NLG

ⓘ *... reviewing machine learning literature about NLG (Natural Language Generation)*

- *how to generate text (automatic commentary) from*
    1. *numbers (fund data from clients and system)*
    2. *short text (user and system commentary, and corporate actions)*
    3. *long text (news articles)*

Language Models are Few-Shot Learners

- Goal : introduce GPT-3 model, an autoregressive language model with 175 billion parameters, train the model, and test its performance in the few-shot setting
- Link : https://arxiv.org/abs/2005.14165 (05/2020)
- Code : https://github.com/openai/gpt-3
- Credible source: Open AI

GPT(Generative Pre-trained Transformer)

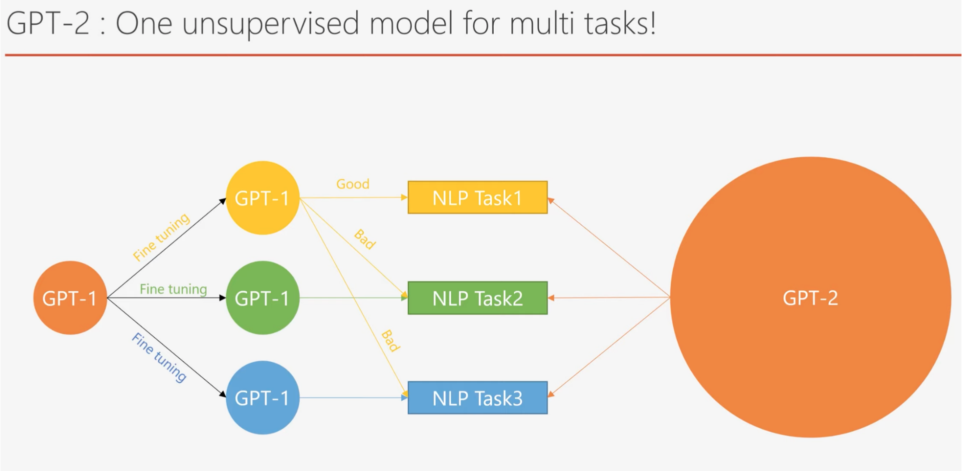
- task-specific vs task-agnostic
    - trained specifically on a particular task like sentiment classification, textual entailment, etc. using supervised learning
    - limitations : need large amount of annotated data for learning a particular task, fail to generalise for tasks other than what they have been trained for
- autoregressive language model
    - language model predicts next token using given tokens
    - autoregressive language model : language model's current output depends on previous output (i.e. RNN, transformer)
- generative model requires enough data to represent real distribution
- pre-trained transformer language model
    - use transformer decoder
    - transformer models focus on machine translation
    - GPT-1 focus on more NLP tasks such as
        - natural language inference,
        - question answering,
        - semantic similarity,
        - classification

GPT-1	GPT-2

1. learning objectives and concepts
    - unsupervised pre-training followed by supervised fine-tuning
        - a. unsupervised language modelling (pre-training)
            - standard language model objective was used
        - b. supervised fine-tuning
            - fine tuning without additional task specific model
            - just adding a linear and a softmax layer to get the task labels for downstream tasks
        - c. task specific input transformation
            - start and end tokens added to the input sequences
            - delimiter token added between different parts of example
2. dataset
    - BooksCorpus dataset - 7000 unpublished books
3. performance and summary
    - performed better than specifically trained supervised state-of-the-art models in 9 out of 12 tasks
    - proved that language model served as an effective pre-training objective which could help model generalise well
    - facilitated transfer learning and could perform various NLP tasks with very little fine-tuning
4. original paper
    - Improving Language Understanding by Generative Pre-Training

1. learning objectives and concepts
    - task conditioning
        - aimed at learning multiple tasks using the same unsupervised model (without fine-tuning)
        - the model is expected to produce different output for same input for different tasks
        - task conditioning for language models is performed by providing examples or natural language instructions to the model to perform a task
        - LM objective function : P(output|input, task) vs P(output|input)
2. dataset
    - WebText dataset
        - scraped the Reddit platform and pulled data from outbound links of high upvoted articles
        - 40GB of text data from over 8 million documents
3. performance and summary
    - achieve state-of-the-art results on 7 out of 8 tested language modelling datasets
    - showed that training on larger dataset and having more parameters improved the capability
    - with increase in the capacity of the model, the performance increased in log-linear fashion
    - perplexity of language models decrease with increase in number of parameters
4. original paper
    - Language Models are Unsupervised Multitask Learners

size	GPT-1	GPT-2
parameters	117 millions	1.5 billions
layers	12	48
states dimension	768	1600
context token size	512	1024
batch size	64	512


GPT-2 : One unsupervised model for multi tasks!

GPT-3

learning objectives and concepts

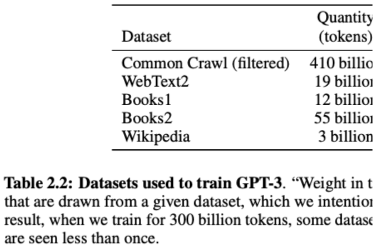
- in-context learning
  - while learning the primary objective of predicting the next word given context words, the language models also start recognising patterns in data
- few-shot learning
  - the model is provided with task description and as many examples as fit into the context window of model

model architecture and implementation details

- 175 billion parameters
- context window size was increased to 2048 tokens
- size of word embeddings was increased to 12888
- 96 layers with each layer having 96 attention heads

training dataset

- Common Crawl dataset
- 3 steps to improve the average quality of datasets
  - filtered based on similarity to a range of high-quality reference corpora
  - performed fuzzy deduplication at the document level
  - added known high-quality reference corpora
- trained on a mix of five different corpora
  - WebText2, Books1, Books2 and Wikipedia
- 300 billion tokens, some datasets are seen up to 3.4 times during training

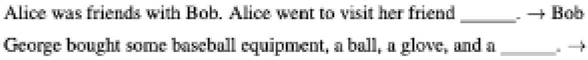


Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billio
WebText2	19 billio
Books1	12 billio
Books2	55 billio
Wikipedia	3 billion

Table 2.2: Datasets used to train GPT-3. "Weight in t that are drawn from a given dataset, which we intentio result, when we train for 300 billion tokens, some datas are seen less than once.

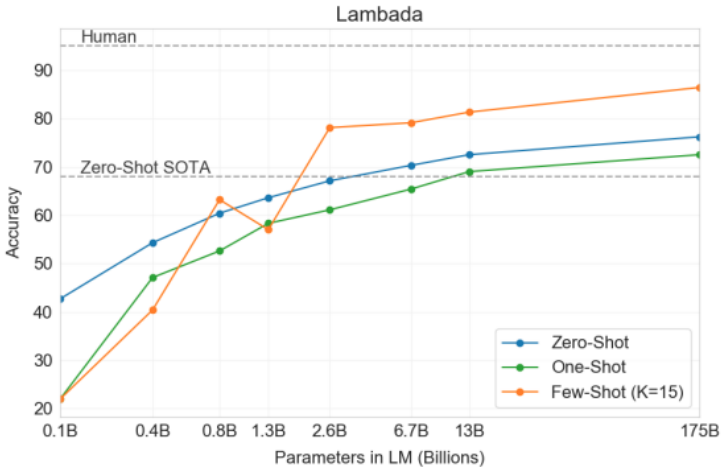
3.1 Language Modelling task

- LAMBADA dataset - test how well the model predict the last word of sentences



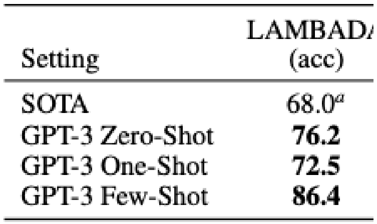
Alice was friends with Bob. Alice went to visit her friend ______. → Bob

George bought some baseball equipment, a ball, a glove, and a ______. →



achieves 86.4% accuracy in the few-shot setting, an increase of over 18% from the previous state-of-the-art

- HellaSwag dataset - test for picking the
- StoryCloze dataset - test for selecting



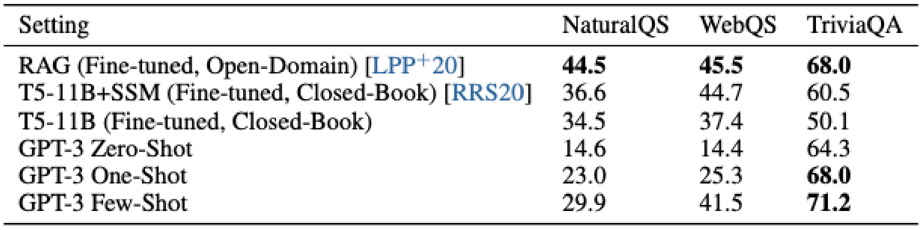
Setting	LAMBAD (acc)
SOTA	68.0[a]
GPT-3 Zero-Shot	**76.2**
GPT-3 One-Shot	**72.5**
GPT-3 Few-Shot	**86.4**

human's accuracy is 95.6%

HellaSwag : achieves 79.3% accuracy in th achieved by the fine-tuned multi-task model

StoryCloze : achieves 87.7% accuracy in th using a BERT based model

3.2 Closed Book Question Answering

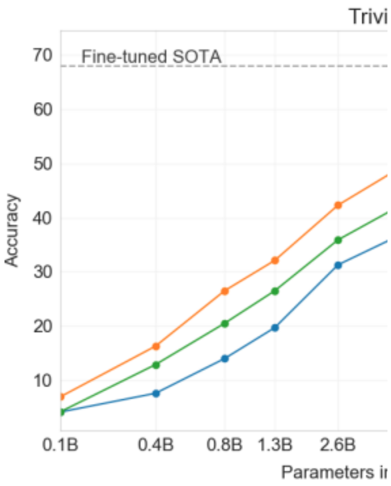


Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

one of the three datasets GPT-3's one-shot matches the open-domain fine-tuning SOTA

other two datasets it approaches the performance of the closed-book SOTA despite not using fine-tuning

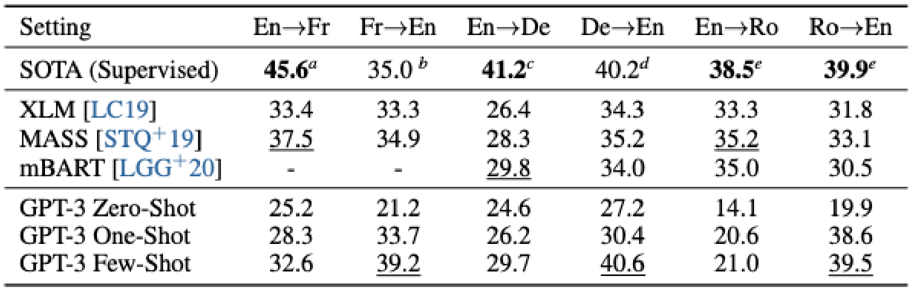
all 3 datasets, we find that performance scales very smoothly with model size possibly reflecting the idea that model capacity translates directly to more 'knowledge' absorbed in the parameters of the model



for Trivia QA dataset, achieves 71.2% in th

zero-shot result already outperforms the fir

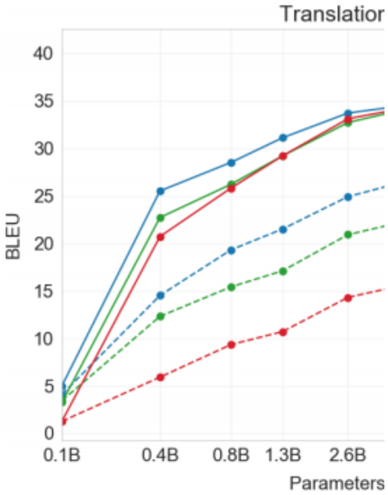
3.3 Translation



Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

BLEU score

few-shot setting outperforms on FrenchEnglish, GermanEnglish, RomanianEnglish

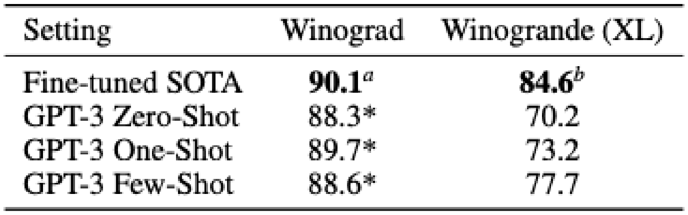


few-shot translation performance on 6 lang

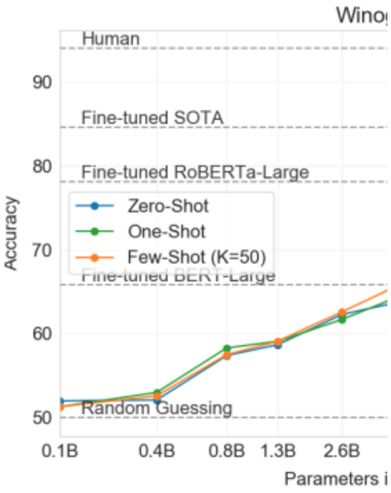
improvement across all dataset as the moc

3.4 Winograd-Style Tasks

- classical task in NLP that involves determining which word a pronoun refers to
- when the pronoun is grammatically ambiguous but semantically unambiguous to a human
- normally bi-directional model is better
- i.e. https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.xml,
- i.e. https://www.tensorflow.org/datasets/catalog/winogrande



Setting	Winograd	Winogrande (XL)
Fine-tuned SOTA	**90.1**[a]	**84.6**[b]
GPT-3 Zero-Shot	88.3*	70.2
GPT-3 One-Shot	89.7*	73.2
GPT-3 Few-Shot	88.6*	77.7

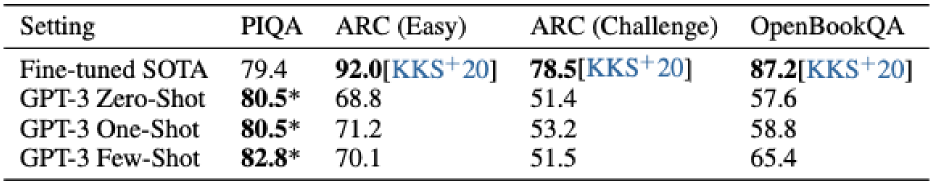


achieved 77.7% in the few-shot setting

fine-tuned RoBERT model achieved 79%, 94.0%

3.5 Common Sense Reasoning

- ask about physical or scientific reasoning



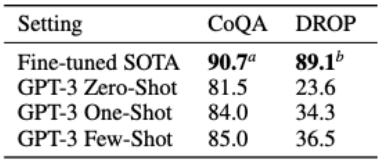
Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

PhysicalQA dataset asks common sense questions about how the physical world works

achieves 82.8% accuracy on few-shot setting better than fine-tuned RoBERT model, 79.4% which is still over 10% worse than human performance

3.6 Reading Comprehension

- difficult task for NLP



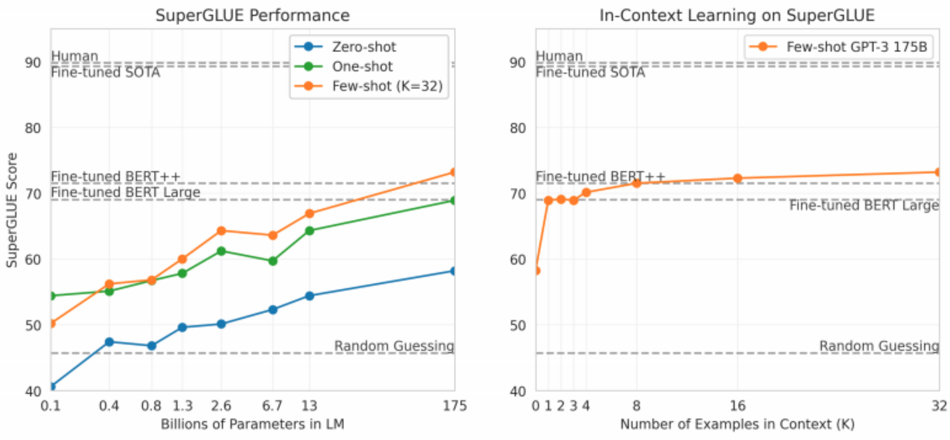
Setting	CoQA	DROP
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]
GPT-3 Zero-Shot	81.5	23.6
GPT-3 One-Shot	84.0	34.3
GPT-3 Few-Shot	85.0	36.5

GPT-3 performs best on CoQA, a free-form

performs weakly compare to the fine-tuned

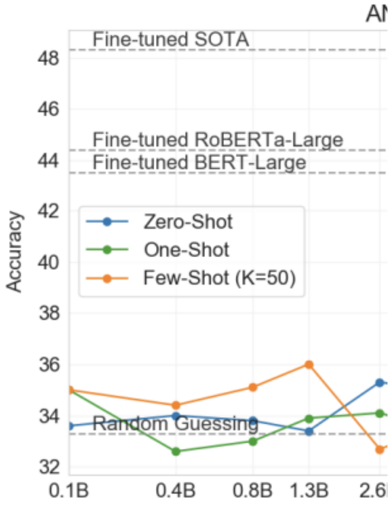
3.7 SuperGLUE

- a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard
- evaluate GPT-3 on a standardised collection of datasets, the SuperGLUE benchmark by comparing to BERT models



3.8 NLI (Natural Language Inference)

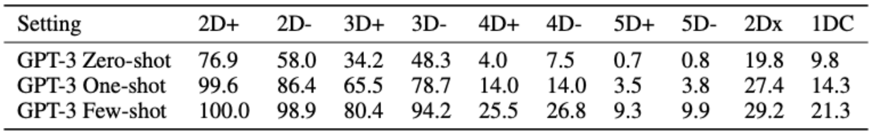
- the ability to understand the relationsh



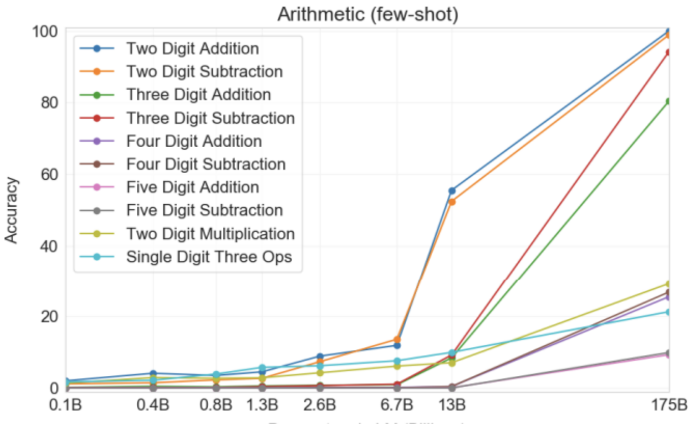
NLI is still a very difficult task for language

3.9 Synthetic and Qualitative Tasks

1. Arithmetic



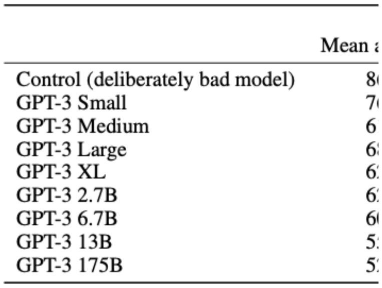
Setting	2D+	2D-	3D+	3D-	4D+	4D-	5D+	5D-	2Dx	1DC
GPT-3 Zero-shot	76.9	58.0	34.2	48.3	4.0	7.5	0.7	0.8	19.8	9.8
GPT-3 One-shot	99.6	86.4	65.5	78.7	14.0	14.0	3.5	3.8	27.4	14.3
GPT-3 Few-shot	100.0	98.9	80.4	94.2	25.5	26.8	9.3	9.9	29.2	21.3



good at simple math

2. News Article Generation

- arbitrarily selected 25 article titles and
- generated completions of these titles a
- presented around 80 US-based partici written by a human", "more likely writte machine", or "very likely written by a m



	Mean a
Control (deliberately bad model)	8
GPT-3 Small	7
GPT-3 Medium	6
GPT-3 Large	6
GPT-3 XL	6
GPT-3 2.7B	6
GPT-3 6.7B	6
GPT-3 13B	5
GPT-3 175B	5

mean human accuracy at detecting that the

52% of articles are detected as model gene

Limitations

- loses coherency while formulating long sentences and repeats sequences of text over and over again
- does not perform very well on tasks like natural language inference (determining that if a sentence implies other sentence), fill in the blanks, some rea
- in this paper, it cites unidirectionality of GPT models as the probable cause for these limitations and suggests training bidirectional models at this scale
- potential risks of misuse of its human-like text generating capability for phishing, spamming, spreading misinformation or performing other fraudulent a
- text generated by GPT-3 possesses the biases (gender, ethnicity, race or religion) of the language it is trained on

perplexity

- a measure of how easy a probability distribution is to predict
- a measure of prediction error
- the lower the better model

GLUE (General Language Understanding Evaluation) benchmark

- collection of resources for training, evaluating, and analysing natural language understanding systems
- https://gluebenchmark.com/

ref.

- The Journey of Open AI GPT models
- Better Language Models and their Implications
- How GPT3 Works
- GPT-3 projects