Language Models are Few-Shot Learners

OpenAI

NeurIPS 2020

Presenter: Shu Zhao

smz5505@psu.edu

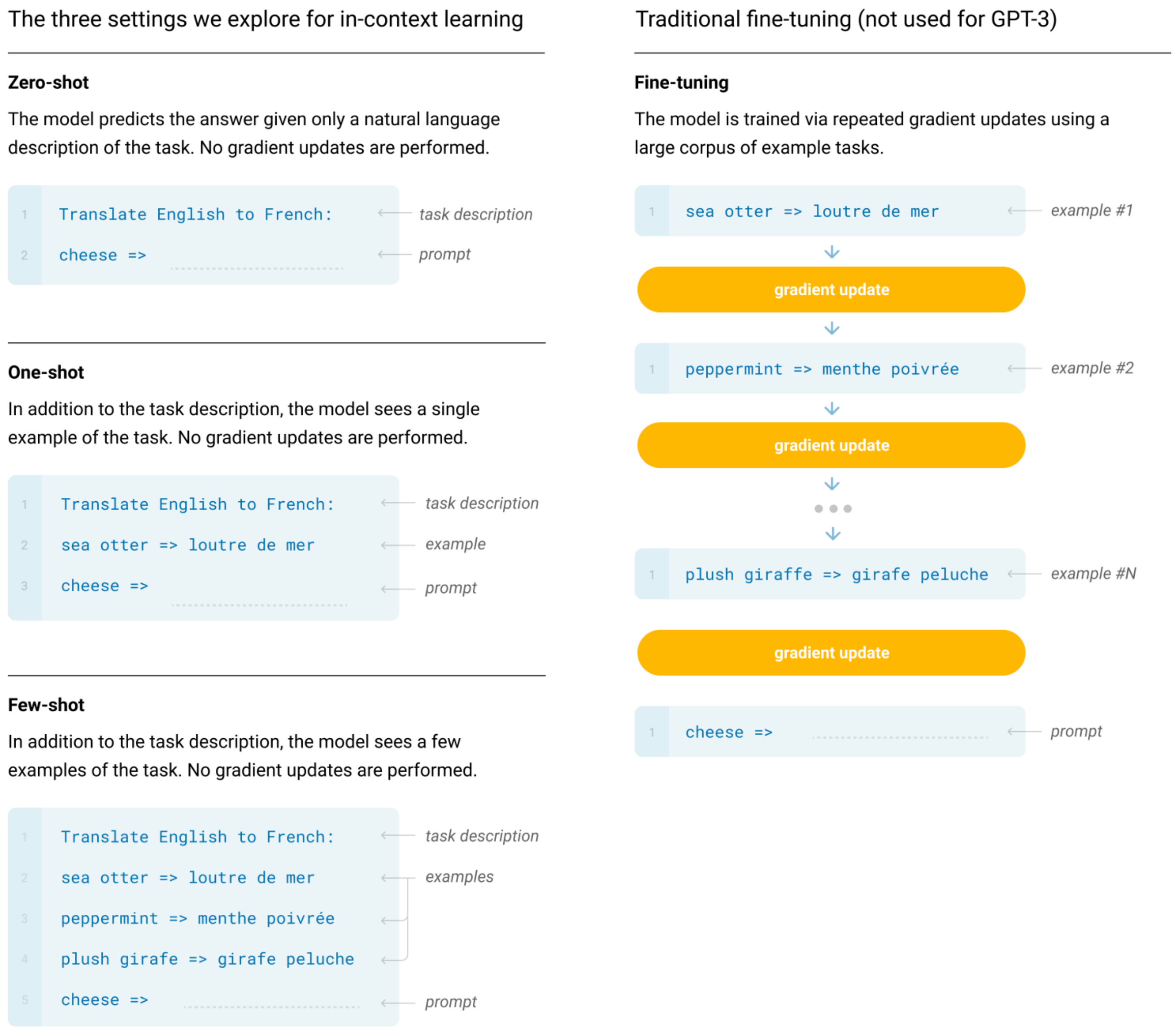
Outline

- Motivation
- Related Work
- Method
- Experiments
- Limitations
- Conclusion
- Feature Work

Motivation

- Remove the requirement of task-specific datasets and task-specific fine-tuning when applying large language models to downstream tasks

Why?



Motivation

Why?

- Expensive data collection and fine-tuning on each task
- Some works pointed out pre-training+fine-tuning paradigm may have poor generalization ability
- Simulate humans that don’t require large supervised datasets

Hendrycks, Dan, et al. Pretrained transformers improve out-of-distribution robustness. ACL 2020.
Yogatama, Dani, et al. Learning and Evaluating General Linguistic Intelligence. 2019.
McCoy, Tom, et al. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019.

Related Work

- GPT-1 (117M)
  - Unsupervised Generative Pre-Training + Supervised Fine-Tuning
  - Transformer Decoder Architecture
  - BooksCorpus + Word Benchmark for unsupervised pre-training, 5GB

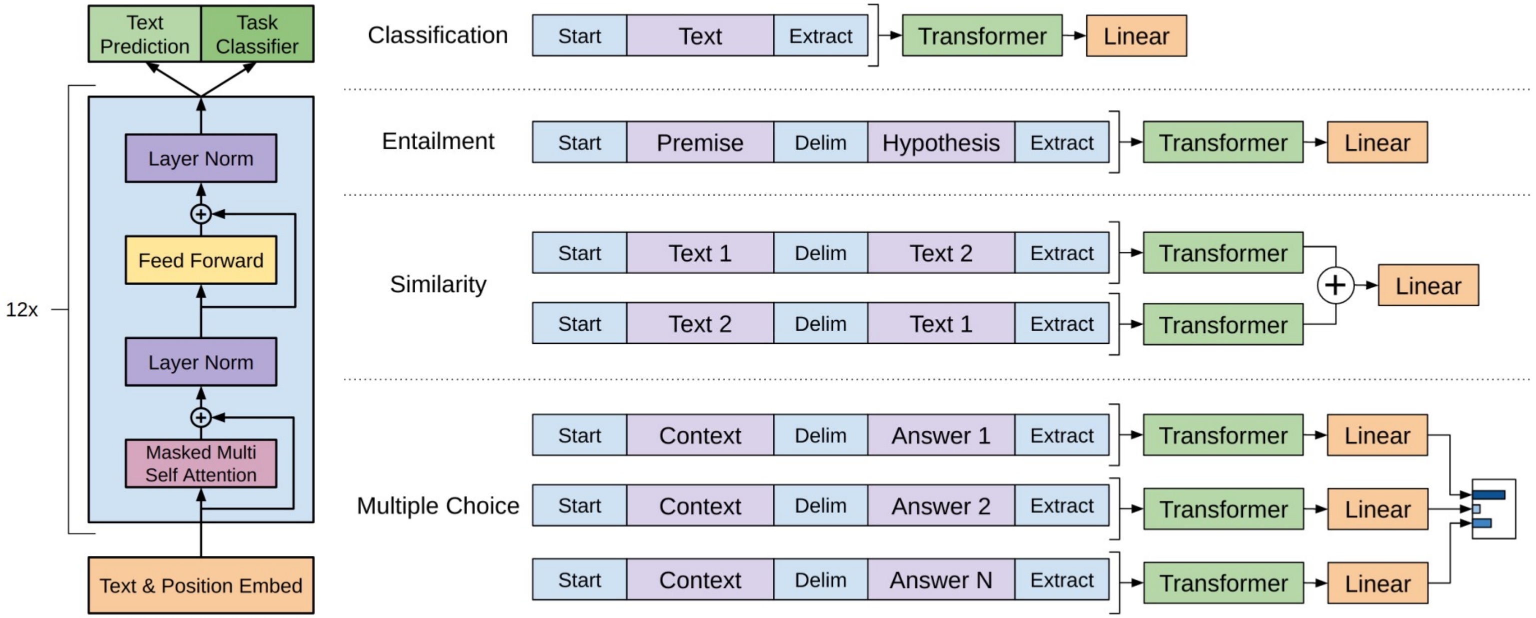


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer

Table 1: A list of the different tasks and datasets used in our experiments.

Task	**Datasets**
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

Radford, Alec, et al. Improving Language Understanding by Generative Pre-Training. 2018.

Related Work

- GPT-2 (1.5B)
  - Unsupervised Generative Pre-Training
  - GPT-1 + Modified layer normalization and initialization
  - WebText, 40GB

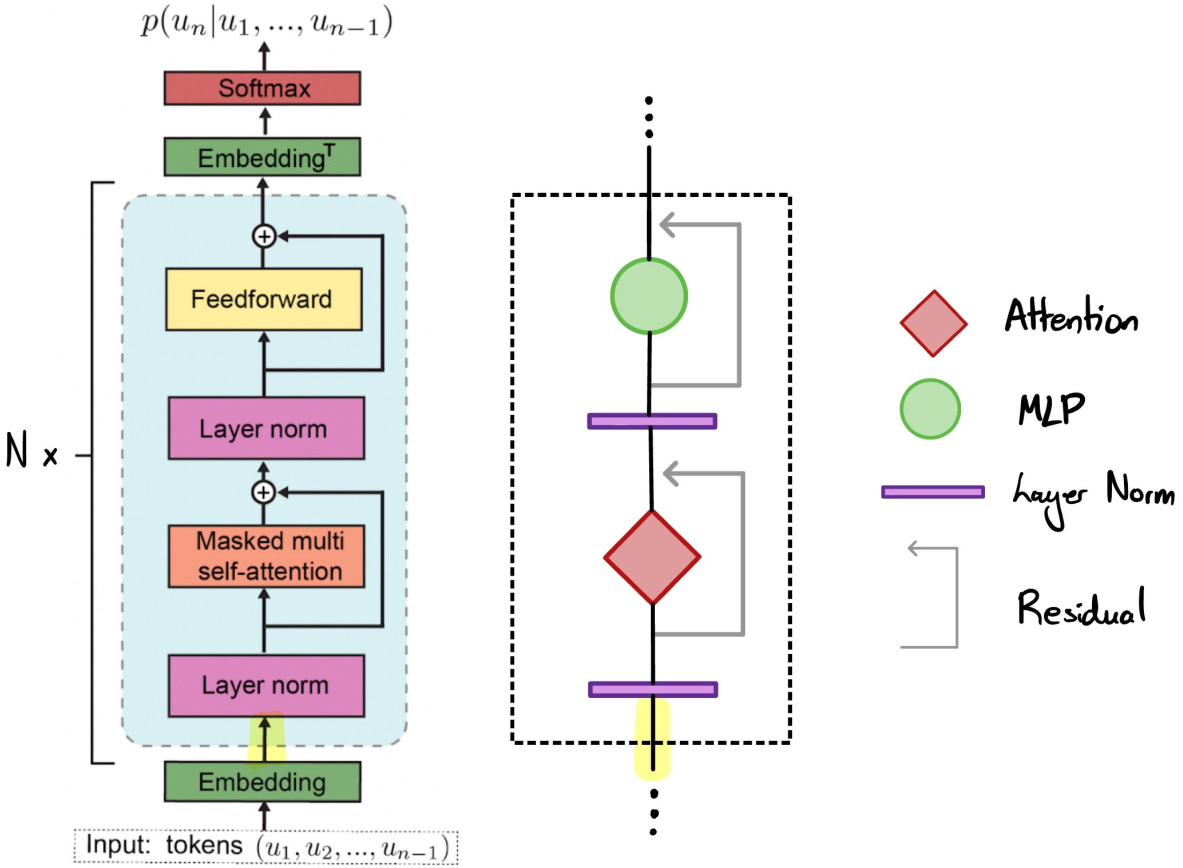


Image from https://www.lesswrong.com/posts/qxvihKpFMuc4tvuf4/recall-and-regurgitation-in-gpt2

Radford, Alec, et al. Language Models are Unsupervised Multitask Learners. 2019.

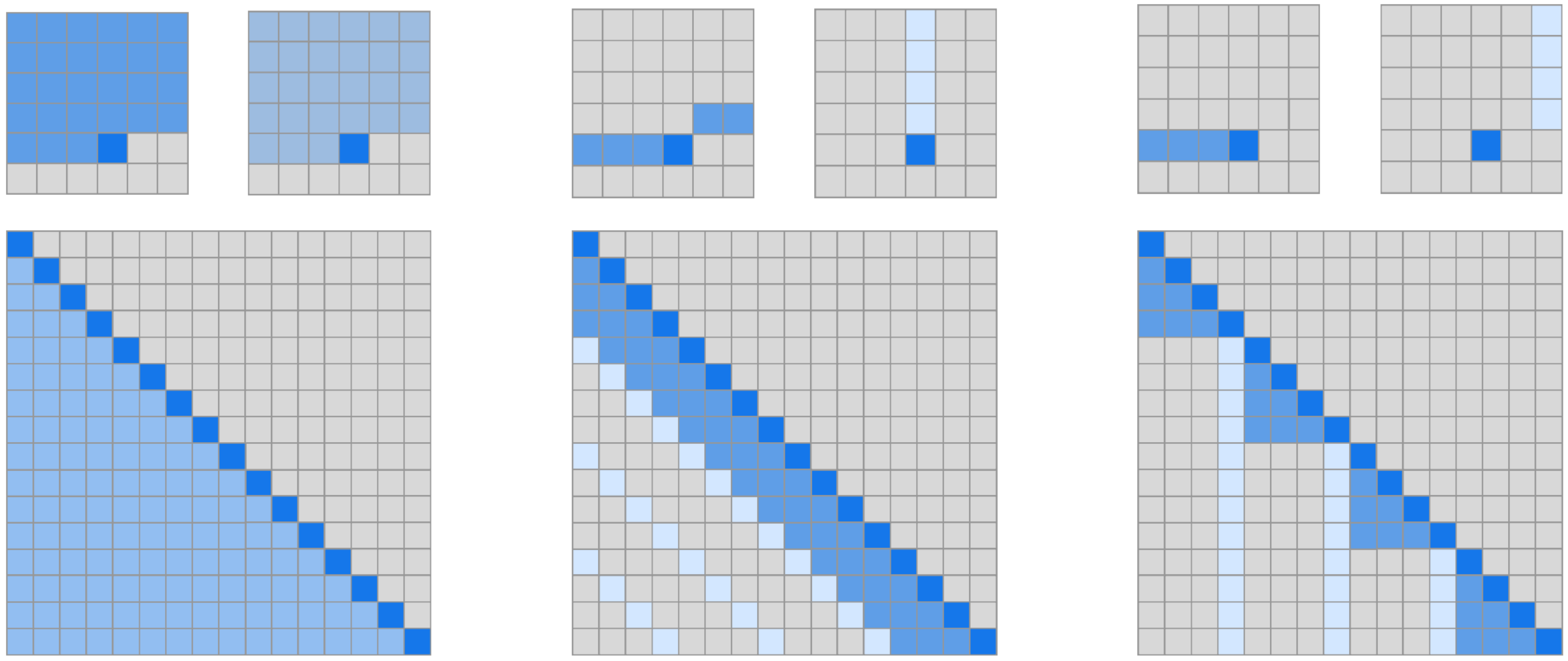
Method

- Parameters

  175B v.s. 1.5B (GPT-2)

- Architecture

  Sparse self-attention layer described in the Sparse Transformer



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

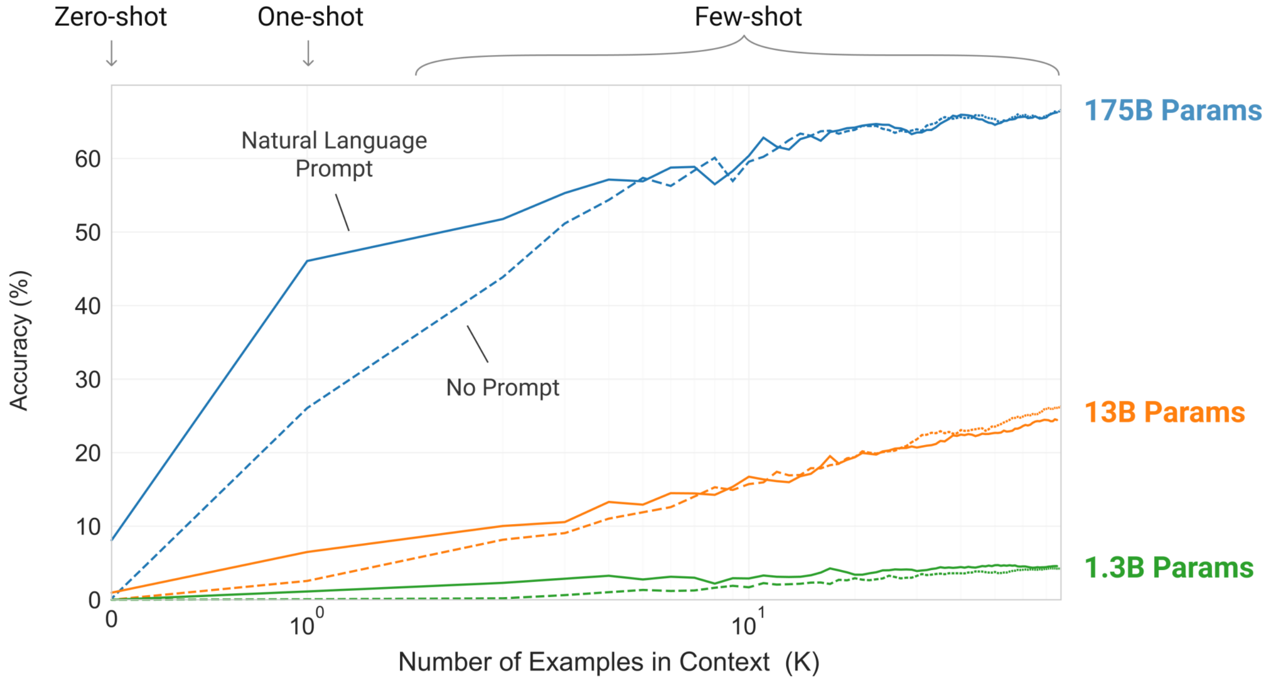
$O(N^2)$ -> $O(N\sqrt{N})$

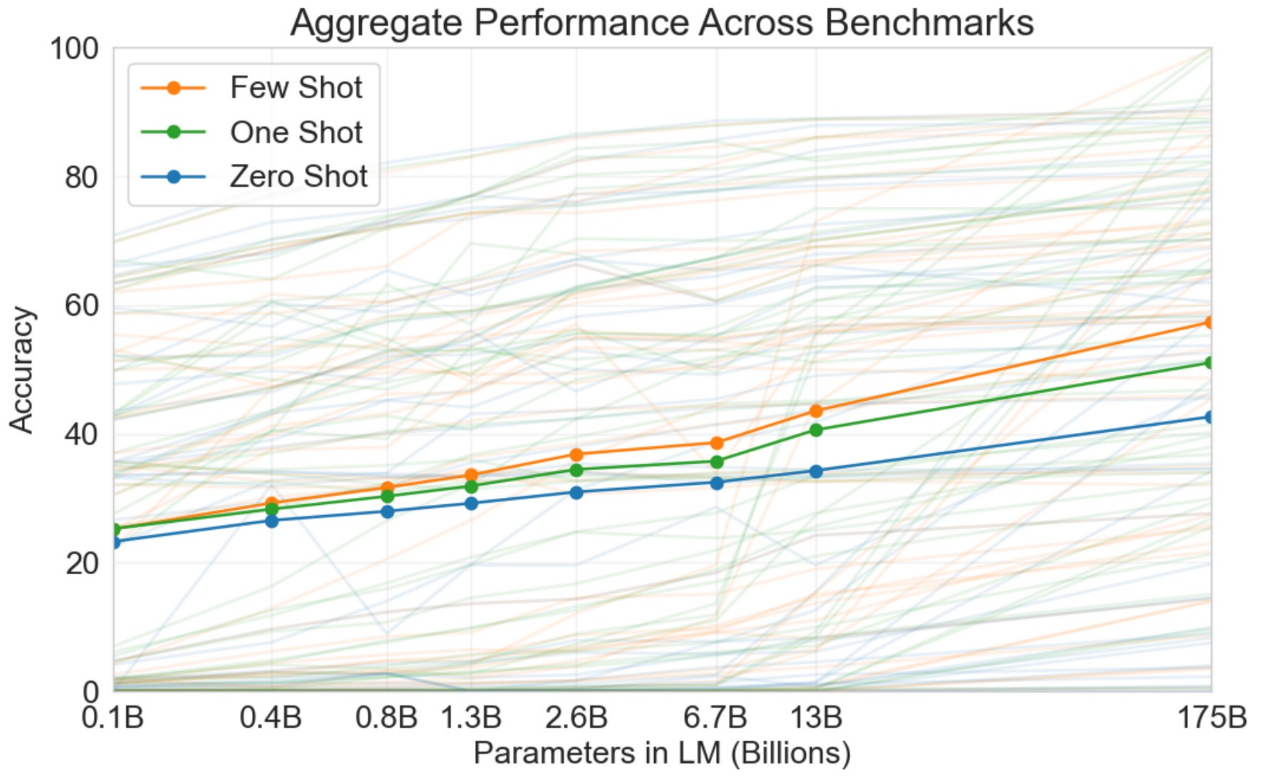
Child, Rewon, et al. Generating Long Sequences with Sparse Transformers. 2019.

Method

- The bigger, the better

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}





Method

- Dataset

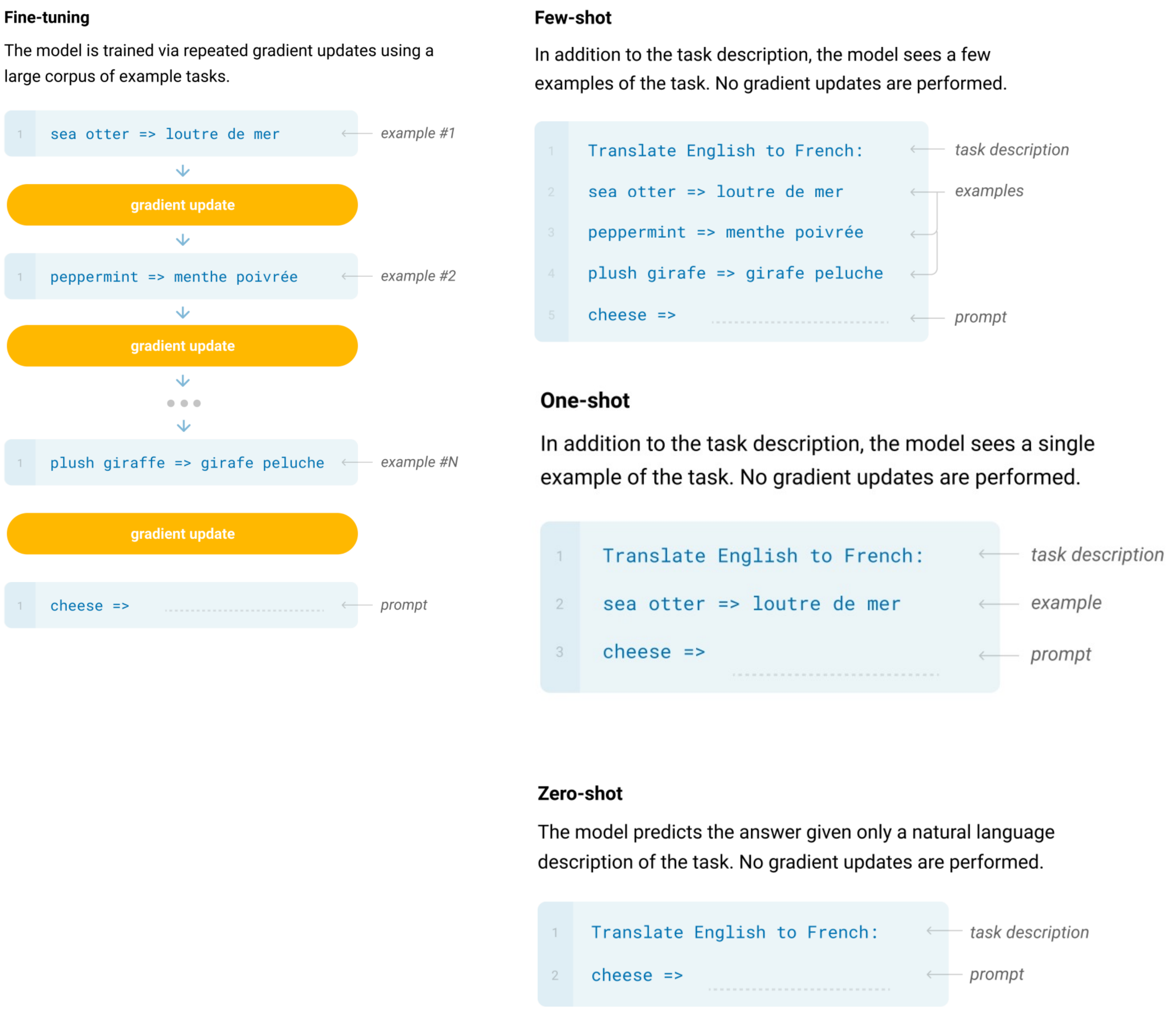
Common Crawl (filtered) + WebText2 + Books1 + Books2 + Wikipedia (45TB)

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

<- higher-quality are sampled more frequently

Experiments

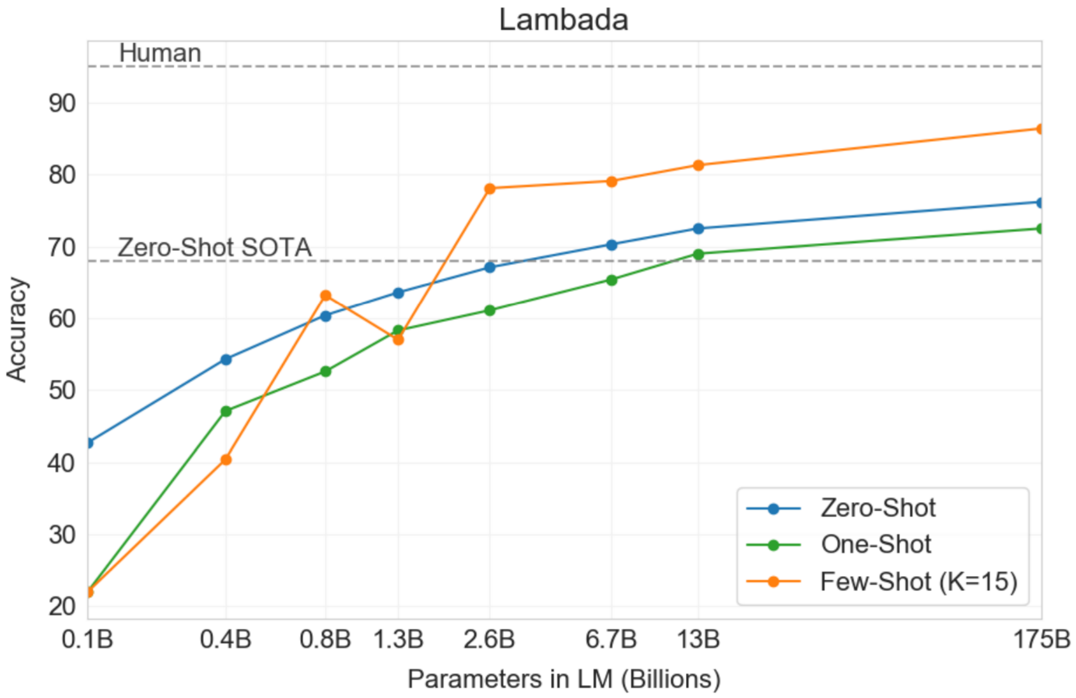
- Evaluation Settings
  - Fine-tuning
  - Few-shot
  - One-shot
  - Zero-shot
- Tasks
  - Completion
  - Question Answering
  - Translation
  - Common Sense Reasoning
  - Reading Comprehension
  - Natural Language Inference
  - …



Experiments

- Language Modeling, Cloze, and Completion Tasks
  - LAMBADA: Predict the last word of sentences
  - HellaSwag: Pick the best ending to a story or set of instructions
  - StoryCloze: Select the correct ending sentence for five-sentence long stories

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



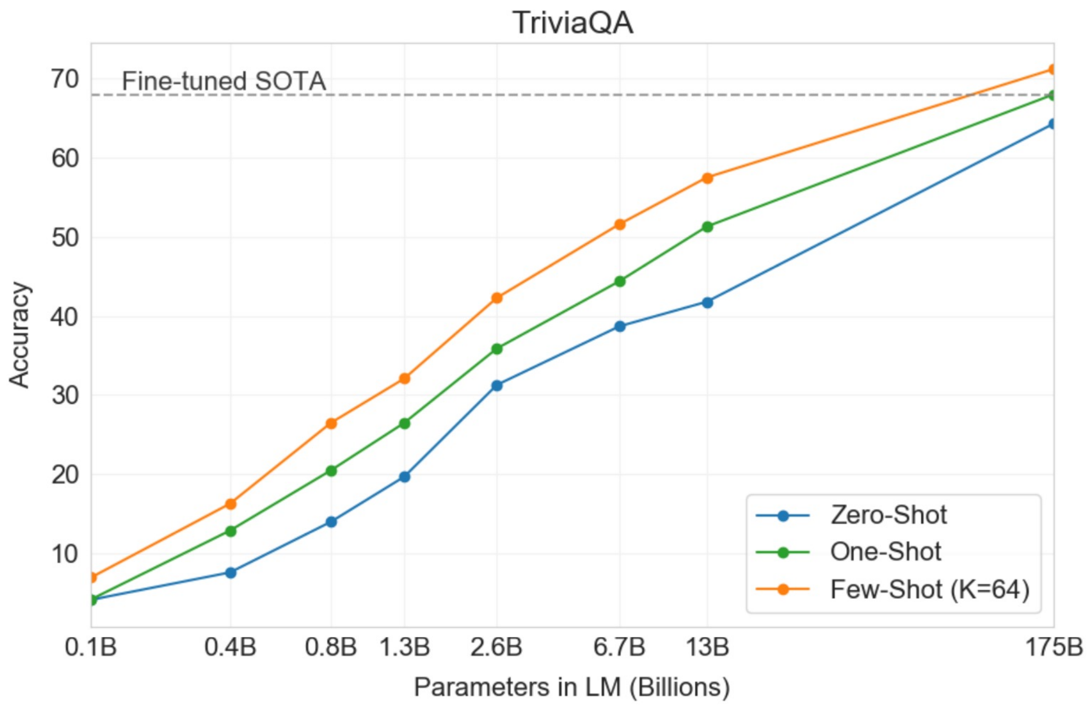
Lack reasoning ability

Paperno, Denis, et al. The lambada dataset: Word prediction requiring a broad discourse context. 2016.
Zellers, Rowan, et al. Hellaswag: Can a machine really finish your sentence?. 2019.
Mostafazadeh, Nasrin, et al. A corpus and evaluation framework for deeper understanding of commonsense stories. 2016.

Experiments

- Closed Book Question Answering
  - Compare to Open-Book: Answer the questions without conditioning on auxilliary information
  - NaturalQS: Read and comprehend an entire Wikipedia article that may or may not contain the answer to the question
  - WebQS: Answer questions based on Freebase
  - TriviaQA: Contain Question-answer-evidence triples and require more cross sentence reasoning to find answers

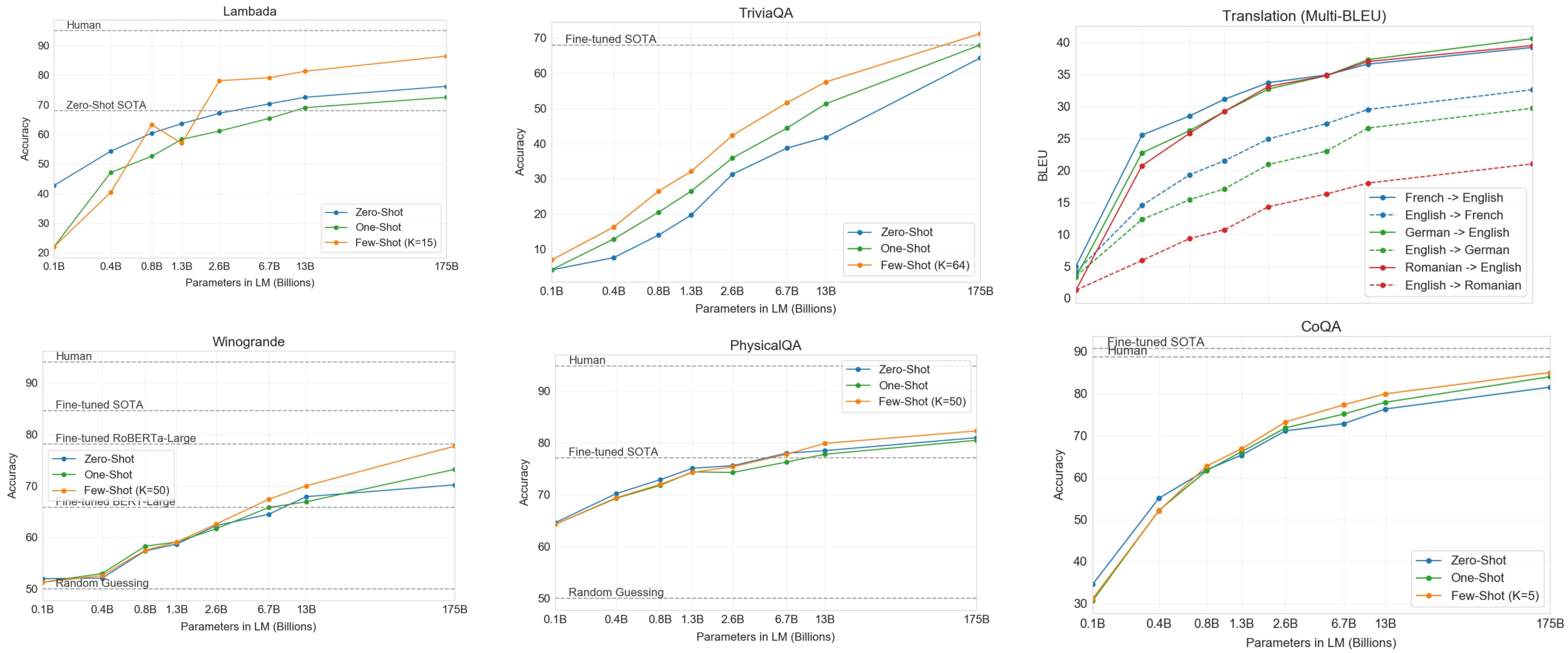
Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**



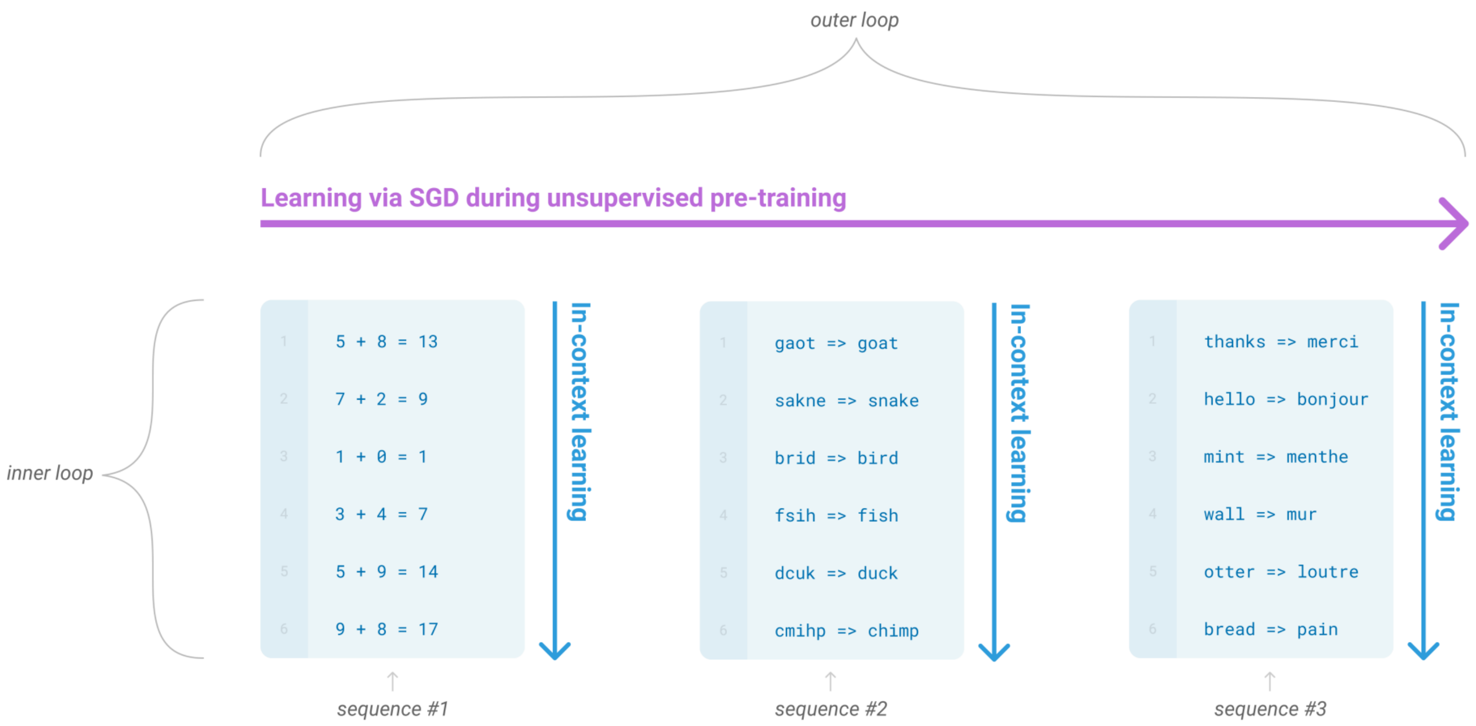
Kwiatkowski, Tom, et al. Natural questions: a benchmark for question answering research. TACL 2019.
Berant, Jonathan, et al. Semantic parsing on freebase from question-answer pairs. EMNLP 2013.
Joshi, Mandar, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. 2017.

Experiments

- Model capacity translates directly to more ‘knowledge’ absorbed in the parameters of the model



Why does GPT-3 have good generalization performance?



Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples <- inner loop
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ <- outer loop
9: **end while**

Finn, Chelsea, et al. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.

Limitations

- Weaknesses in Text Synthesis
  - Repeat
  - Lose coherence
  - Contradict
- Difficulty in Commonsense
- Weight every token equally
- Poor sample efficiency during pre-training
- Lack of explainability
- Closed source

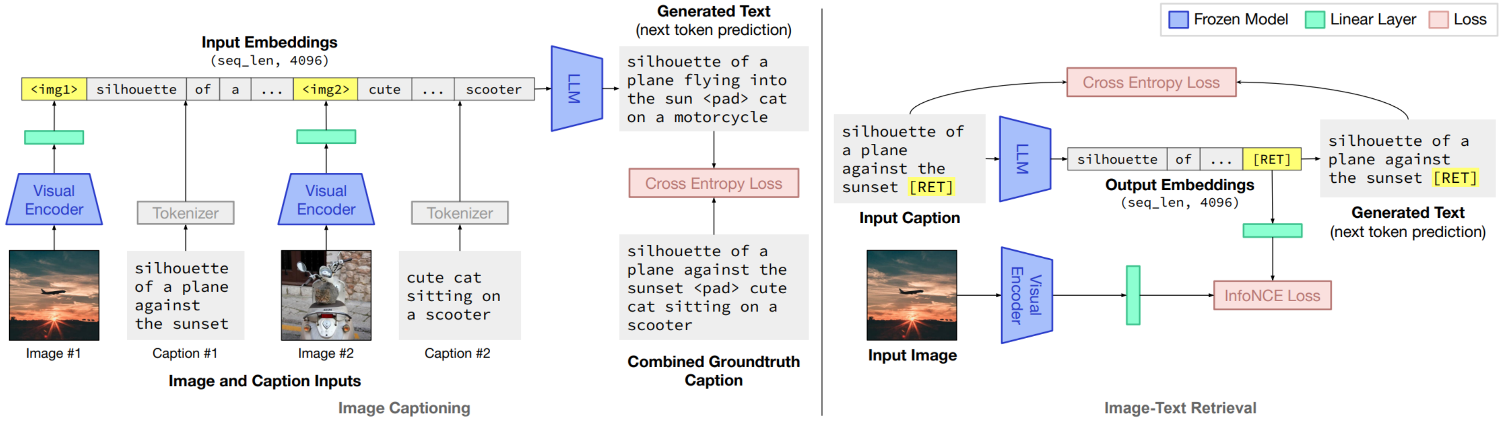
West, Peter, et al. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. NAACL 2022.
Wei, Jason, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022.

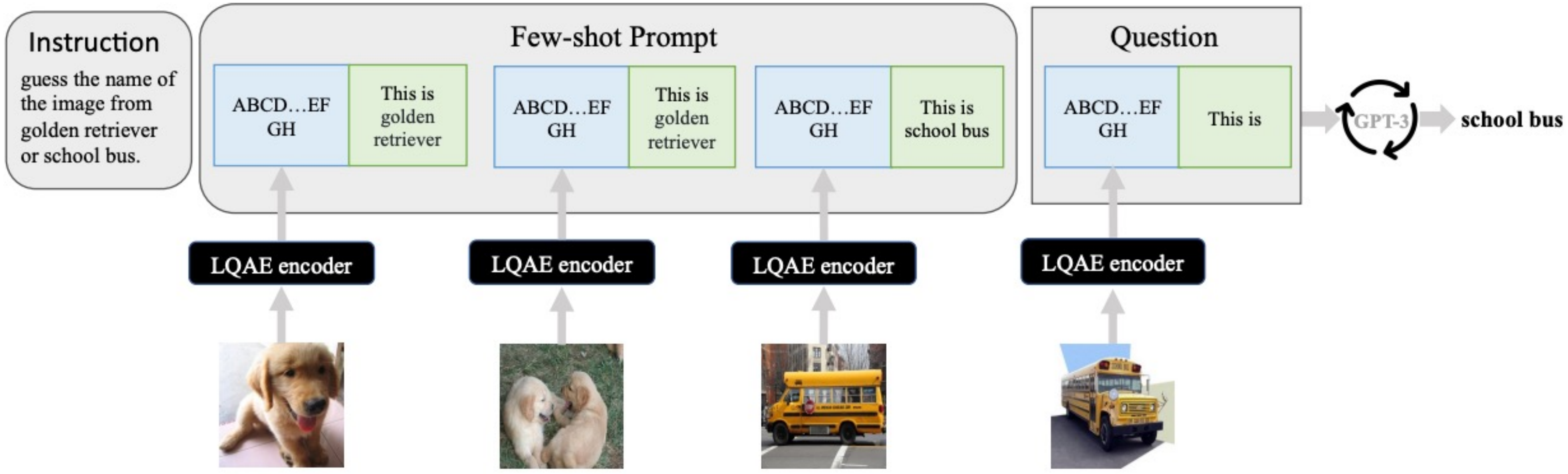
Conclusion

- 175 billion parameter language model which shows strong performance on many NLP tasks and benchmarks in the zero-shot, one-shot, and few-shot settings
- Still have some limitations and weaknesses to overcome
- More data, larger model, more ability

Future Work

- Inject knowledge into vision from LLM





Koh, Jing, et al. Grounding Language Models to Images for Multimodal Generation. 2023.
Hao, Liu, et al. Language Quantized AutoEncoders: Towards Unsupervised Text-Image Alignment. 2023.

Thanks

Shu Zhao
smz5505@psu.edu