Seoul National University
Graduate School of Data Science

GPT-3

Language Models are Few-Shot Learners

성주용, Causality Lab, SNU

4월 3일, 2025년

Abstract

- Recent work requires task-specific fine-tuning datasets.
- In contrast, humans can generally perform a new language task from only a few examples
  - Something which current NLP systems still struggle
- Scaling up language models greatly improves task-agnostic, few-shot performance => GPT-3
  - Without any gradient updates or fine-tuning

Introduction

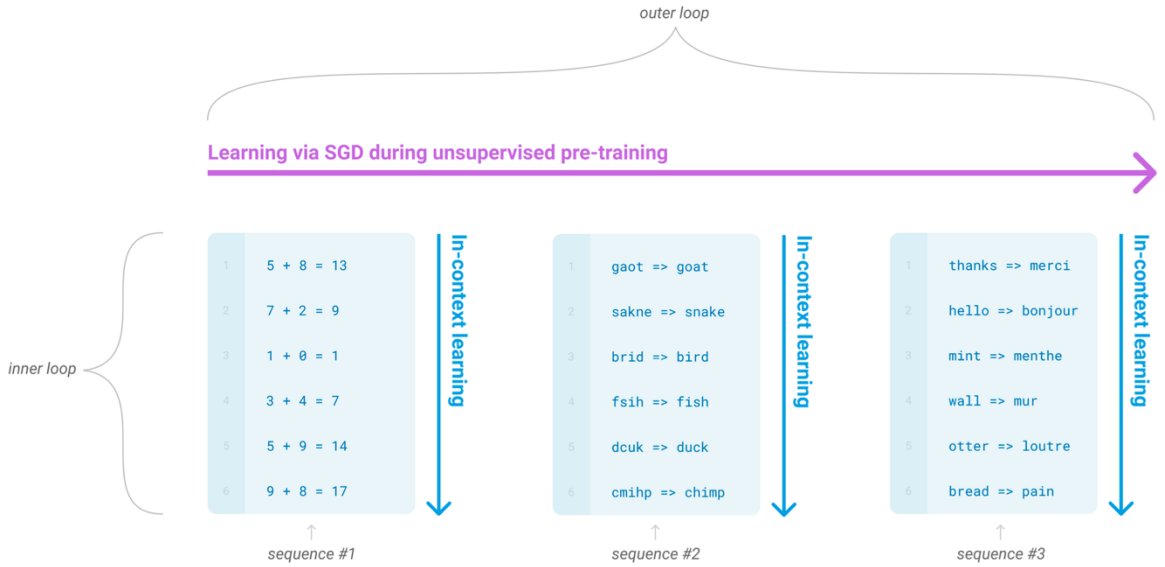
Removing limitation requiring task-specific datasets

- Practical perspective
  - Limits the applicability of language models
  - Difficult to collect a large supervised training dataset
- Potential of spurious correlations in training data
  - Absorb information during pre-training, fine-tuned on very narrow task distributions
  - Does not generalize well
- Human does not act like that
  - Allows humans to seamlessly mix together or switch between many tasks and skills
  - Fluidity and generality

Introduction

Meta-learning

- In-context learning
  - Model is conditioned on a natural language instructions and/ or a few demonstrations of the task
  - Expected to complete further instances of the task
  - Strong gains with model scale



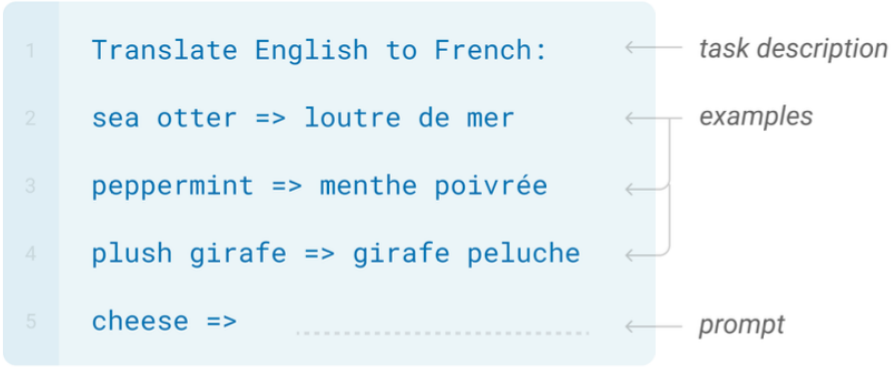
Approach

Few-Shot

- Model is given a few demonstrations at inference time
- K examples of context and completion
  - K is 10 to 100 (that fit model's context window)
- Major reduction for task-specific data
- Much worse than fine-tuning

Few-shot

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



Approach

One-Shot

- Only one demonstration is allowed
- Closely matches the way in which some tasks are communicated to humans
- Sometimes difficult to communicate the content or format of a task

One-shot

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

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

Approach

Zero-Shot

- No demonstrations are allowed
  - Only given a instruction describing the task

- For at least some settings is closest to how humans perform tasks
  - Translations

Zero-shot

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

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

Approach

Model and Architectures

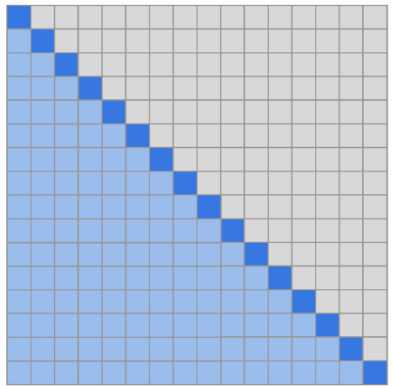
- Use same model and architecture as GPT-2
  - Include modified initialization, pre-normalization, reversible tokenization

- Alternating dense and locally banded sparse attention patterns in the layers of the transformer

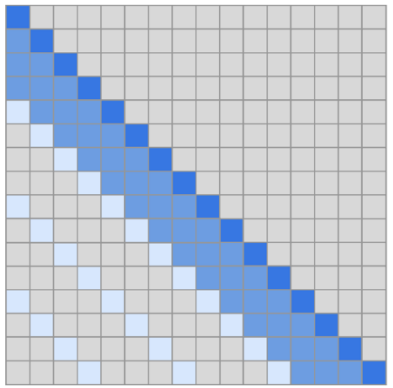
Approach

Model and Architectures

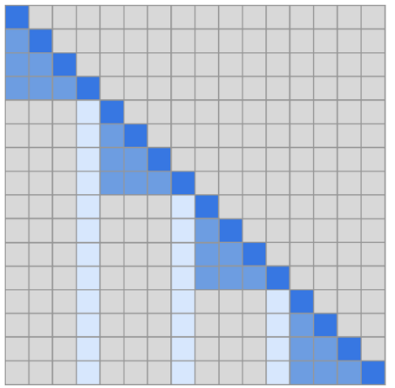
- Sparse Transformer
  - Restructured residual block and weight initialization
  - A set of **sparse attention** kernels which efficiently compute subsets of the attention matrix
  - Recomputation of attention weights during the backwards pass to reduce memory usage



(a) Transformer



(b) Sparse Transformer (strided)



(c) Sparse Transformer (fixed)

Approach

Evaluation

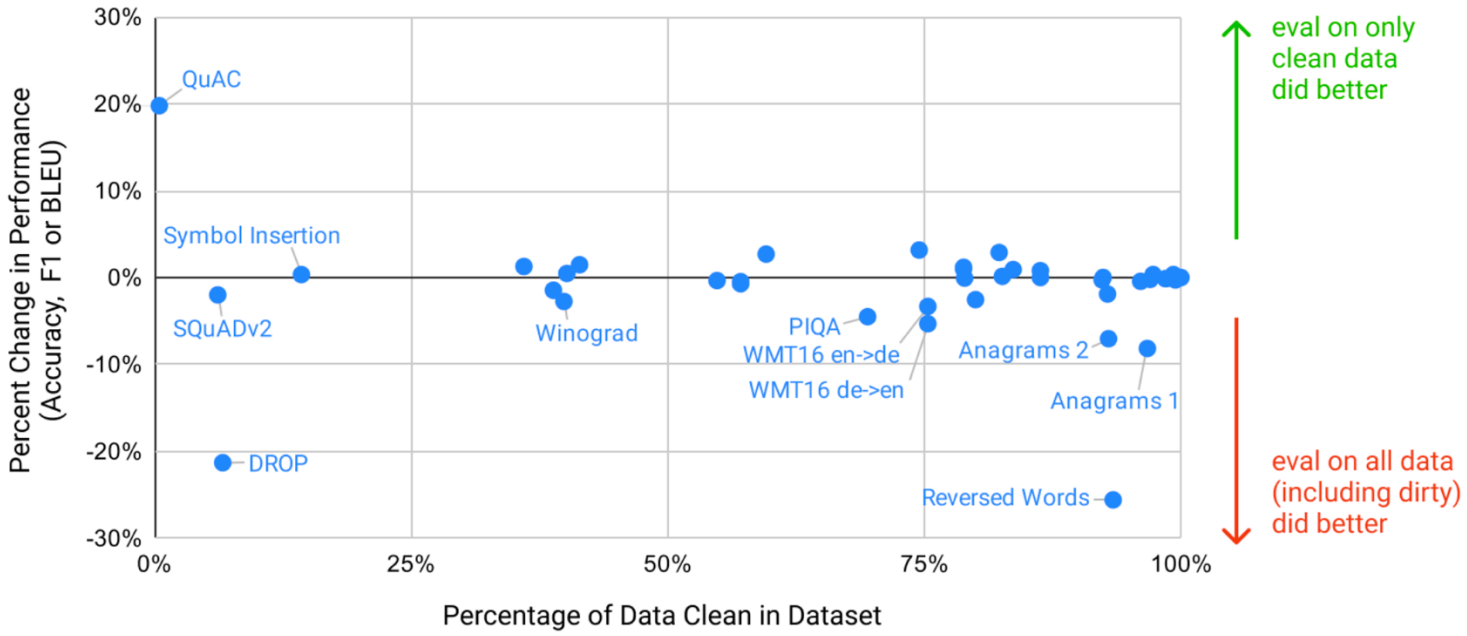
- Choosing one correct completion from several options
  - For most tasks, compare the per-token likelihood
  - For small datasets, compute $\frac{P(completion|context)}{P(completion|answer\_context)}$
- For binary classification, used True/False than 0/1
  - Treat as multiple choice
- For free-form completion, used beam search

Memorization Of Benchmarks

Contamination

- Overlap between training and testing

- Investigated how remaining overlap impacts results

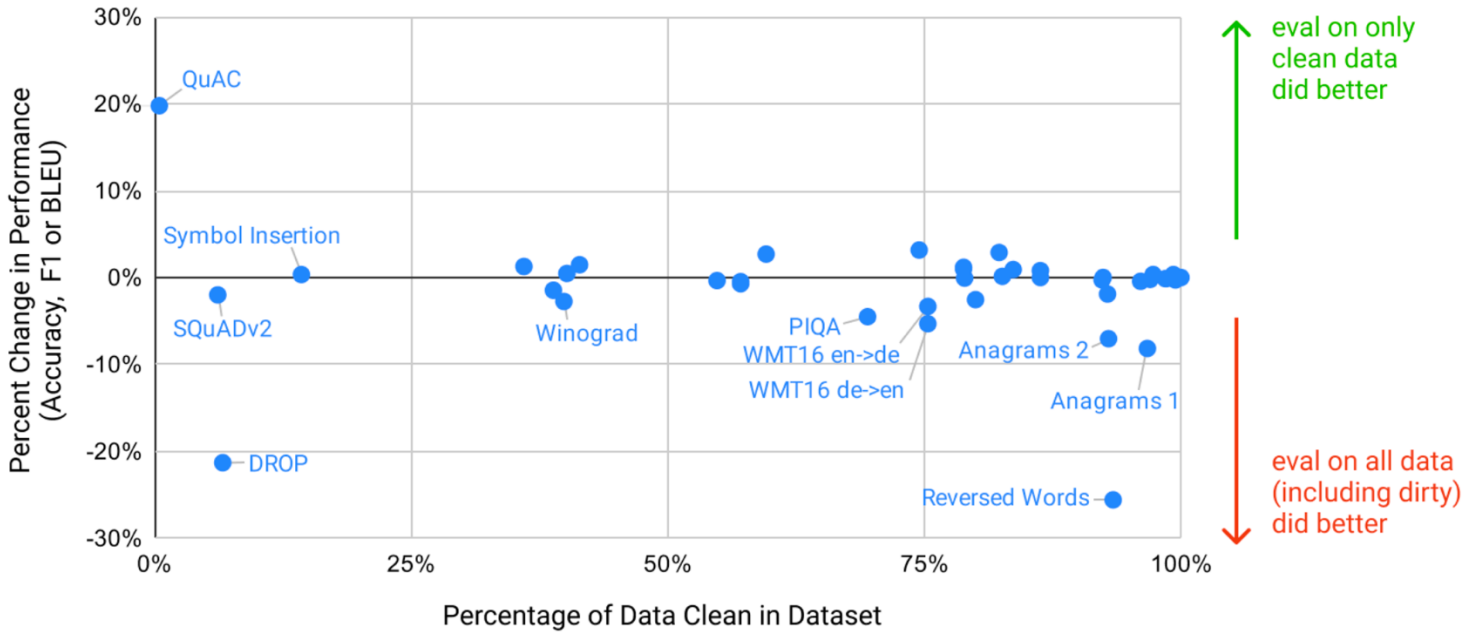


Memorization Of Benchmarks

Contamination

- Either overestimate contamination or has little effect on performance

- Cannot sure that the clean subset is drawn from the same distribution as the original dataset



Limitations

Weaknesses in text synthesis and several NLP tasks

- Text synthesis
  - Lose coherence over long passages
  - Contradict themselves
  - Contain non-sequitur sentences

- Discrete NLP
  - Common sense physics
    - "If I put cheese into the fridge, will it melt?"

Limitations

Structural and algorithmic limitations

- Does not include any bidirectional architectures / denoising

- WIC
  - Compare the use of a word in two sentences

- ANLI
  - Compare two sentences to see if one implies the other

- Reading comprehension tasks

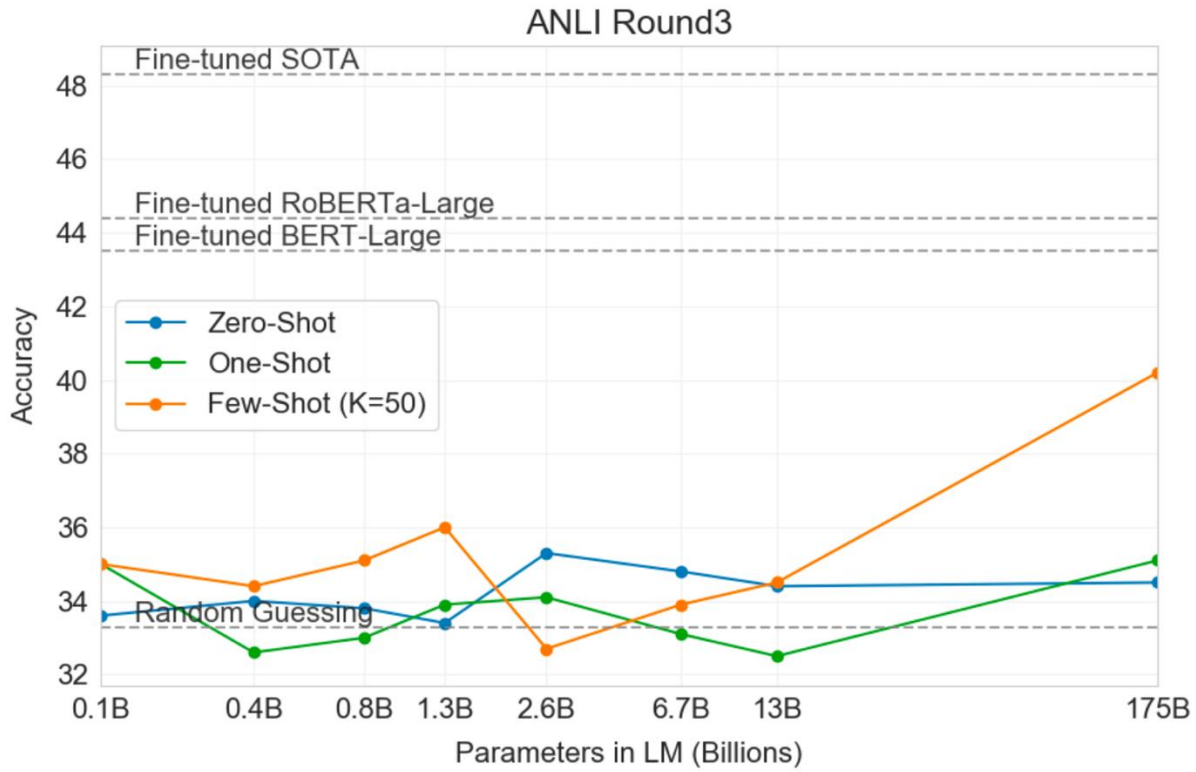
Limitations

WIC Result

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Limitations

ANLI Result



Limitations

Reading comprehension task Result

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Limitations

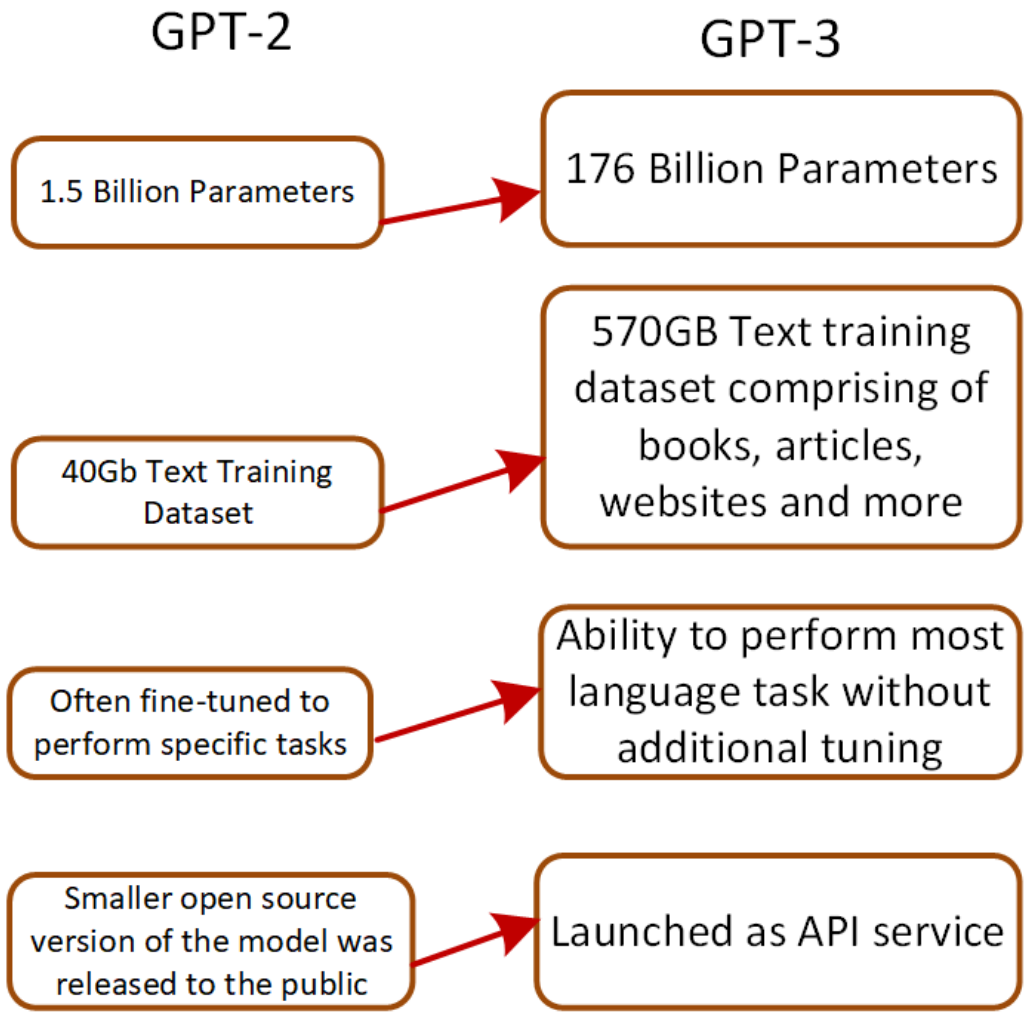
Self-supervised prediction limitation

- Learn the objective function from humans
- Fine-tuning with RL
- Add additional modalities such as images

Limitations

- Improve poor sample efficiency
- Understand precisely how few-shot learning works
- Expensive and inconvenient to perform inference on
  - Challenge of practical applicability
  - Distillation of large models
- Decisions are not easily interpretable

GPT-3 vs GPT-2


GPT-2
GPT-3
1.5 Billion Parameters
176 Billion Parameters
40Gb Text Training Dataset
570GB Text training dataset comprising of books, articles, websites and more
Often fine-tuned to perform specific tasks
Ability to perform most language task without additional tuning
Smaller open source version of the model was released to the public
Launched as API service

End of Document