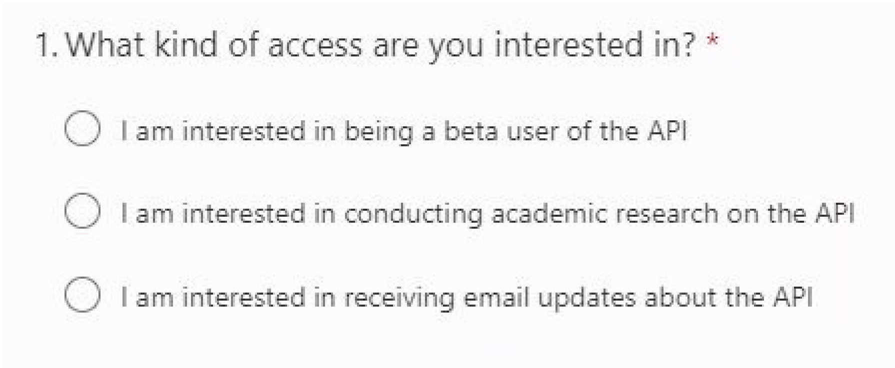
1. **OpenAI GPT-3 API开放给研究人员申请**

在OpenAI 官方有人提出疑问，**为什么OpenAI选择公开API而不是直接开源他们的模型？**他们也很诚实的给出了最重要的原因，1，技术的商业化有助于正在进行的AI研究，安全和政策工作。2，许多模型非常大，需要大量的专业知识来进行开发和部署，且运行起来非常昂贵。这就使得除了大公司以外的任何人都很难从基础技术中受益, 比较像近阶段美国国会就三大巨头质询时提到的技术垄断，OpenAI希望该API使小型企业和组织更容易使用到他们研发的强大的AI系统。3，API模型可以更轻松地应对技术滥用。由于很难预测模型的下游用例，因此通过API释放它们并随着时间的推移扩展访问范围，这样做相对来说更安全，而不是仅仅发布一个开源模型，因为如果发现它具有有害的应用程序，则该模型无法进行访问调整。

其实相反也可以看出他们也在探索更多的商业模式，并且想主宰技术的所有权，控制模型的API访问权限在一方面控制了下游的具体商业有害APP，防止技术滥用，但是7月以来对申请API的限制也可以看出OpenAI对其技术的控制力度。

下面是申请API权限步骤和一些心得：

首先第一个问题：


1. What kind of access are you interested in? *

- I am interested in being a beta user of the API
- I am interested in conducting academic research on the API
- I am interested in receiving email updates about the API

Beta user 和 conducting academic research 的区别是，研究机构的不同，第一个更多的以企业商用角度出发，除了填写个人信息还需要填写企业信息，需要注明商业落地具体user case 这点有点类似Megengine 天元和百度飞桨。

Your use case. Provide a brief description of the way that you or your organization might want to use the API (or if you're not sure, describe your current product or service). Where applicable, please include any initial thoughts you have around potential benefits or risks from your use case or any other information that you'd like us to know.

OpenAI 也在积极探索更多的商业模式，就像其在官网说的，他们在招募更多的研究人员探索GPT-3的更多的可能性。如果作为一个开发人员，所需的信息就更多了，在这除了个人信息之外，还需要填写研究机构，以及6大研究领域任选一个Fairness and Representation，Robustness，Model Exploration，Interdisciplinary Research，Misuses Potential，other。由于个人对模型可解释性更加感兴趣，故选择这个，同时需要回答他们的以下问题：*Model Exploration: Models like those served by the API have a variety of capabilities which we are yet to explore. We're excited by investigations in many areas including linguistic properties, commonsense or grounded reasoning, and potential uses for many other NLP problems (especially those involving generation).*

- Within the research area selected above, please describe the type of work you plan on doing with the API and your approach.
- What is the anticipated timeline for completing this research (i.e. 3 months, 6 months)?
- Would you be the sole investigator on this project, or would you require technical access for multiple individuals? If so, please specify how many.
- What past research by you would you like us to read as we consider your submission (if applicable)?

针对上述问题，我结合Cellnet 的可解释诸如不适用Pointwise conv,利用线性变换同样可以在保证相对精度的范围下提取图像特征，想试看这一原则是否适用于基于语言模型的model，这样是否对GPT-3的参数过于冗余有一定的帮助。Timeline我选择的是基于业余兴趣6个月或者更长，在项目的负责上选择需要协助以及订阅了GPT-3相关news。

2. 众多Developer 关于OpenAI GPT-3的研究和探讨

纵观现有的国外的技术博客或者已经出的关于GPT-3 的youtube 教程, 热点时间是GPT-3 刚出**6月份**，热点演示领域是**fake news**所带来的震惊。关键词GPT-3检索到的视频顶流阅读量在英文世界达**20万**次，最高阅读量来自Barış Özcan
Barış Özcan 土耳其出生，美国科技博主，**133万次播放量**，关于GPT-3的科普。

Source from：MIT Tech review 有对近期GPT-3 有意思的研究成果进行过一次综述，其中有引用到 Arram Sabeti 三藩市的AI创业者， 他个人在GPT-3领域比较活跃，撰写的博客阅读数量也非常惊人。 故 MIT Tech review 在写这篇文章时有引用了很多来自Arram的开发案例和demo。
Arram比较多的是在使用GPT-3在诗歌和新闻采访方面进行训练和演示：诸如：
Recent Posts

- *AI Tim Ferriss Interviews AI Marcus Aurelius (GPT-3)*
- *AI Fan Fiction or: Barry by Terry Pratchett (GPT-3)*
- *GPT-3 Predicts the Rest of 2020*
- *GPT-3: Using Fiction to Demonstrate How Prompts Impact Output Quality*
- *Are Humans Intelligent? A Salty AI Op-Ed*
- *Teaching GPT-3 to Identify Nonsense*
- ***GPT-3 Predicts the Rest of 2020***
- ***Elon Musk By Dr. Seuss (GPT-3)***

3. **OpenAI GPT3 官方和研究者合作具体应用**

OpenAI和众多合作者、Startup在6月Release以来开发出了一系列的具体场景应用，并分别有各自的APP名称，在语义搜索，聊天，客户服务，文字生成，内容理解，**Polyglot（多语言转换）**以及启发工具七大领域展示了包括 **Casetext，Algolia，AI Channels，MessageBird，Sapling，AI Dungeon，AI Weirdness，Replika，Quizlet，Art of Problem Solving，Koko，Ross Intelligence，Translation 16款应用。（点击文字既可访问应用官网）。**

这些具体的应用程序有的是OpenAI自己在建设开发中的，有些是提供API接口或技术人员支持，拿Art of Problem Solving 公司举例，今年的疫情也加速了北美在线教育市场的迅猛发展，AoPS正在帮助有效地培养下一代STEM专业人员，今年也打破了10年来所有课程的入学记录。在过去10年中，AoPS训练的学生几乎是美国国际数学奥林匹克团队的所有成员以及成长成为OpenAI的研究人员/开发人员。AoPS根据专家老师的现有反馈对OpenAI的API进行训练，并使用该API快速生成有关学生作业的反馈的初稿，以供评分者改进和发送。教师的最终版本也与GPTAPI 共享以帮助进一步改进，而教师本人则负责评估该工具并确定其使用范围。

在内容理解模块，KOKO这款应用借助OpenAI，基于文本的分类器得到了大幅改进且无需进行预处理。分类器的F1分数从76%上升到86%，综合Acc.准确度上升到96%。在文字内容生成方面，GPT-3简直可以做到以假乱真的地步，请阅读下面一篇博客：https://maraoz.com/2020/07/18/openai-gpt3/）OpenAI's GPT-3 may be the biggest thing since bitcoin JUL 18, 2020

Summary: I share my early experiments with OpenAI's new language prediction model (GPT-3) beta. I explain why I think GPT-3 has disruptive potential comparable to that of blockchain technology.

这篇技术博客就是早期获得openAI API的资深算法工场师写的预言，他预言GPT-3将会是媲美比特币的技术，当读者读完整篇文章后，会在末尾发现这样一行字：

Now for the fun part, *I have a confession:* ***I did not write the above article****. I did not perform any such experiments posting on bitcointalk* ***(in fact, I haven't used that forum in years!)****. But I did it on my own blog!* ***This article was fully written by GPT-3. Were you able to recognize it?*** *I received access to OpenAI API yesterday.*

4. **GPT-3 论文亮点和引入**

GPT-3 的论文有67页左右，其中有很多晦涩难懂的逻辑推导，

*Yannic Kilcher 有专门利用**一个小时**的时间把所有的亮点进行一次讲解，:这是一个非常有帮助的视频，也是 paper workshop 的 onsite video 讲解。之后，在看完整篇文章之后 我会依照这个顺序罗列出具体技术细节。（下面文字链接至视频）*

0:00 - Intro & Overview
1:20 - Language Models
2:45 - Language Modeling Datasets
3:20 - Model Size
5:35 - Transformer Models

7:25 - Fine Tuning
10:15 - In-Context Learning
17:15 - Start of Experimental Results
19:10 - Question Answering
23:10 - What I think is happening
28:50 - Translation
31:30 - Winograd Schemes
33:00 - Commonsense Reasoning
37:00 - Reading Comprehension
37:30 - SuperGLUE
40:40 - NLI
41:40 - Arithmetic Expressions
48:30 - Word Unscrambling
50:30 - SAT Analogies
52:10 - News Article Generation
58:10 - Made-up Words
1:01:10 - Training Set Contamination
1:03:10 - Task Examples

---Qiang's report on 01.09.2020 regarding at GPT-3 research study

GPT-3 (Generative Pre-training Transformer)
Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches

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, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic

GPT-3 faces methodological issues related to training on large web corpora.

Finally, we find that GPT-3 can generate samples of news articles in which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

can be poor because the model is overly specific to the training distribution and does not generalize [YdC+19, MPL19]. Thus, the performance of fine-tuned models on specific benchmarks, even when it ~~human-level, may exaggerate actual performance~~ on the underlying task [GSL+18, NK19].

Third, humans do not require large supervised ~~datasets to learn most language tasks~~ – a brief direc language (e.g. "please tell me if this sentence describes something happy or something sad") or at most of demonstrations (e.g. "here are two examples of people acting brave; please give a third example of bra

语言模型

为啥复杂语言模型？

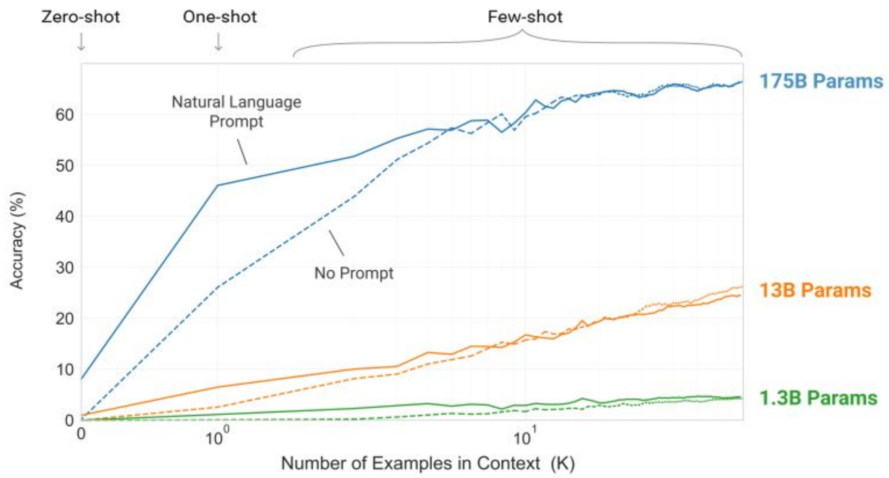


Figure 1.2: Larger models make increasingly efficient use of in-context information.

Open AI has been in the race for a long time now. In comparison, GPT-3 has a whopping 175 BN parameters, 10 times more than the next largest LM, the Turing NLG, developed by Microsoft with 17 BN parameters.

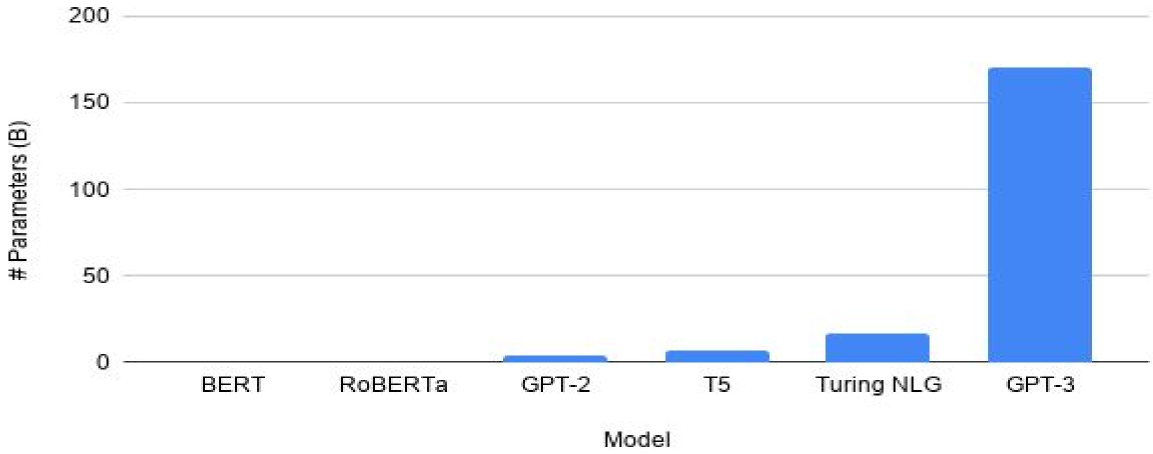


Figure 1.3: Comparison of all available language models (LMs) parameter wise
[Source: TowardsDataScience]

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 is 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.

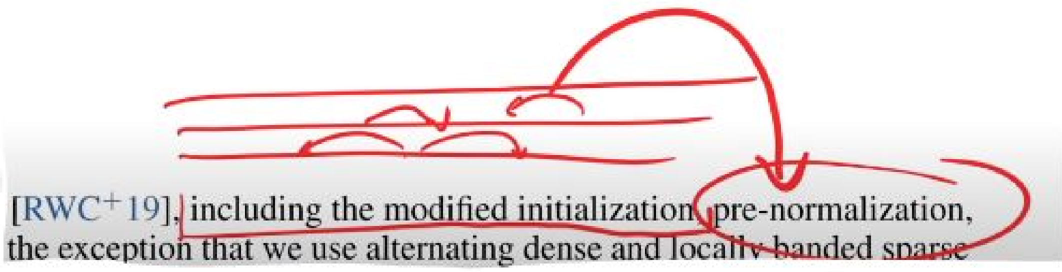
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-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. layer stands for attention layer and each head represents for attention head

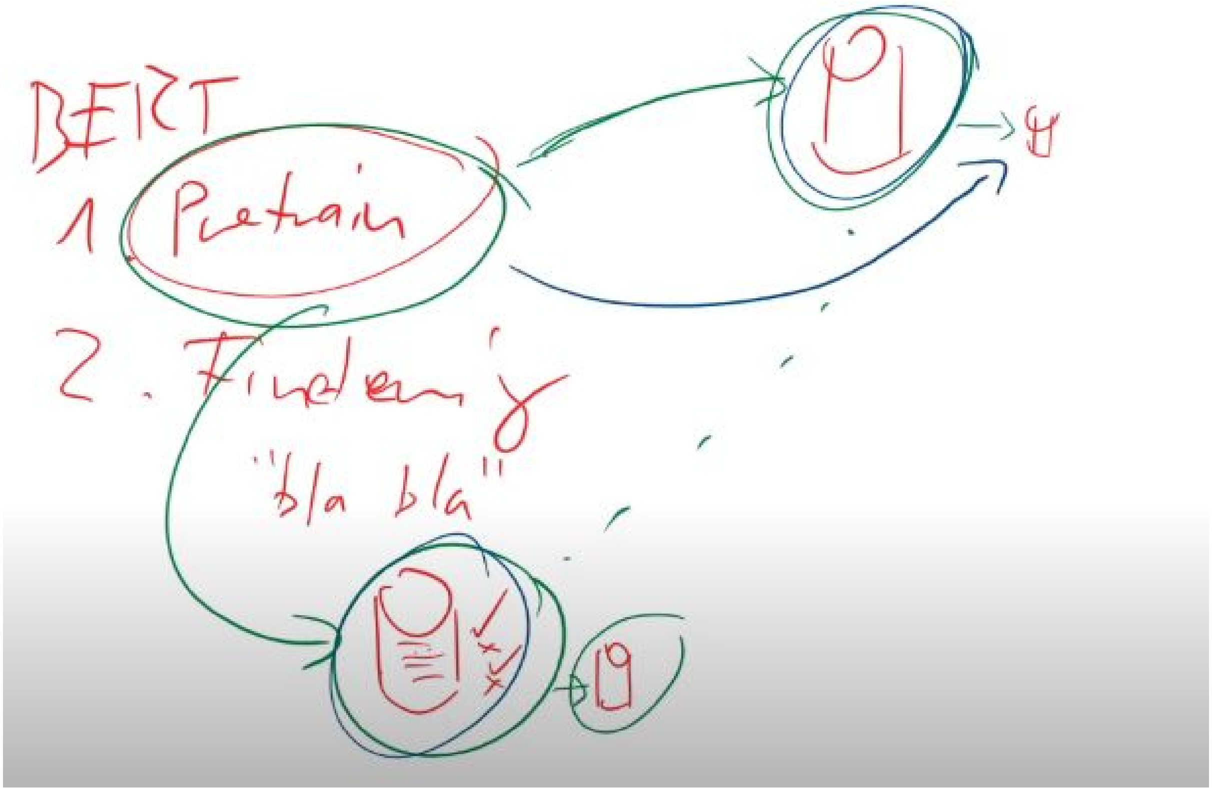
Before we deep dive, it may be useful to define some commonly used terminologies:

- NPL Tasks: These are tasks which have something to do with human languages, example — Language Translation, Text Classification (e.g. Sentiment extraction), Reading Comprehension, Named Entity Recognition (e.g. recognizing person, location, company names in text)
- Language Models: These are models which can predict the most likely next words (and their probabilities) given a set of words (think something like Google query auto-complete). Turns out these type of models are useful for a host of other tasks although they may be trained on mundane next word prediction
- Zero / One / Few shot learning: Refers to model's ability to learn a new task by seeing zero / one / few examples for that task
- Transfer Learning: Refers to the notion in Deep Learning where you train a model for one task (example object detection in images) , but the ability to leverage and build upon that for some other different task (example assessing MRI scans). After massive success in Computer Vision, its in vogue in NLP these days.
- Transformer Models: Deep learning family of models, used primarily in NLP, which forms the basic building block of most of the state-of-the-art NLP architectures these days. You can read more about Transformers at one of my earlier blog

how does the transformer model look like:


[RWC+19], including the modified initialization, pre-normalization,
the exception that we use alternating dense and locally banded sparse

it's not like Bert always predicts left and right, it is an order regressive model, and it always predicts from left to right.
We use the same model and architecture as GPT-2 [RWC+19], including the modified initialization, pre-normalization, and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer
just have more layers and wider layers and more data to train it on a language modeling training way.



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*

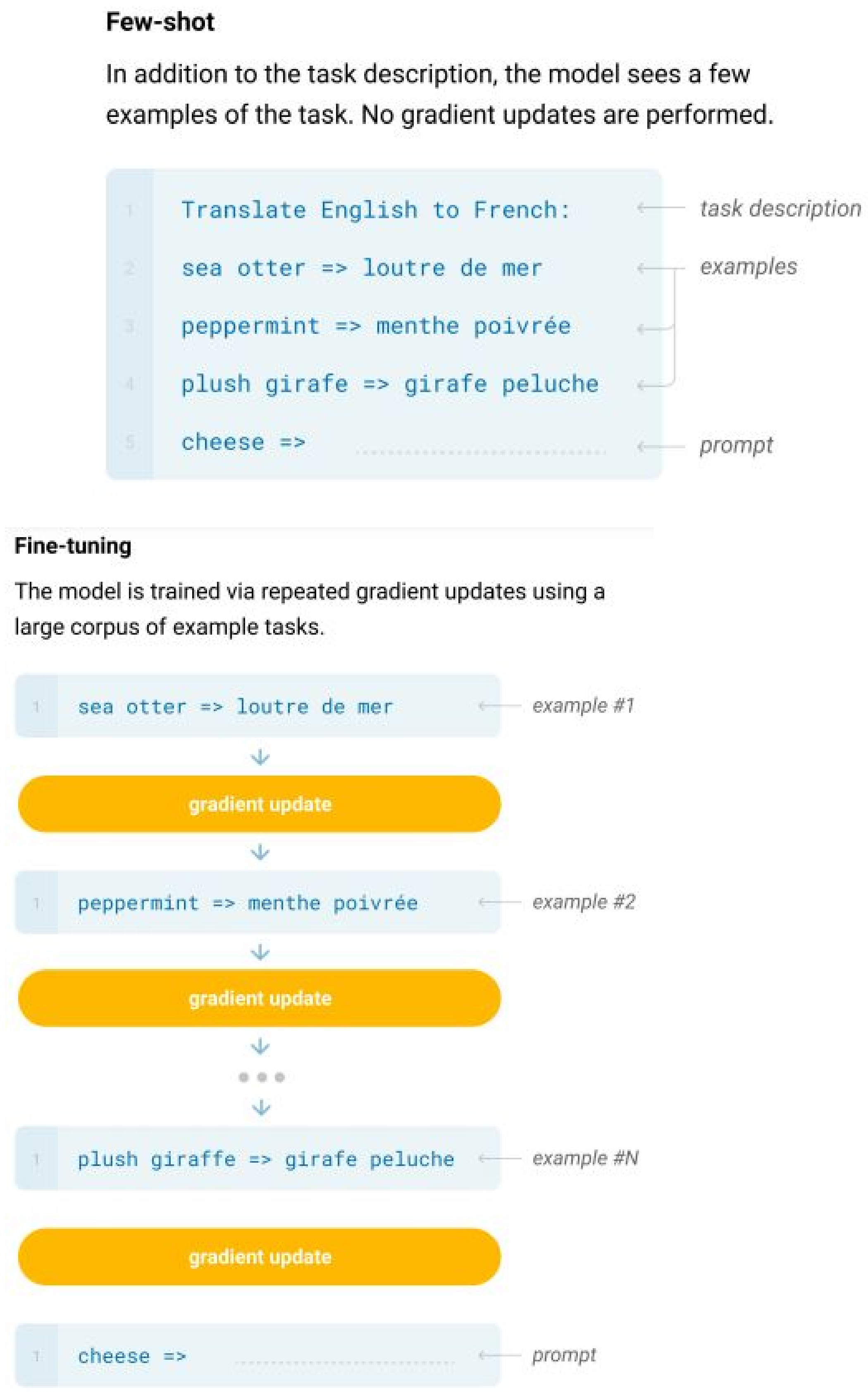


Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning.

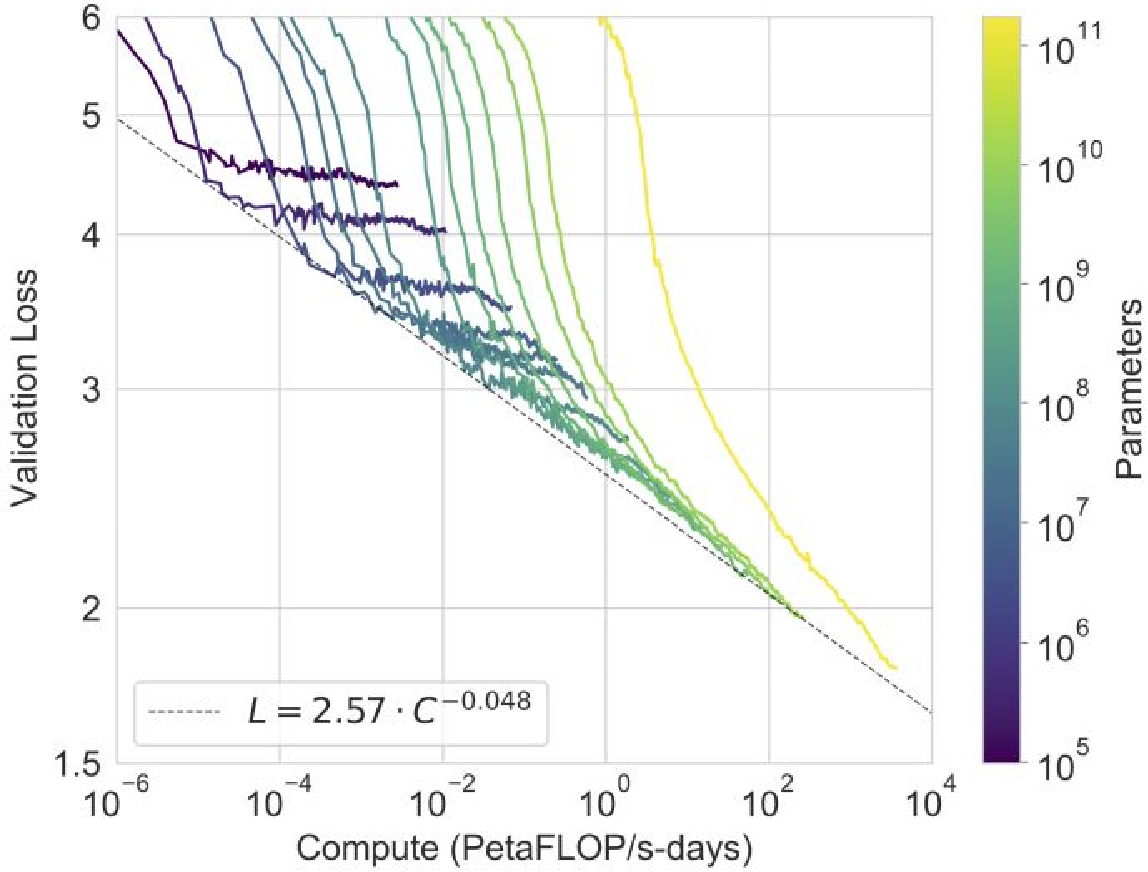


Figure 3.1: Smooth scaling of performance with compute. Performance (measured in terms of cross-entropy validation loss) follows a power-law trend with the amount of compute used for training.

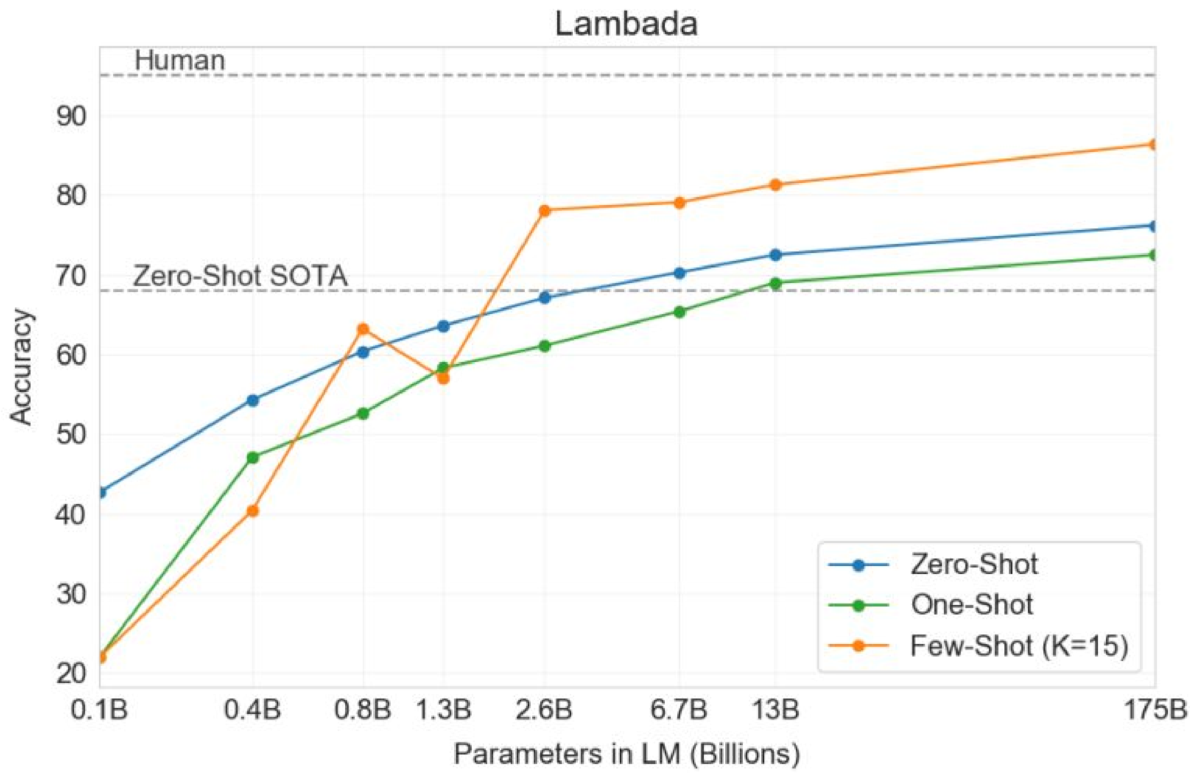


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text. **the model is asked to predict the last word of sentences which requires reading a paragraph of context.**

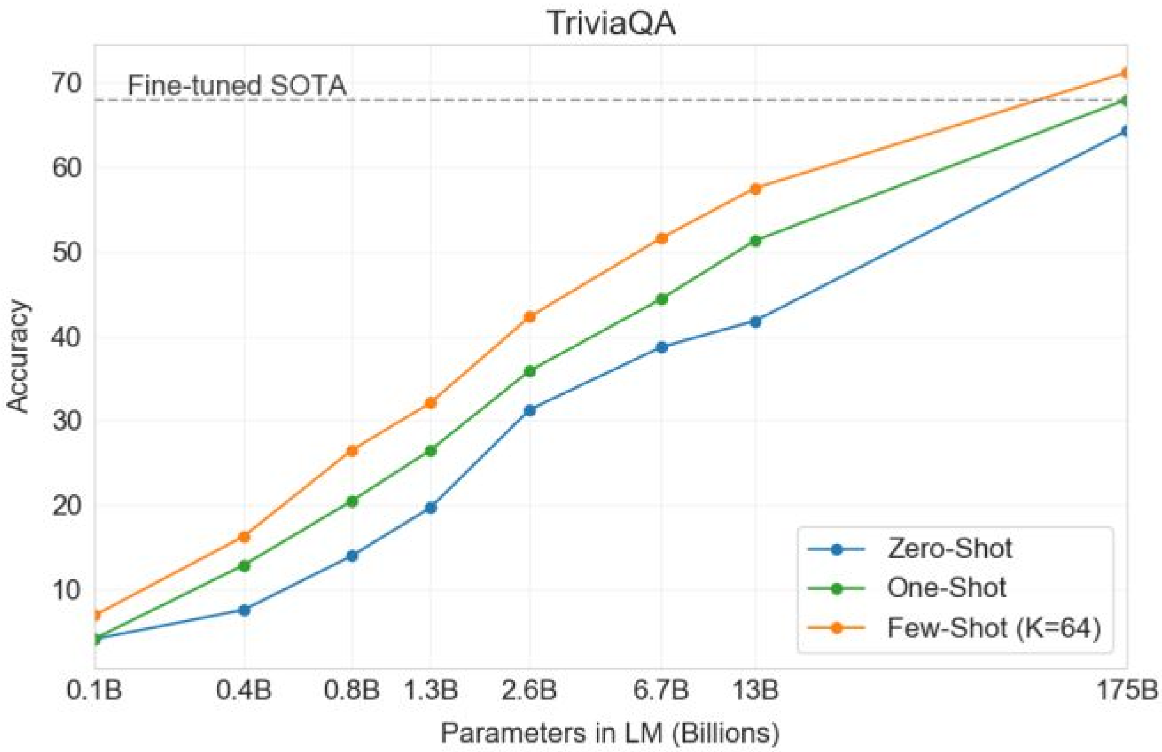


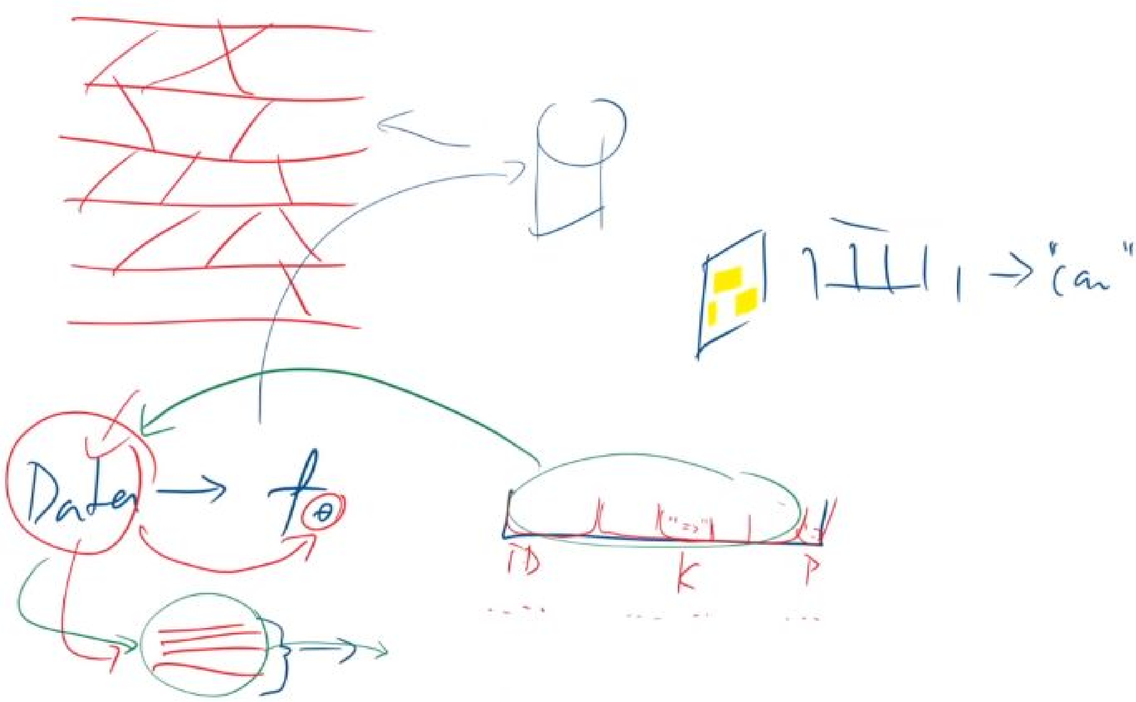
Figure 3.3: On TriviaQA GPT3's performance grows smoothly with model size, suggesting that language models continue to absorb knowledge as their capacity increases. One-shot and few-shot performance make significant gains over zero-shot behavior, matching and exceeding the performance of the SOTA fine-tuned open-domain model, RAG measure **GPT-3's ability to answer questions about broad factual knowledge no access to wiki closed booking setting**

Answer the Q.
Who climbed Everest 1st => Hilary
How . . . . => . . .
Who Q. of Eng => ____

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Table 3.3: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and

zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. **TriviaQA few-shot result is evaluated on the wiki split test server.**



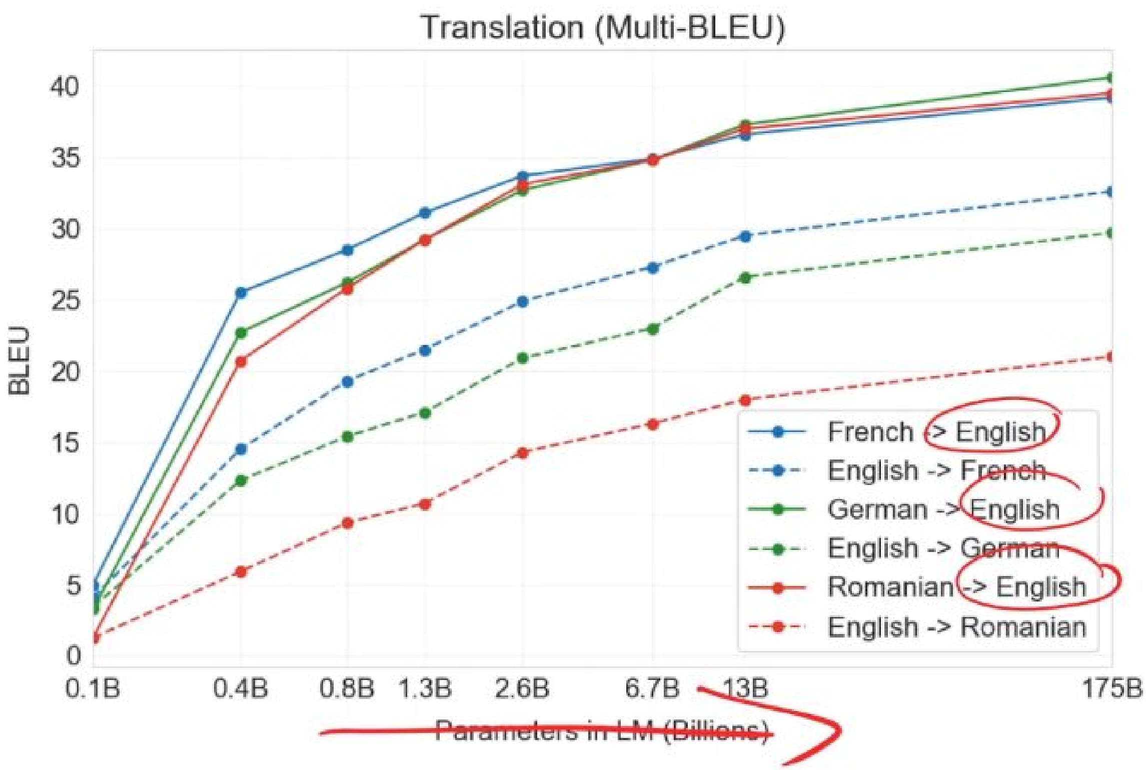
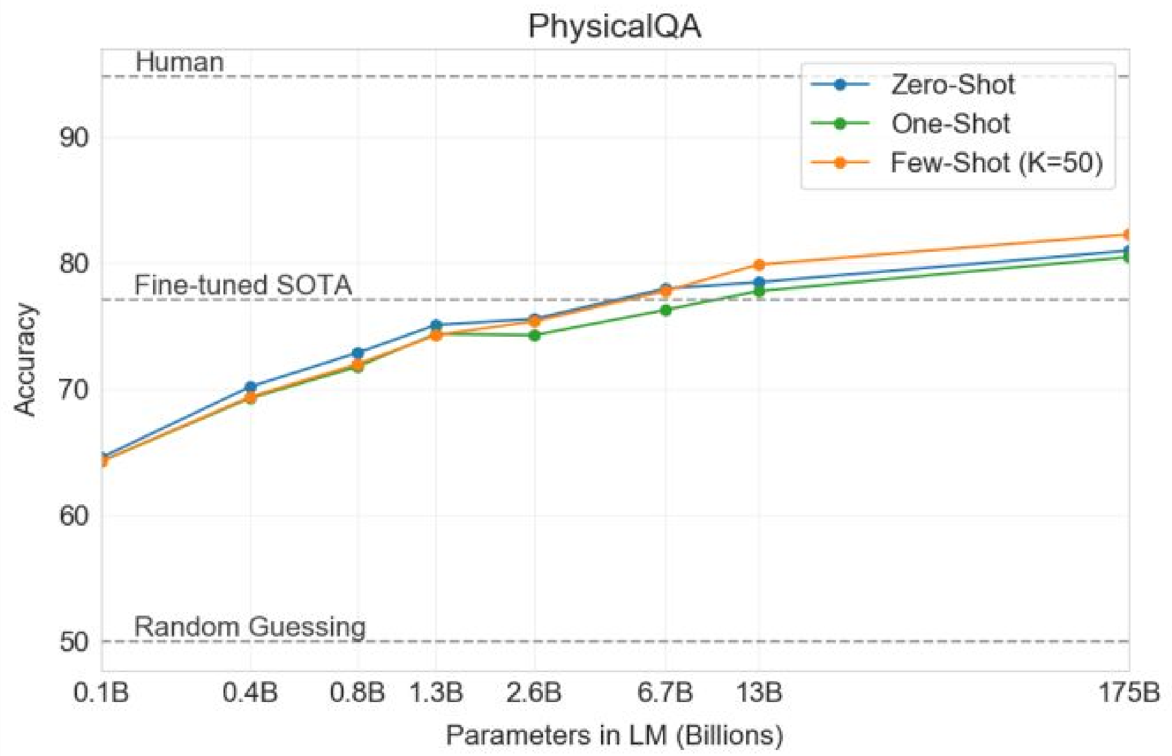


Figure 3.4: Few-shot translation performance on 6 language pairs as model capacity increases. There is a consistent trend of improvement across all datasets as the model scales, and **as well as the tendency for translation into English to be stronger than translation from English.**



Common Sense Reasoning, consider three datasets which attempt to capture physical or scientific reasoning, as distinct from sentence completion, reading comprehension, or broad knowledge question answering The first, **PhysicalQA (PIQA) [BZB+19], asks common sense questions about how the physical world works** and is intended as a probe of grounded understanding of the world. GPT-3 achieves 81.0% accuracy zero-shot, 80.5% accuracy one-shot, and 82.8% accuracy few-shot (the last measured on PIQA's test server).

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Table 3.6: GPT-3 results on three commonsense reasoning tasks, PIQA, ARC, and OpenBookQA. GPT-3 Few-Shot PIQA result is evaluated on the test server. See Section 4 for details on **potential contamination issues on the PIQA test set.**

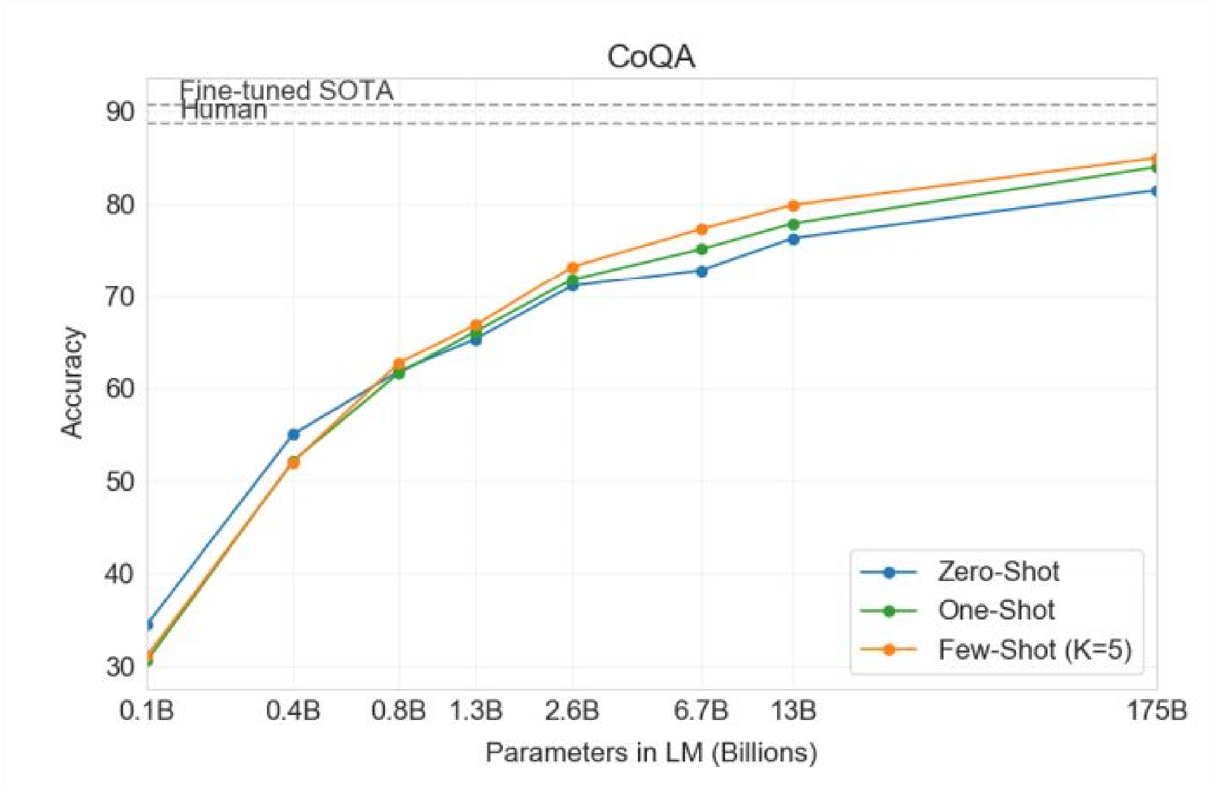
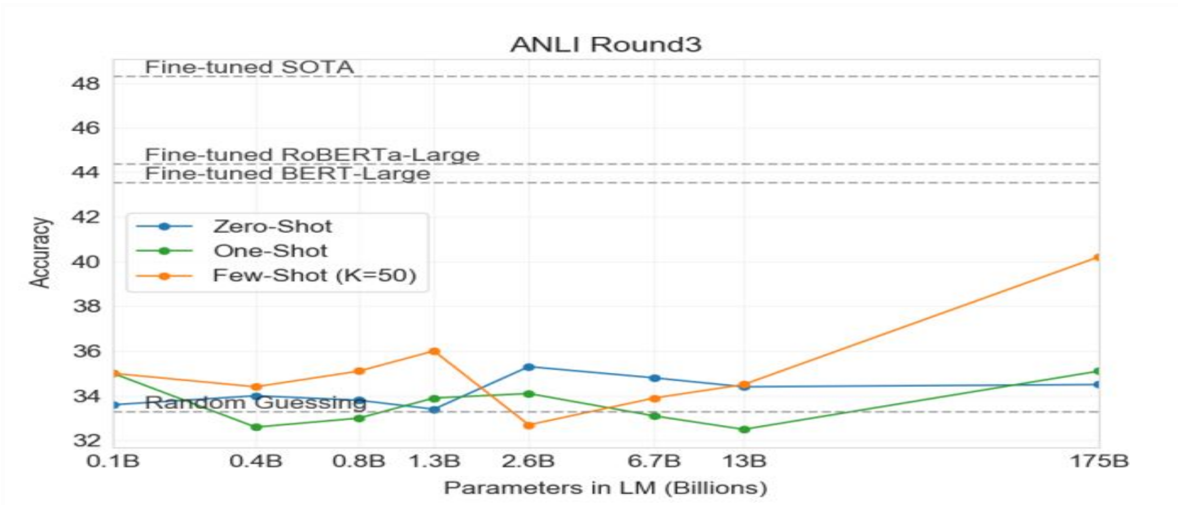


Figure 3.7: GPT-3 results on CoQA reading comprehension task. GPT-3 175B achieves 85 F1 in the few-shot setting, only a few points behind measured human performance and state-of-the-art fine-tuned models. Zero-shot and one-shot performance is a few points behind, with the gains to few-shot being largest for bigger models. We use a suite of 5 datasets including abstractive, multiple choice, and span based answer formats in both dialog and single question settings, **read a piece of text, and answer a question above this text.**

NLI Natural Language Inference concerns the ability to understand the relationship between two sentences.



• 2 digit addition (2D+) – The model is asked to add two integers sampled uniformly from [0, 100), phrased in the form of a question, e.g. "Q: What is 48 plus 76? A: 124." • 2 digit subtraction (2D-) – The model is asked to subtract two integers sampled uniformly from [0, 100); the answer may be negative. Example: "Q: What is 34 minus 53? A: -19".

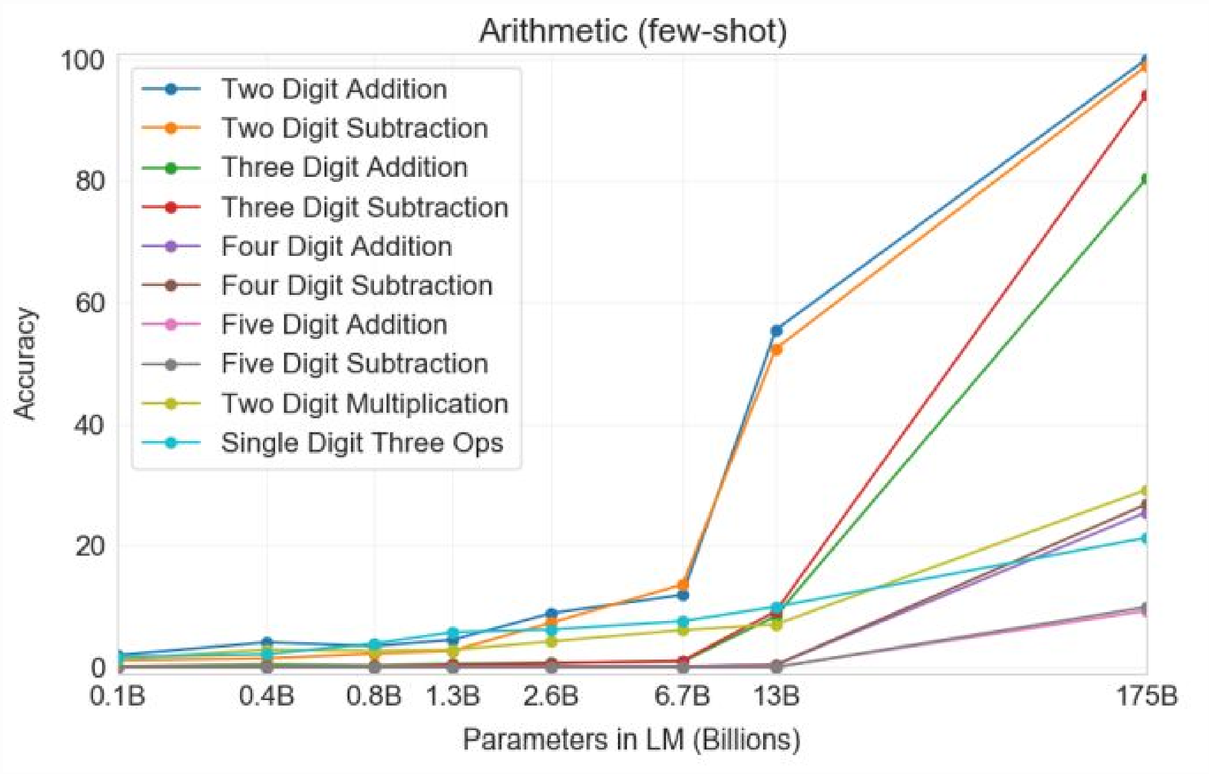


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

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

Google 98 45 143 18 55 73 72 46 118 12 89 101

Alle | Maps | Shopping | Bilder | Videos | Mehr | Einstellungen | Suchfilter

Ungefähr 963'000'000 Ergebnisse (0.61 Sekunden)

Hinweise zum Datenschutz bei Google

SPÄTER ERINNERN | ANSEHEN

books.google.ch › books - Diese Seite übersetzen

Manpower Report of the President

04 **89** . **72 118** . 08 108 . 41 132 . 06 102 . 97 Manufacturing 129 . 51 117 . **18** 131 . ... **73 143** . 05 164 . 56 122 . 51 132 . 07 135 . 71 104 . 34 100 . 28 124 . **98** 147 . ... **72 101** . 59 80 . 39 122 . 09 133 . 77 91 . 80 91 . **72** 114 . 24 138 . 09 122 . ... 14 74 . 24 74 . 48 91 . **46** 112 . 19 96 . **12** 102 . 92 **89** . 10 107 . **45** 91 . 39 **73** .

United States. President - 1970 - Labor supply

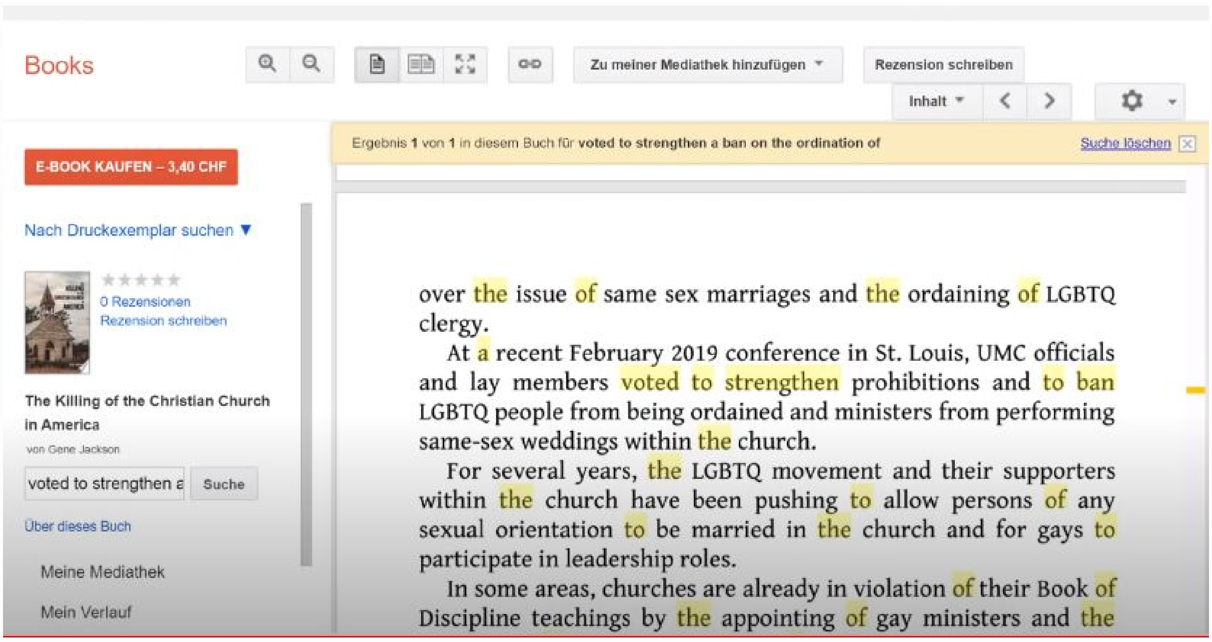
Mean human accuracy (the ratio of correct assignments to non-neutral assignments per participant) at detecting that the intentionally bad articles were model generated was $\sim 86\%$ where 50% is chance level performance. By contrast, mean human accuracy at detecting articles that were produced by the 175B parameter model was barely above chance at $\sim 52\%$ (see Table 3.11).[5] Human abilities to detect model generated text appear to decrease as model size increases: there appears to be a trend towards chance accuracy with model size, and human detection of GPT-3 is close to chance.[6] This is true despite the fact that participants spend more time on each output as model size increases (see Appendix E).

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.

Figure 3.14: The GPT-3 generated news articles that humans had the greatest difficulty distinguishing from a human-written article (accuracy: 12%).

Title: Star’s Tux Promise Draws Megyn Kelly’s Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: **A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can’t change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that’s a really special thing, to not change yourself. I think it’s a really special thing to say, ‘This is what’s inside of me, I’m proud of it, and I’m not going to be ashamed because of the way that someone else thinks I should be.’" Now, it’s the Oscars, and Phoenix is at it again. But this time, his publicist is saying he’ll be wearing a tux no matter what.**
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you’re saying you’re a shape-shifter. I don’t know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn’t stick. "I was like, ‘Okay, I’m going to wear a tuxedo to this thing.’ And then I thought, ‘I don’t want to wear a tuxedo to this thing.’" Kelly goes on to encourage him to change his mind again, but Phoenix says it’s too late: "I’m committed to wearing this."

Figure 3.15: The GPT-3 generated news articles that humans found the easiest to distinguish from a human-written article (accuracy: 61%).



ST. LOUIS (AP) — The United Methodist Church, America's second-largest Protestant denomination, faces a likely surge in defections and acts of defiance after delegates at a crucial conference voted Tuesday to strengthen the faith's divisive bans on same-sex marriage and ordination of LGBT clergy.

Emotions were high throughout the third and final day of the UMC's meeting. Some supporters of greater LGBT inclusion were in tears, while others vented their anger when, midway through the session, delegates defeated a proposal that would have let regional and local church bodies decide for themselves on gay-friendly policies.

"Devastation," was how former Methodist pastor Rebecca Wilson of Detroit described her feelings. "As someone who left because I'm gay, I'm waiting for the church I love to stop bringing more hate."

---Qiang's report on 04.09.2020 regarding at GPT-3 research study