” (There was a widely commented upon news article generation task however, leading to almost 50/50 human detection acc (GPT-3))
- “**common sense physics**” (already discussed) - questions of the type “If I put cheese into the fridge, will it melt?”
- very specific things like **determining if two words are the same**, if one sentence implies another, etc. (p. 33)

algorithmically/structurally:

- **autoregressive** model, “our design decision comes at the cost of potentially worse performance on tasks which empirically benefit from bidirectionality” (p. 33)
    - “This may include fill-in-the-blank tasks, tasks that involve looking back and comparing two pieces of content, or tasks that require re-reading or carefully considering a long passage and then generating a very short answer.” (p. 33)
    - => making a bidirectional model at the scale of GPT-3 is a promising future direction for research

# Language Models are Few-Shot Learners: Limitations

algorithmically/structurally:

- LLMs not grounded in other domains of experience, such as video or real-world physical interaction: “thus lack a large amount of context about the world [...] For all these reasons, scaling pure self-supervised prediction is likely to hit limits, and augmentation with a different approach is likely to be necessary. Promising future directions in this vein might include learning the **objective function from humans [...], fine-tuning with reinforcement learning, or adding additional modalities such as images** to provide grounding and a better model of the world [...]” - highlighted parts all came with later models (p. 34)
- crucially for this presentation: “**A limitation, or at least uncertainty, associated with few-shot learning in GPT-3 is ambiguity about whether few-shot learning actually learns new tasks “from scratch” at inference time, or if it simply recognizes and identifies tasks that it has learned during training.** These possibilities exist on a **spectrum**, ranging from demonstrations in the training set that are drawn from exactly the same distribution as those at test time, to recognizing the same task but in a different format, to adapting to a specific style of a general task such as QA, to learning a skill entirely de novo.” (p. 34)

other than this: size (distilled versions feasible?), interpretability, biases from trained data (and stereotyped/prejudiced content) - these last issues leading on to a whole Broader Impacts part in the paper

# Summary

- Contrasted the Brown et al. GPT-3 Open AI Paper
  **Language Models are Few-Shot Learners**
  - with: **Language Models are Unsupervised Multitask Learners**
    (Radford et al. GPT-2 Open AI GPT-2 Paper)
  - => Persistent themes and modes of evaluating Language Models?
- **Themes:** evolution of the models: data quality (GPT-2) and model expressiveness (GPT-3)
- **… implications** for how to use these models, transition from fine-tuning to in-context, with **obvious advantages** of in-context if performance suffices, which this remains a **(task-specific!) open question - and we talked about limitations**
- **Evaluating Models** next slide

# Summary

*Interesting to me in terms of the Barrier-free PDF Documents project:*

- **Evaluating Models** (in the in-context regime)
    - skipped the slide today: but topics like beam search in evaluating free form answers against some given acceptable label (by human worker, e.g. someone annotating PDFs for accessibility)

Can stop here: final slides have discussion + open questions/possible directions

# Summary

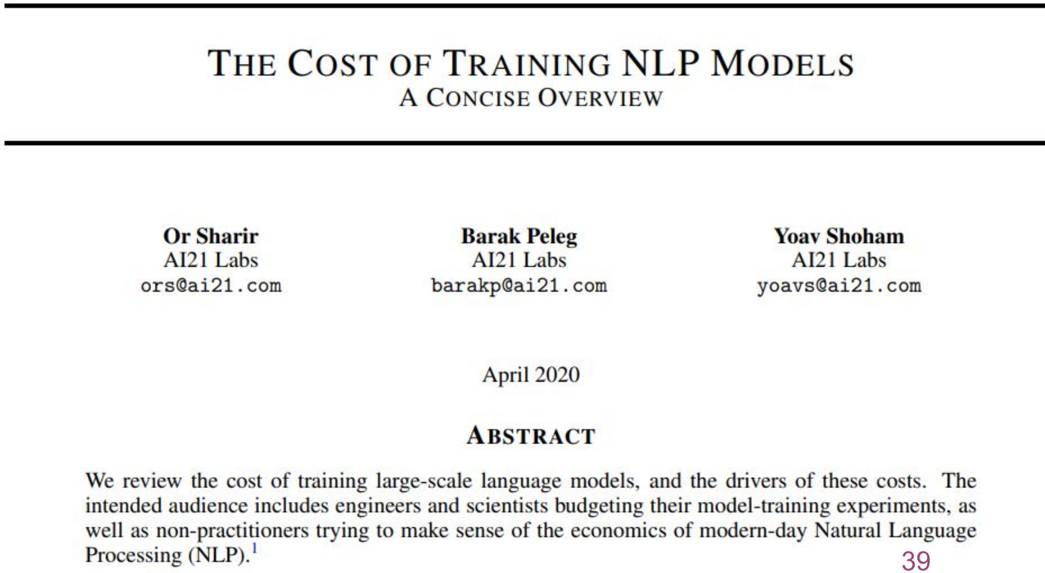
Discussion Questions

- **Shift** going from the Unsup. Multitask Learners Paper to the Few Shot Learners Papers?
  - GPT-3 and later models show significant **improvements** in natural language understanding and generation, enabling more coherent and contextually relevant responses. They are better at "few-shot" or "zero-shot" learning, meaning they can perform tasks with little to no specific examples provided, a significant leap from GPT-2.
  - **Extension** of the Unsup. Multitask idea, that by training on a wide range of text sources, the model learns a variety of language tasks without needing task-specific training data.
- Is the **model/architecture new**? Appears standard: the difference is size and data quality

# Summary

We looked at one aspect, the core idea of zero shot transfer, and didn’t cover other important aspects, just one economic example being …

- Sharir et al., The Cost of Training NLP Models: A Concise Overview (2020)
  - $2.5k - $50k (110 m params)
  - $10k - $200k (340 m params)
  - $80k - $1.6m (1.5 b params)
- … to take this in a different direction, particularly looking at the different strategies under a **(dollar, environmental) cost** aspect



THE COST OF TRAINING NLP MODELS
A CONCISE OVERVIEW

**Or Sharir**
AI21 Labs
ors@ai21.com

**Barak Peleg**
AI21 Labs
barakp@ai21.com

**Yoav Shoham**
AI21 Labs
yoavs@ai21.com

April 2020

ABSTRACT

We review the cost of training large-scale language models, and the drivers of these costs. The intended audience includes engineers and scientists budgeting their model-training experiments, as well as non-practitioners trying to make sense of the economics of modern-day Natural Language Processing (NLP).[1]

# Summary

We looked at one aspect, the core idea of zero shot transfer, and didn't cover other important aspects, just one more, performance-oriented example being …

- Ouyang et al. (2022): Training language models to follow instructions with human feedback
  - **fine-tuning the overall behavior** of the GPT model using RLHF (Reinforcement Learning from Human Reinforcement)

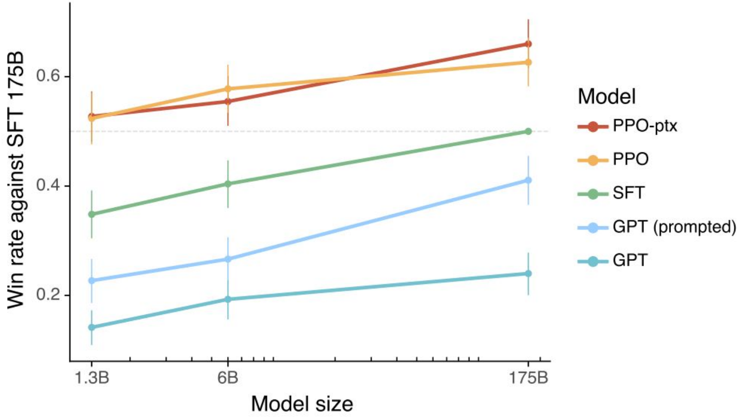


Figure 1: Human evaluations of various models on our API prompt distribution, evaluated by how often outputs from each model were preferred to those from the 175B SFT model. Our InstructGPT models (PPO-ptx) as well as its variant trained without pretraining mix (PPO) significantly outperform the GPT-3 baselines (GPT, GPT prompted); outputs from our 1.3B PPO-ptx model are preferred to those from the 175B GPT-3. Error bars throughout the paper are 95% confidence intervals.

# Sources

- Papers linked throughout
- Helpful (paper explanation type) videos out there:
  - L19.5.2.5 GPT-v3: Language Models are Few-Shot Learners - YouTube
  - GPT-3: Language Models are Few-Shot Learners (Paper Explained) - YouTube
  - (GPT-2) Language Models are Unsupervised Multitask Learners | Paper Explained - YouTube
  - Beam Search in NLP: C5W3L03 Beam Search - YouTube