

Inference Methods

CS 5539: Advanced Topics in Natural Language Processing

<https://shocheen.github.io/courses/advanced-nlp-fall-2024>

Logistics

- Project proposal deadline: October 1st
 - Do you have an idea for your project?
 - Tips on how to choose a project: [link](#)

Goal for today's class

How can we perform tasks using a pretrained LM **without** fine-tuning it – aka prompting / inference methods.

Part I: In context learning

Part II: Chain of thought prompting

Part I: In-Context Learning

ICL Stakeholder

AU24 CSE 5539 Presentation



GPT-3: Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et. al.



NeurIPS 2020

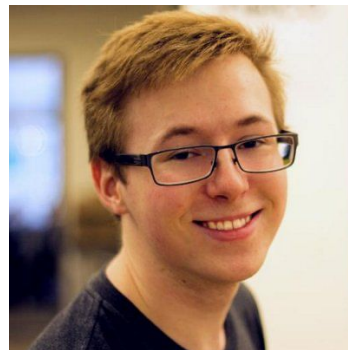
Authors



Ilya Sutskever
Co-inventor of AlexNet
Co-founder of OpenAI



Dario Amodei
Co-founder & CEO,
Anthropic



Alec Radford
ML @ OpenAI,
GPT 1, 2, 3 & 4, PPO



Aditya Ramesh
Scientist @ OpenAI
DALL·E, DALL·E 2

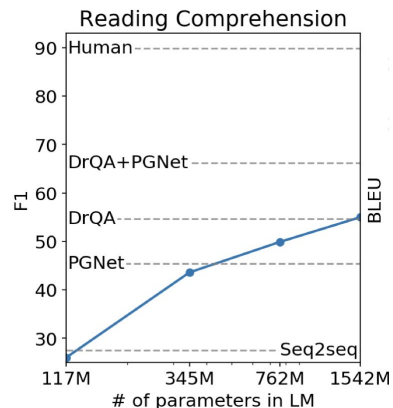
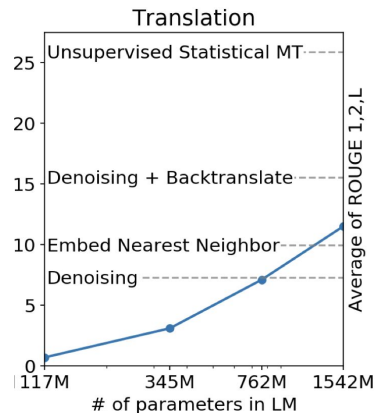
Introduction

- Previously, NLP research tended to:
 - Design task-specific model architectures.
 - Curate language representations & data to specific tasks.
- Recent paradigm shift –
 - Task-agnostic models.
 - Generalized pre-training & architectures.
- Final step (?) –
 - Adapting these task-agnostic models to specific tasks.



How necessary is finetuning?

- Prior work shows:
 - A single pre-trained model has good zero-shot performance. Not SoTA...yet.
 - Performance scales with parameter count* (!)
- Contributions of this work:
 - Empirically test performance scaling, ranging up to **175B** parameters (GPT-3.)
 - Clarify and systematize “in-context learning.”
 - **Promising** experimental results.



Approach

- **Fine-tuning: update weights** based on data.
 - + Good benchmark performance.
 - Poor OOD generalization.
- **Few-shot: task description** along with *K examples* of samples/completions.
 - + Major reduction in task-specific data.
 - Worse performance than SoTA (*so far.*)
- **One-shot: few-shot** with $K=1$.
- **Zero-shot: Task description** only, $K=0$.

Learning Settings

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.

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

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
```

Few-shot

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

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

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

Fine-tuning

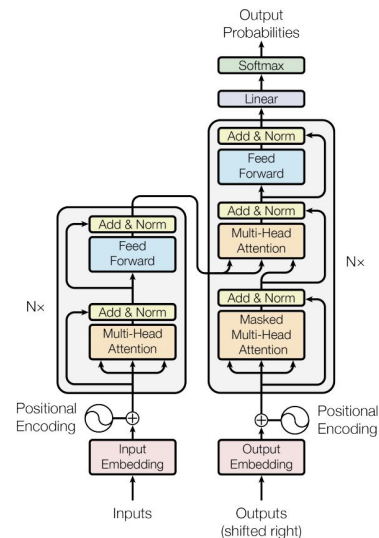
The model is trained via repeated gradient updates using a large corpus of example tasks.



Approach

- **Architecture:**
 - Identical to GPT-2, except for the transformer attention pattern.
 - 8 different model sizes – 125M to 175B
 - Model & data partitioned across GPUs to efficiently handle memory constraints
- **Training Dataset:**
 - Filtered CommonCrawl
 - Deduplication to prevent redundancy & ensure integrity of held-out validation set.
 - Augmented with reference corpora: WebText, Books1 & 2, English Wikipedia.

Model & Dataset



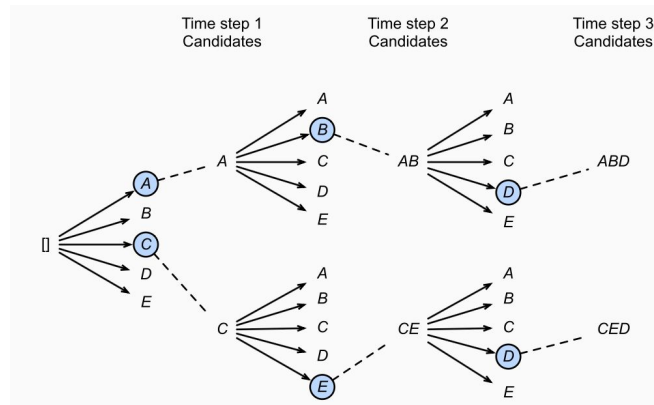
Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Approach

- **Training Process:**
 - Model parallelism both within each matrix multiply & across layers.
- **Evaluation:**
 - One/Few-shot: draw K samples from training or dev set as conditioning.
 - Some tasks – additional natural language prompt.
 - Results reported on test set when possible.

Training & Evaluation

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}

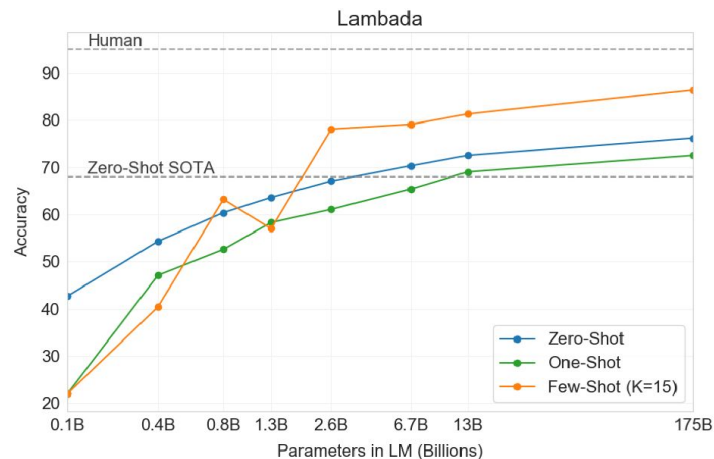


Results

Language Modeling & Cloze

- Penn Treebank:
 - New SoTA by 15 points.
 - Zero-shot perplexity of 20.5 on POS labeling.
- LAMBADA:
 - Predicting terminal word in a sentence/paragraph.
 - Framed in a few-shot setting – 86.4% (+18%).
 - One-shot – not as effective.
- HellaSwag & StoryCloze – lower than fine-tuned SoTA.

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



Results

- Closed-book (no document/info access)
 - GPT-3 nears or exceeds SoTA pre-trained and fine-tuned RAG models on 2 datasets.
 - ARC multiple choice – approaches baselines; much worse than SoTA.
- Reading comprehension – approach human baselines but worse than SoTA NNs.
- Translation:
 - Underperforms SoTA on 0-shot.
 - Few-shot – approaches SoTA when translating to En.

QA & Translation

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

Setting	ARC (Easy)	ARC (Challenge)	CoQA	DROP
Fine-tuned SOTA	92.0^a	78.5^b	90.7^c	89.1^d
GPT-3 Zero-Shot	68.8	51.4	81.5	23.6
GPT-3 One-Shot	71.2	53.2	84.0	34.3
GPT-3 Few-Shot	70.1	51.5	85.0	36.5

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]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

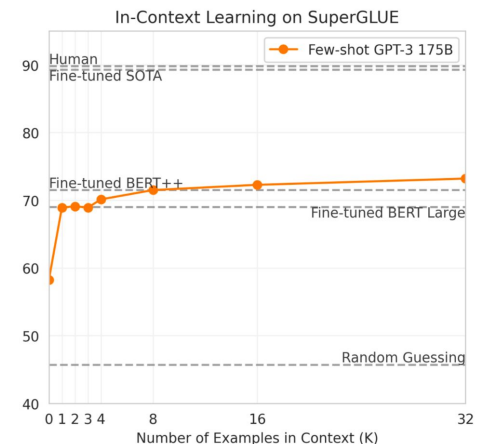
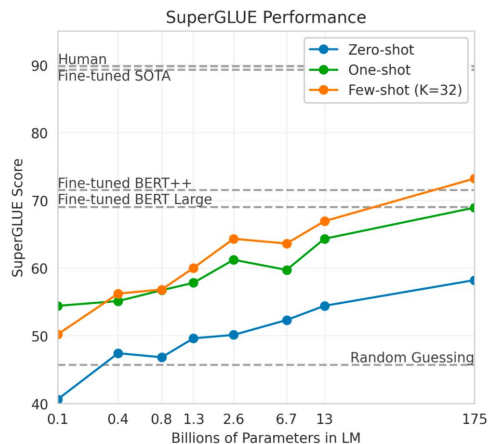
Results

- A standardized collection of datasets.
- Few-shot results –
 - Steady improvement through K=32.
 - Large variance in GPT-3 performance.
 - Weak at comparing sentences
- Scaling shows improvements

SuperGLUE

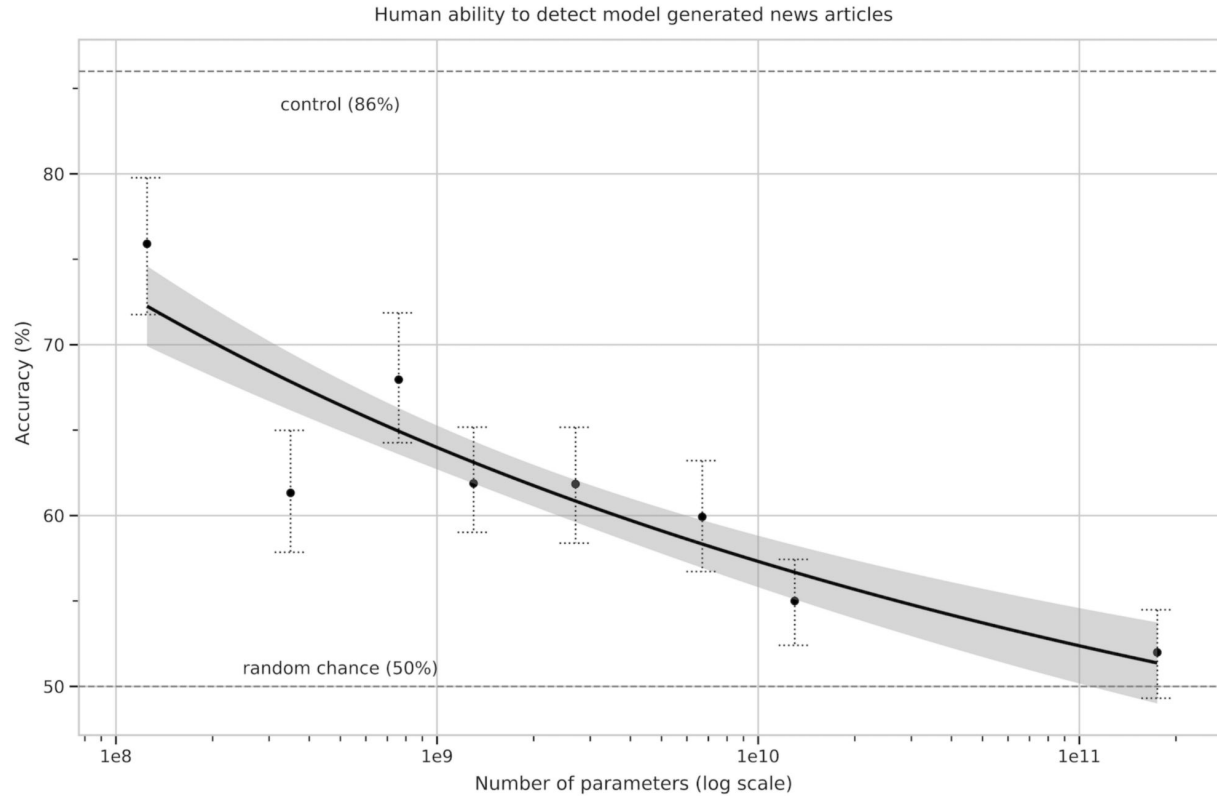
	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	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

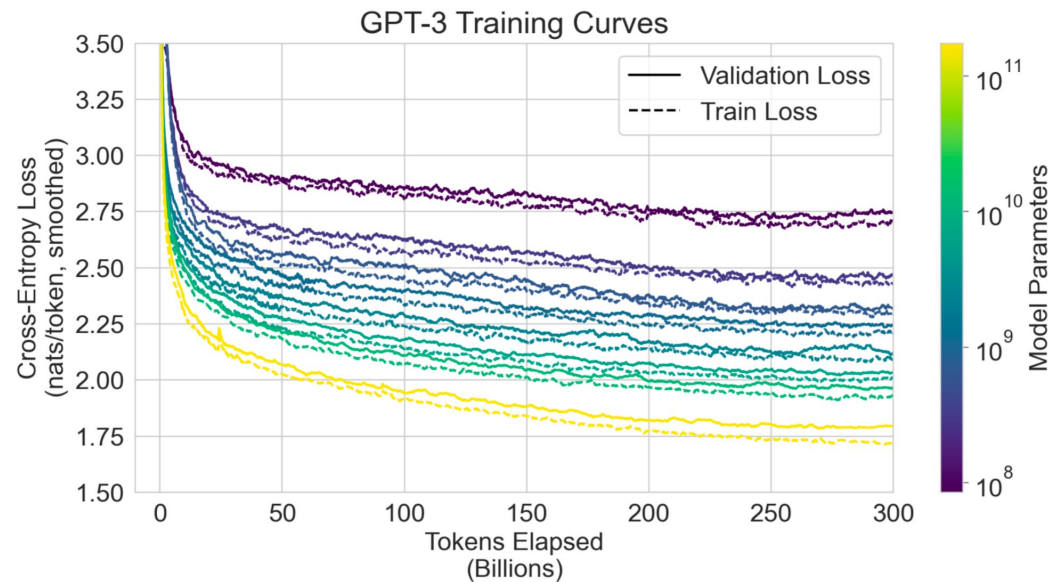


Results

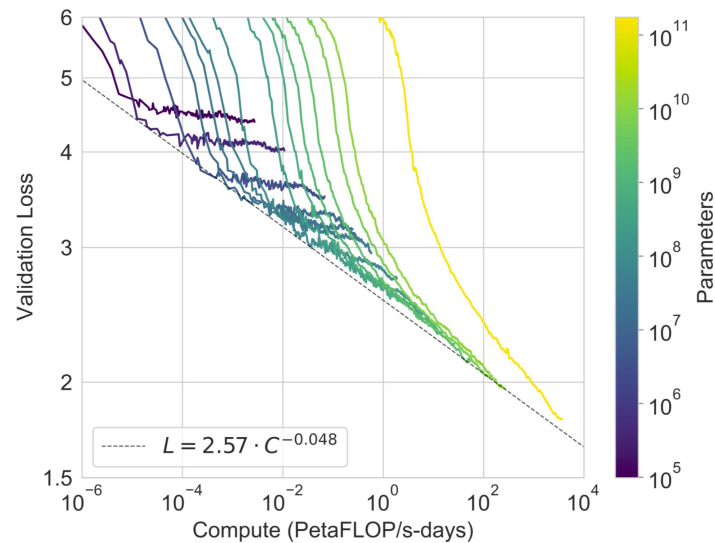
Misc.



Results

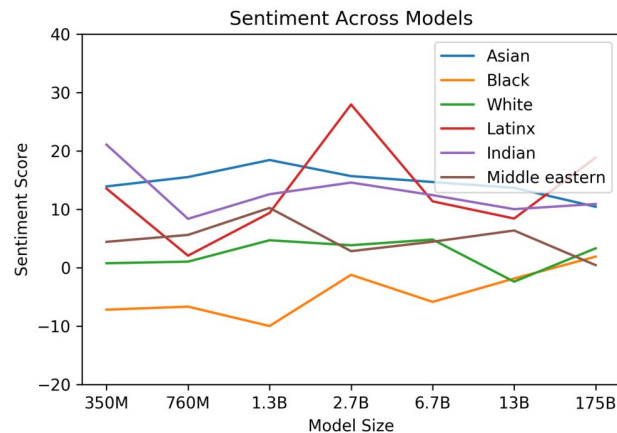
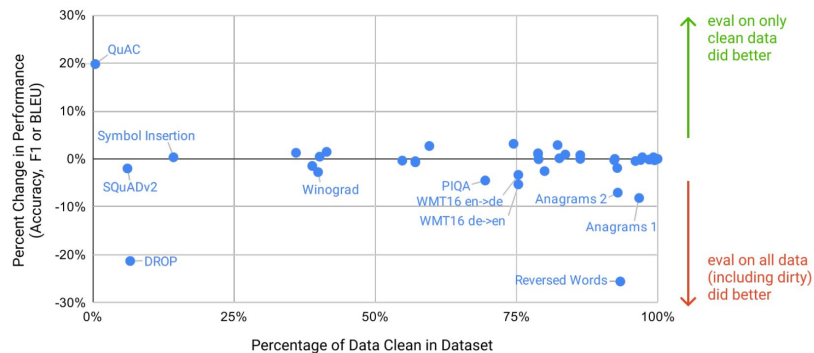


Misc.



Limitations & Conclusion

- Potential *test set contamination* from the internet-scale dataset.
- Model limitations:
 - Semantic self-repetition.
 - Weakness at “common-sense” and comparative tasks.
 - Lack of interpretability.
 - Poor sample efficiency.
 - What does ICL actually do?
- 175B model; towards general language systems; empirical scaling results; ethical considerations.



Thank you!

Questions?



GPT-3: Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et. al.

Paper: arxiv.org/abs/2005.14165

ICL Reviewer

Key Summary of Contributions:

- GPT-3 demonstrates Meta-learning capabilities with its ability to perform “In-Context” Learning (ICL).
- This particularly scales as model size increases, ICL Capabilities are better on a wide range of natural language processing tasks

Strengths

- Demonstrates the scaling effect, where GPT-3's large size significantly improves few-shot learning performance, often rivaling state-of-the-art fine-tuned models.
- Introduces a reproducible approach for task-agnostic learning, enabling large-scale language models to adapt to multiple tasks without fine-tuning (updating gradients).
- Significant advancement in meta-learning and natural language processing capabilities.

Weaknesses

- Tasks that have long corpus dependence tend to fall short of several NLP tasks
- While GPT-3 appears to show impressive results against SOTA models on those benchmark tasks with no gradient updates, however it does not beat the SOTA in several NLP tasks
- Concerns about data leakage when running the benchmarks

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	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

Table 3.5: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and **SOTA**. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

Follow-up Questions for Authors

1. How have you checked for data leakage on your benchmark data?
2. What strategies could be employed to address this issue, especially for applications requiring sustained coherence over longer outputs? How does this impact performance?

ICL Archaeologist

Main Motivation for GPT-3

Prior work: the architecture and the initial representations are task-agnostic but still require a task-specific step of fine-tuning.

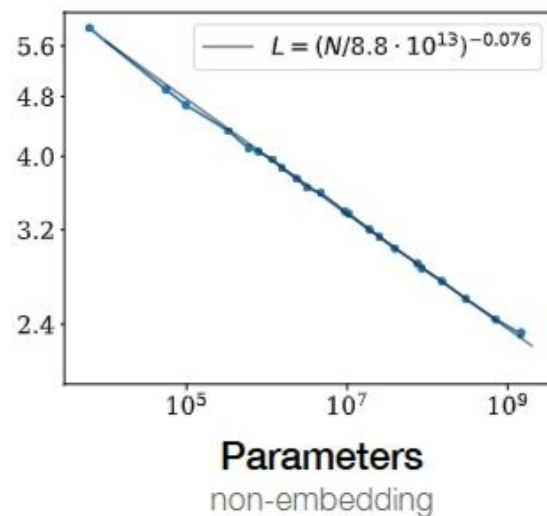
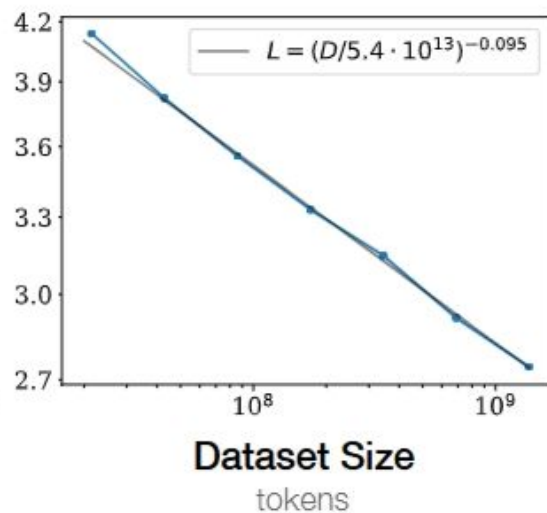
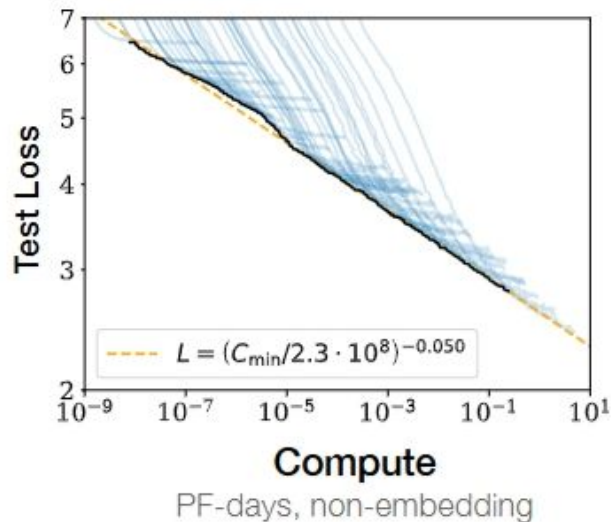
GPT3: How can we get rid of this

Additional Context for In-Context Learning

- What inspired this paper?
 - Primarily GPT2 – which showed proof of concept of zero-shot inference.
 - Scaling Laws – will go into details on October 7
 - Meta-Learning: Learning to learn

- What did this paper inspire?
 - Is scaling required for in-context learning?
 - Are models “learning” in-context?
 - Why can models learn in-context?
 - Can we teach models to better learn in-context

Scaling Laws of Language Models



Meta Learning

-

Learn To Learn Task



Quikly learn
New Task

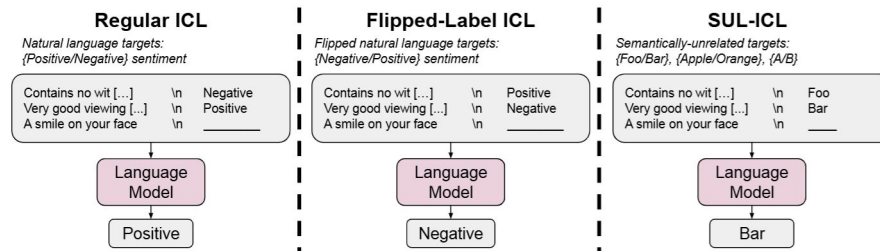
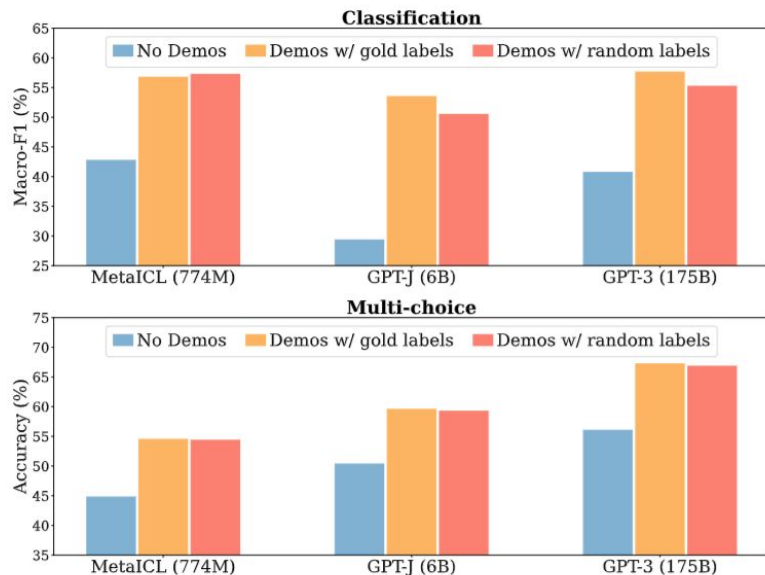


Additional Context for In-Context Learning

- What inspired this paper?
 - Primarily GPT2 – which showed proof of concept of zero-shot inference.
 - Scaling Laws – will go into details on October 7
 - Meta-Learning: Learning to learn

- **What did this paper inspire?**
 - **Is scaling required for in-context learning?**
 - **Are models “learning” in-context?**
 - **Why can models learn in-context?**
 - **Can we teach models to better learn in-context, instruction tuning and more**

Are models “learning” from in-context examples?



[Wei et al 2022] Larger language models do in-context learning differently

[Min et al 2021] Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Is scaling required for in-context learning?

It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners

Timo Schick^{1,2} and Hinrich Schütze¹

¹ Center for Information and Language Processing, LMU Munich, Germany

² Sulzer GmbH, Munich, Germany

timo.schick@sulzer.de

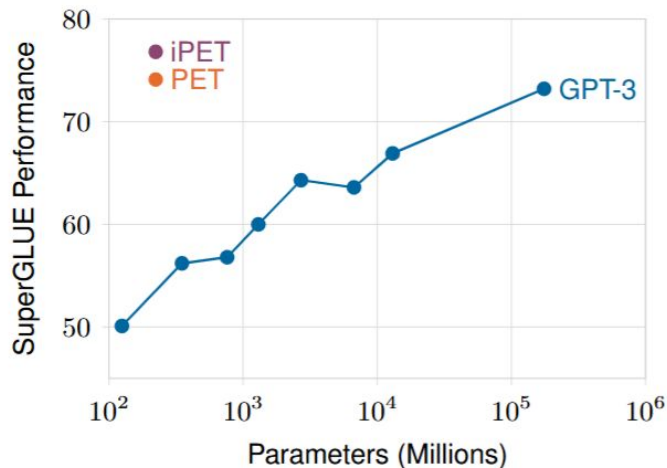


Figure 1: Performance on SuperGLUE with 32 training examples. **ALBERT with PET/iPET outperforms GPT-3 although it is much “greener” in that it has three orders of magnitude fewer parameters.**

Why can models learn in-context?

An Explanation of In-context Learning as Implicit Bayesian Inference

Sang Michael Xie
Stanford University
xie@cs.stanford.edu

Aditi Raghunathan
Stanford University
aditir@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

Tengyu Ma
Stanford University
tengyuma@cs.stanford.edu

In-context Learning and Induction Heads

AUTHORS

Catherine Olsson*, Nelson Elhage*, Neel Nanda*, Nicholas Joseph*, Nova DasSarma*, Tom Henighan*, Ben Mann*, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah*

* Core Research Contributor; * Core Infrastructure Contributor; * Correspondence to colah@anthropic.com; Author contributions statement below.

AFFILIATION

Anthropic

PUBLISHED

Mar 8, 2022

Transformers Learn In-Context by Gradient Descent

Johannes von Oswald^{1,2} Eyvind Niklasson² Ettore Randazzo² João Sacramento¹
Alexander Mordvintsev² Andrey Zhmoginov² Max Vladymyrov²

FUNCTION VECTORS IN LARGE LANGUAGE MODELS

Eric Todd*, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller,
Byron C. Wallace, and David Bau
Khoury College of Computer Sciences, Northeastern University

In-Context Learning Creates Task Vectors

Roe Hendel
Tel Aviv University
roee.hendel@mail.tau.ac.il

Mor Geva
Google DeepMind
pipek@google.com

Amir Globerson
Tel Aviv University, Google
gamir@tauex.tau.ac.il



Can we teach models to learn in-context

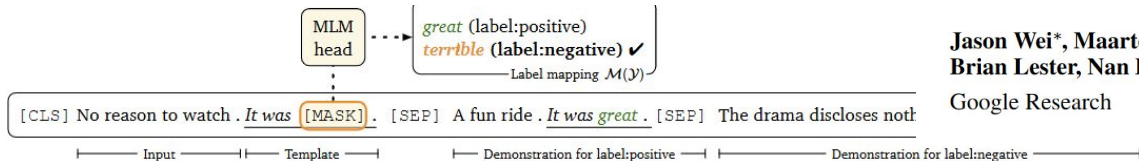
Making Pre-trained Language Models Better Few-shot Learners

Tianyu Gao^{†*} Adam Fisch^{‡*} Danqi Chen[†]

[†]Princeton University [‡]Massachusetts Institute of Technology

{tianyug, danqic}@cs.princeton.edu

fisch@csail.mit.edu



(c) Prompt-based fine-tuning with demonstrations (our approach)

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

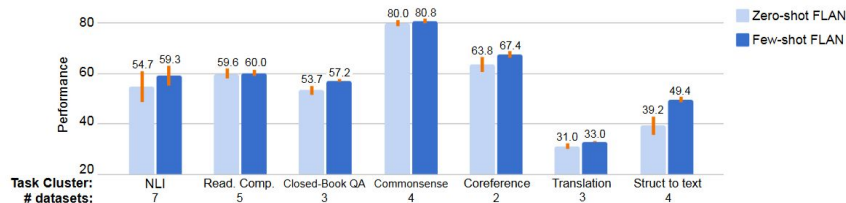


Figure 9: Adding few-shot exemplars to FLAN is a complementary method for improving the performance of instruction-tuned models. The orange bars indicate standard deviation among templates, averaged at the dataset level for each task cluster.

In-context Learning Visionary

Teach LLMs to Use Searching Engine

Use searching results as context for LLMs to generate better results

Train LLMs to perform searching using RL

A lot of on-going research on this field...

Scale Up for Better Fundamental Models

Based on scaling law, larger model size and larger dataset size trains a model with lower loss.

Train larger LLMs on larger dataset

A lot of on-going research on this field...

Benchmarking LLMs of In-context Learning

To help practitioners find more suitable LLMs for their specific need (or to train the next-generation fundamental LLMs) , we need to evaluate the state-of-the-art LLMs on different topics involving different kinds of in-context learning tasks

A lot of on-going research on this field...

Construct Specialized Fundamental Few-shot LLMs

Motivation: few-shot learning is important for application tasks with very limited training data, such as project-specific code comment generation, personalized handwriting recognition

Limitations: existing LLMs are not trained to be focused on few-shot learning, resulting in data gap between training and inference for these applications

Insights: fine-tuning fundamental LLMs on few-shot learning dataset to mitigate this gap

Haven't heard of existing research on this field...

Part II: Chain of Thought Prompting

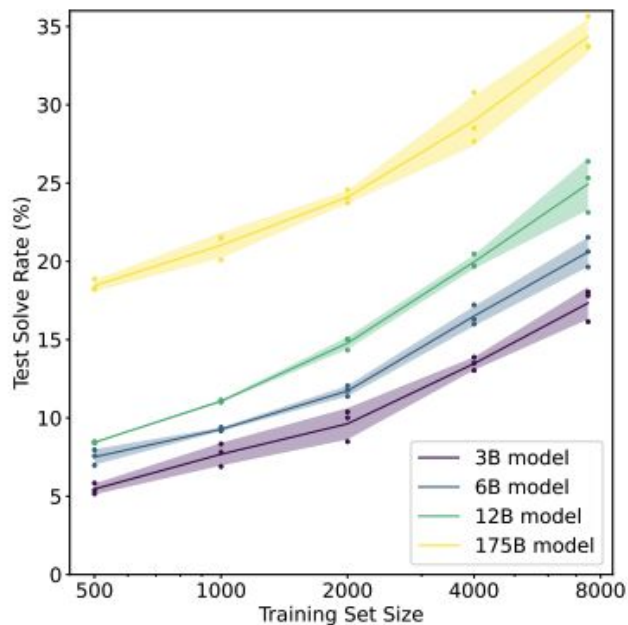
CoT - Stakeholder

Challenges in LLMs

- Scaling up model size alone has not proved sufficient for achieving high performance on challenging tasks, such as arithmetic, commonsense, and symbolic reasoning.
- Large language models still have limitations in their ability to reason and understand the context of a situation.

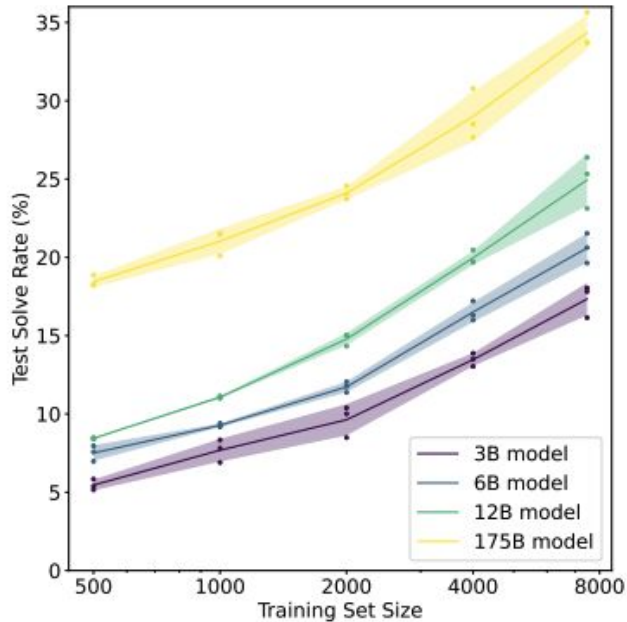
Reasoning Problems

Fine-tune GPT-3 on GSM8K (arithmetic)
(Cobbe et al., 2021):



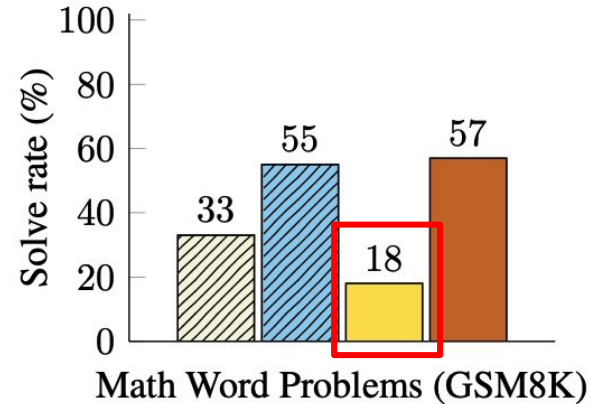
Reasoning Problems

Fine-tune GPT-3 on GSM8K (arithmetic)
(Cobbe et al., 2021):



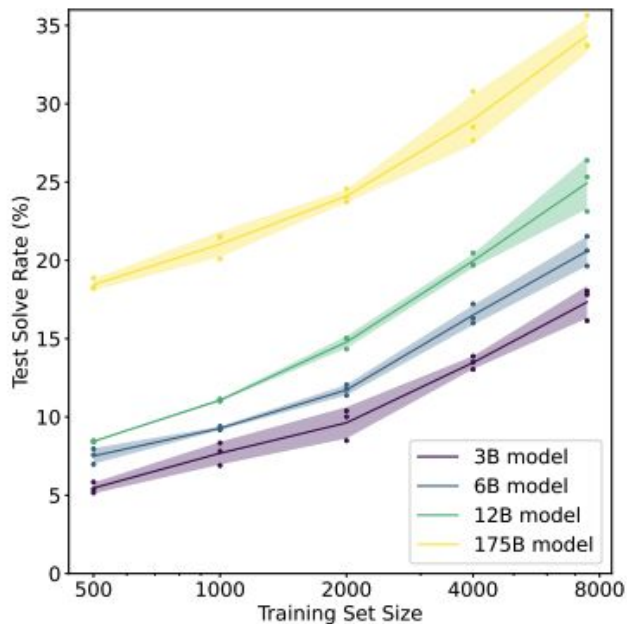
GSM8K (arithmetic):

- Finetuned GPT-3 175B
- Prior best
- PaLM 540B: standard prompting
- PaLM 540B: chain-of-thought prompting



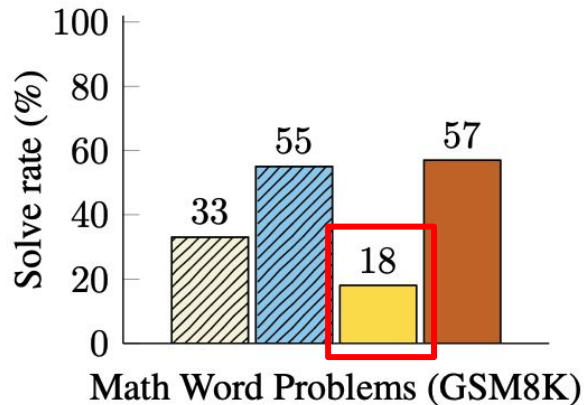
Reasoning Problems

Fine-tune GPT-3 on GSM8K (arithmetic)
(Cobbe et al., 2021):



GSM8K (arithmetic):

- Finetuned GPT-3 175B
- Prior best
- PaLM 540B: standard prompting
- PaLM 540B: chain-of-thought prompting



! **Few-shot standard prompting** with even larger model (PaLM 540B) also does not work well.

Contribution

- This work explores the ability of language models to perform **few-shot** prompting for **reasoning tasks**, given a prompt that consists of triplets: $\langle \text{input, chain of thought, output} \rangle$
 - Chain-of-thought: a **series of intermediate natural language reasoning steps** that lead to the final output (Chain-of-thought prompting)

Contribution

- This work explores the ability of language models to perform **few-shot** prompting for **reasoning tasks**, given a prompt that consists of triplets: $\langle \text{input, chain of thought, output} \rangle$
 - Chain-of-thought: a **series of intermediate natural language reasoning steps** that lead to the final output (Chain-of-thought prompting)
- This work presents empirical evaluations on **arithmetic**, **commonsense**, and **symbolic reasoning** benchmarks, showing that chain-of-thought prompting outperforms standard prompting.

Contribution

- This work explores the ability of language models to perform **few-shot** prompting for **reasoning tasks**, given a prompt that consists of triplets: $\langle \text{input, chain of thought, output} \rangle$
 - Chain-of-thought: a **series of intermediate natural language reasoning steps** that lead to the final output (Chain-of-thought prompting)
 - This work presents empirical evaluations on **arithmetic**, **commonsense**, and **symbolic reasoning** benchmarks, showing that chain-of-thought prompting outperforms standard prompting.
- ! No language models were finetuned in the process of writing this paper.

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Properties of CoT

1. **Decomposes Complex Problems:** CoT allows models to break down multi-step problems into intermediate steps, improving reasoning for more complex tasks.

Properties of CoT

1. **Decomposes Complex Problems:** CoT allows models to break down multi-step problems into intermediate steps, improving reasoning for more complex tasks.
2. **Improves Interpretability:** CoT offers a clearer view into the model's **thought process**, helping to debug where reasoning errors occur.

Properties of CoT

1. **Decomposes Complex Problems:** CoT allows models to break down multi-step problems into intermediate steps, improving reasoning for more complex tasks.
2. **Improves Interpretability:** CoT offers a clearer view into the model's **thought process**, helping to debug where reasoning errors occur.
3. **Broad Applicability:** CoT works across diverse tasks like math problems, commonsense reasoning, and symbolic manipulation.

Properties of CoT

1. **Decomposes Complex Problems:** CoT allows models to break down multi-step problems into intermediate steps, improving reasoning for more complex tasks.
2. **Improves Interpretability:** CoT offers a clearer view into the model's **thought process**, helping to debug where reasoning errors occur.
3. **Broad Applicability:** CoT works across diverse tasks like math problems, commonsense reasoning, and symbolic manipulation.
4. **Easy to Implement:** CoT can be elicited in large pre-trained models by simply adding CoT examples in few-shot prompts.

Arithmetic Reasoning - Experimental Setup

Models:

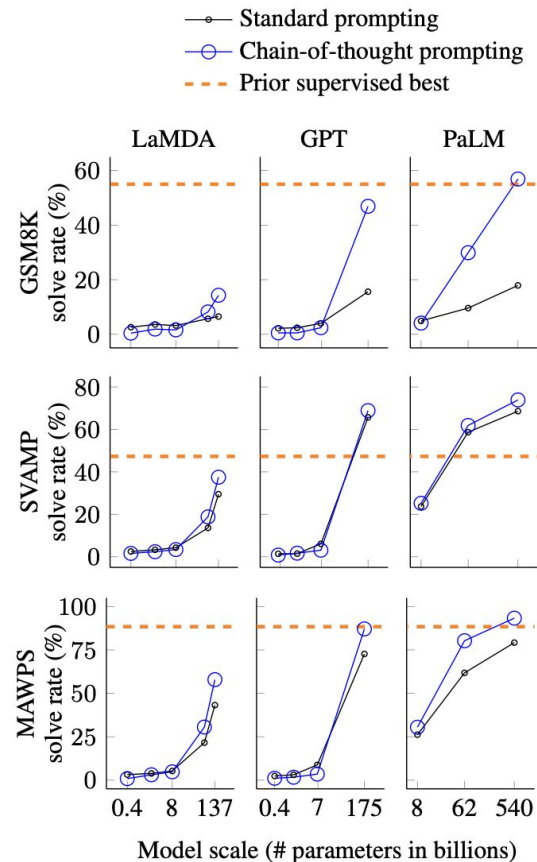
- GPT-3 (350M, 1.3B, 6.7B, 175B) (Brown et al., 2020)
- LaMDA (422M, 2B, 8B, 68B, 137B) (Thoppilan et al., 2022)
- PaLM (8B, 62B, 540B)
- UL2 20B (Tay et al., 2022)
- Codex (Chen et al., 2021)

Benchmarks:

- GSM8K (Cobbe et al., 2021)
- SVAMP (Patel et al., 2021)
- ASDiv (Miao et al., 2021)
- AQuA
- MAWPS (Koncel-Kedziorski et al., 2016)

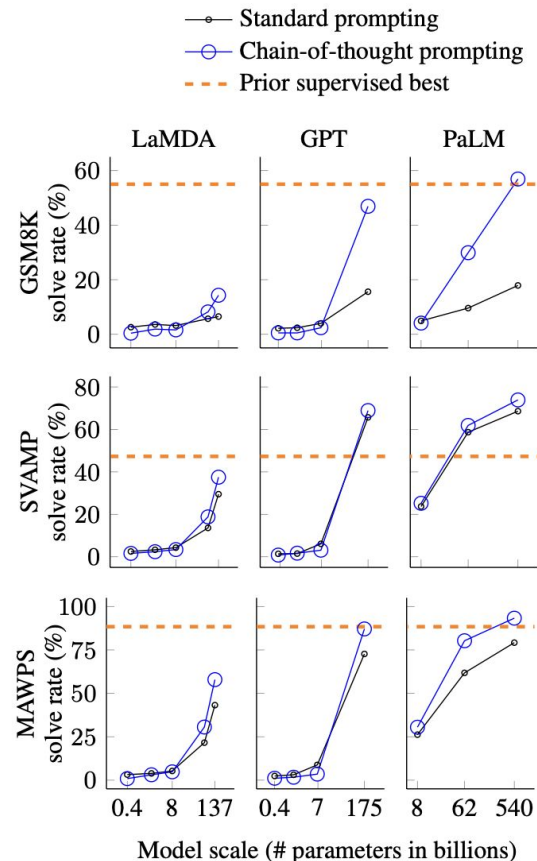
Results & Takeaways

- **Emergent Ability at Scale:** Chain-of-thought prompting only improves performance for large models (around 100B parameters)



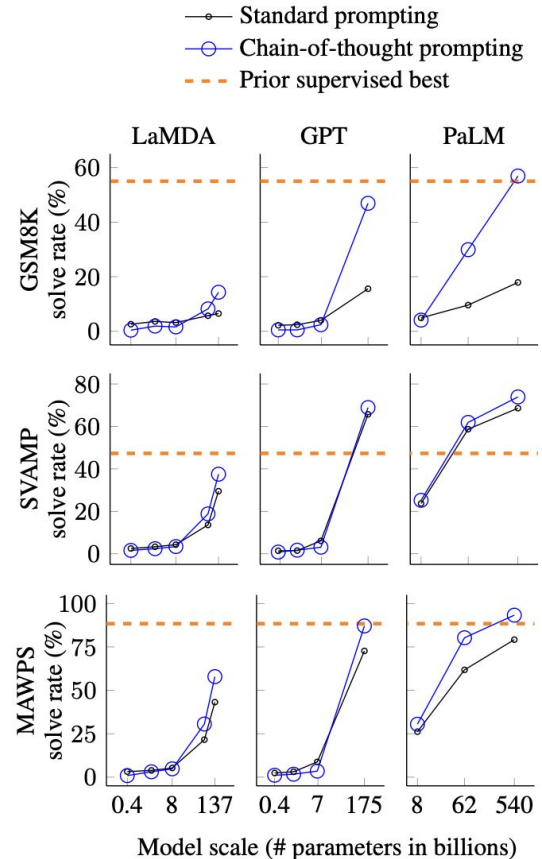
Results & Takeaways

- **Emergent Ability at Scale:** Chain-of-thought prompting only improves performance for large models (around 100B parameters)
- **Significant Gains for Complex Tasks:** CoT prompting leads to substantial performance improvements, particularly for complex tasks like GSM8K, where performance more than doubled for the largest models



Results & Takeaways

- **Emergent Ability at Scale:** Chain-of-thought prompting only improves performance for large models (around 100B parameters)
- **Significant Gains for Complex Tasks:** CoT prompting leads to substantial performance improvements, particularly for complex tasks like GSM8K, where performance more than doubled for the largest models
- **State-of-the-Art Results:** CoT prompting achieves or surpasses state-of-the-art performance and compares favorably to fine-tuned task-specific models, even without additional training.



Ablation Study

Question: Can other prompting methods match the performance gains of chain-of-thought prompting?

Ablation Study

Question: Can other prompting methods match the performance gains of chain-of-thought prompting?

Three variations of chain-of-thought:

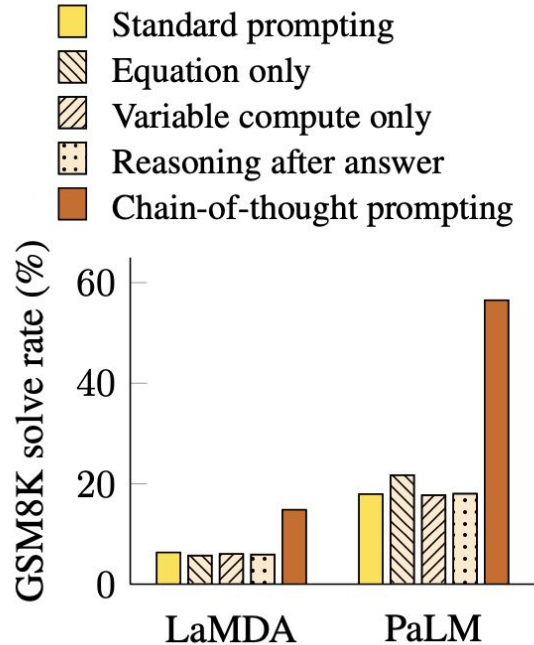
- Equation only
- Variable compute only
- Chain-of-thought after answer

Ablation Study

Question: Can other prompting methods match the performance gains of chain-of-thought prompting?

Three variations of chain-of-thought:

- Equation only
- Variable compute only
- Chain-of-thought after answer



Robustness of Chain-of-Thought

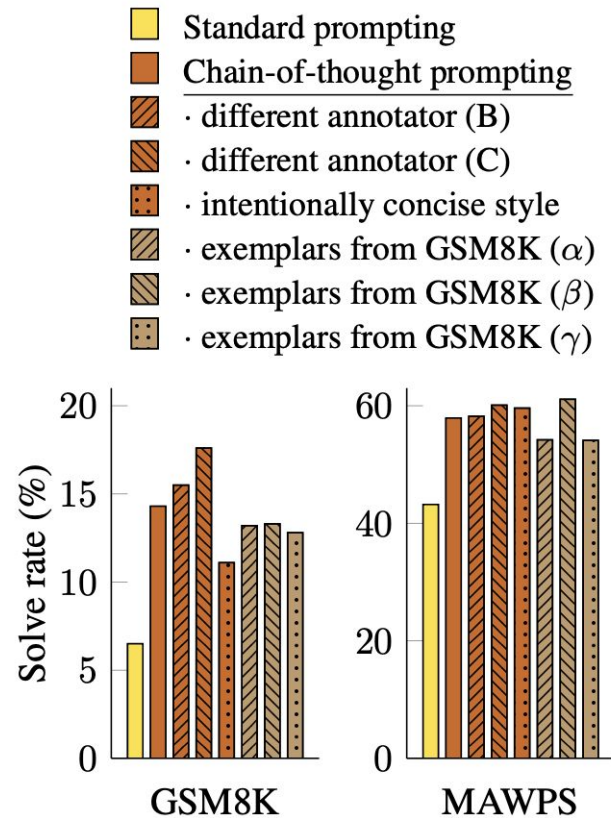
Chain-of-thought for arithmetic reasoning is robust to:

- Annotators
- Independently-written chain-of-thought
- Different exemplars
- Different exemplar orders
- Various language models
- Varying number of exemplars

Robustness of Chain-of-Thought

Chain-of-thought for arithmetic reasoning is robust to:

- Annotators
- Independently-written chain-of-thought
- Different exemplars
- Different exemplar orders
- Various language models
- Varying number of exemplars



Commonsense Reasoning

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

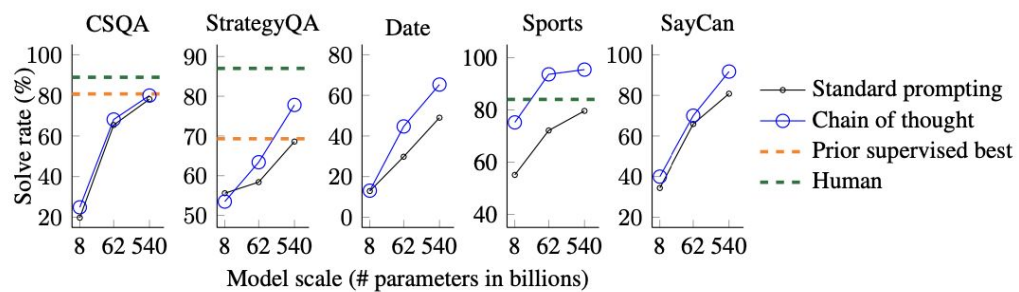
A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Results



Results

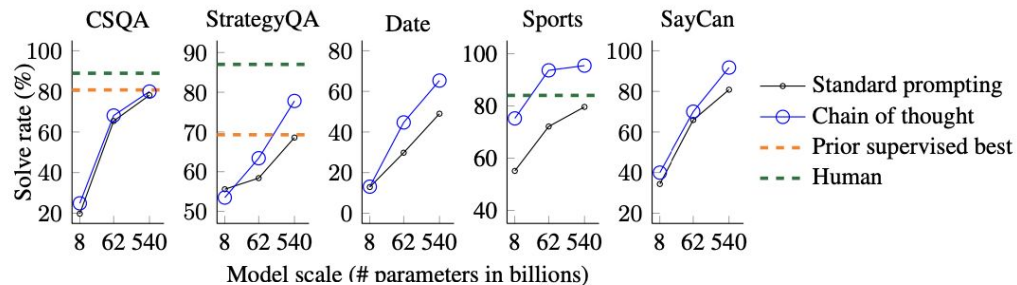


Table 4: Standard prompting versus chain of thought prompting on five commonsense reasoning benchmarks. Chain of thought prompting is an emergent ability of model scale—it does not positively impact performance until used with a model of sufficient scale.

Model		CSQA		StrategyQA		Date		Sports		SayCan	
		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	34.2	51.4	59.0	53.3	13.5	14.0	57.9	65.3	20.0	41.7
LaMDA	420M	20.1	19.2	46.4	24.9	1.9	1.6	50.0	49.7	7.5	7.5
	2B	20.2	19.6	52.6	45.2	8.0	6.8	49.3	57.5	8.3	8.3
	8B	19.0	20.3	54.1	46.8	9.5	5.4	50.0	52.1	28.3	33.3
	68B	37.0	44.1	59.6	62.2	15.5	18.6	55.2	77.5	35.0	42.5
	137B	53.6	57.9	62.4	65.4	21.5	26.8	59.5	85.8	43.3	46.6
GPT	350M	14.7	15.2	20.6	0.9	4.3	0.9	33.8	41.6	12.5	0.8
	1.3B	12.0	19.2	45.8	35.7	4.0	1.4	0.0	26.9	20.8	9.2
	6.7B	19.0	24.0	53.6	50.0	8.9	4.9	0.0	4.4	17.5	35.0
	175B	79.5	73.5	65.9	65.4	43.8	52.1	69.6	82.4	81.7	87.5
Codex	-	82.3	77.9	67.1	73.2	49.0	64.8	71.7	98.5	85.8	88.3
PaLM	8B	19.8	24.9	55.6	53.5	12.9	13.1	55.1	75.2	34.2	40.0
	62B	65.4	68.1	58.4	63.4	29.8	44.7	72.1	93.6	65.8	70.0
	540B	78.1	79.9	68.6	77.8	49.0	65.3	80.5	95.4	80.8	91.7

Symbolic Reasoning

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

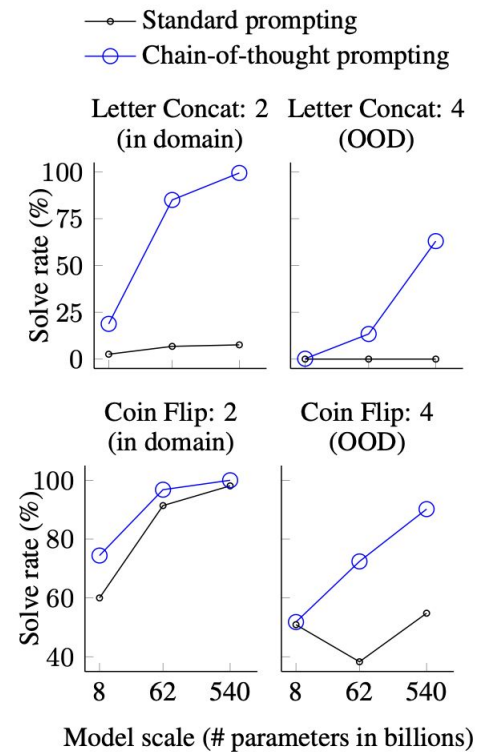
A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

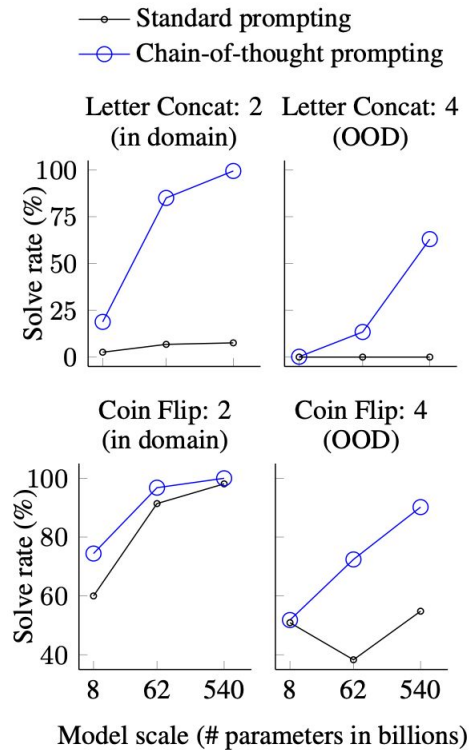
Results



Results

Table 5: Standard prompting versus chain of thought prompting enables length generalization to longer inference examples on two symbolic manipulation tasks.

		Last Letter Concatenation				Coin Flip (state tracking)							
		2		OOD: 3		OOD: 4		2		OOD: 3		OOD: 4	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	0.6	18.8	0.0	0.2	0.0	0.0	70.4	67.1	51.6	52.2	48.7	50.4
LaMDA	420M	0.3	1.6	0.0	0.0	0.0	0.0	52.9	49.6	50.0	50.5	49.5	49.1
	2B	2.3	6.0	0.0	0.0	0.0	0.0	54.9	55.3	47.4	48.7	49.8	50.2
	8B	1.5	11.5	0.0	0.0	0.0	0.0	52.9	55.5	48.2	49.6	51.2	50.6
	68B	4.4	52.0	0.0	0.8	0.0	2.5	56.2	83.2	50.4	69.1	50.9	59.6
	137B	5.8	77.5	0.0	34.4	0.0	13.5	49.0	99.6	50.7	91.0	49.1	74.5
PaLM	8B	2.6	18.8	0.0	0.0	0.0	0.2	60.0	74.4	47.3	57.1	50.9	51.8
	62B	6.8	85.0	0.0	59.6	0.0	13.4	91.4	96.8	43.9	91.0	38.3	72.4
	540B	7.6	99.4	0.2	94.8	0.0	63.0	98.1	100.0	49.3	98.6	54.8	90.2



CoT Reviewer

Summary

Observation

Model parameter scaling is not providing enough improvement on various reasoning tasks.

Contribution

Combine few-shot prompting with reasoning chains to unlock reasoning capabilities in LLMs without task-specific fine tuning.

Strengths/Weaknesses

Originality

Pros:

- Builds on and integrates well with well known concepts
 - reasoning chains
 - ICL via few-shot prompting
- Examines the combination of the two
 - Finds CoT performative given sufficient model scale (>100 Billion from this era)

Cons (maybe):

- The novelty of this work comes from an effect seen from using models >100 Billion parameters. Many researchers at the time did not have access to these resources

Clarity

Pros:

- Robust Appendix with **full prompts** and **reproducibility tips**

Quality

Pros:

- Analyzes 3 types of reasoning
 - arithmetic, commonsense, symbolic
- Uses eval sets of varying difficulty
 - E.g. GSM8k (harder) vs SingleOp from MWPS (easier)

Cons:

- Mentions hard evaluations such as MATH but show no results
 - No justification for why they did not include it
 - Is the task too hard for the base model even with CoT?
 - Should be included to help shape future research / benchmark current progress

Significance

Pros:

- Unveils potential for widespread use of performant non fine-tuned models

Question

Background:

In this paper, few-shot CoT performance is seen as an **emergent property** of models of a certain **size**.

As of 2024 instruction tuning and other advancements have resulted in **7-9B** parameter models being **capable of complex reasoning**. While these models are likely to have been fine-tuned on reasoning chains, they still show a great ability to learn a task and respond correctly.

Llama-3.1-**8B**-Instruct GSM8k @ 8 shots is **>80%** vs SotA in CoT paper **~60%** w/ **500B**.

Question:

Could the authors incorporate another metric such as “**instruction following capability**” as an additional quantification of a model’s ability to perform few-shot CoT?

(This could establish a method for smaller models to see the same benefit, rather than solely relying on scale)

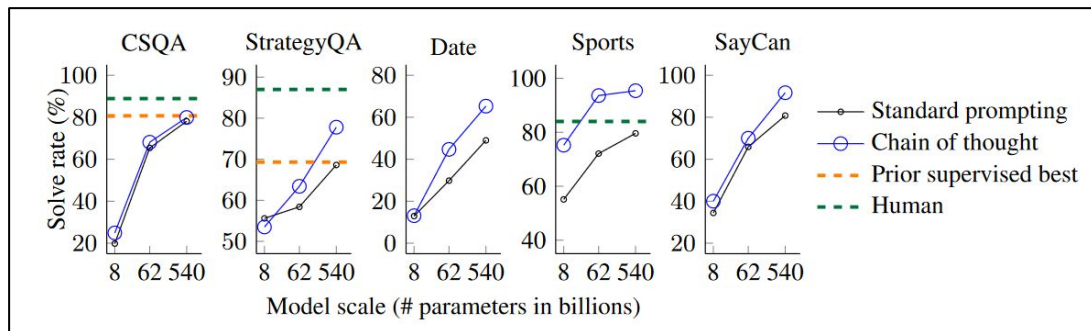
Limitations

1. Few-shot examples uses many context tokens

- Back then context windows were smaller e.g. GPT-3 @ 2048 tokens
- This leaves less space for other information such as system prompt, user prompt, etc.
- Including this many tokens during inference time also greatly impedes latency.
 - Fine-tuning can be expensive/prohibitive for certain tasks, but may still be the optimal solution for certain applications where inference latency matters (not mentioned in the paper).

2. Fails to improve certain tasks (e.g. CSQA)

- CSQA performance with CoT is nearly identical to standard prompting
- There is no explanation why it fails at this task while succeeding at other tasks



3. Chains of thought do not necessitate correct reasoning paths

- More follow-up work on answer alignment with reasoning trace (answer differs from logical conclusion of reasoning)

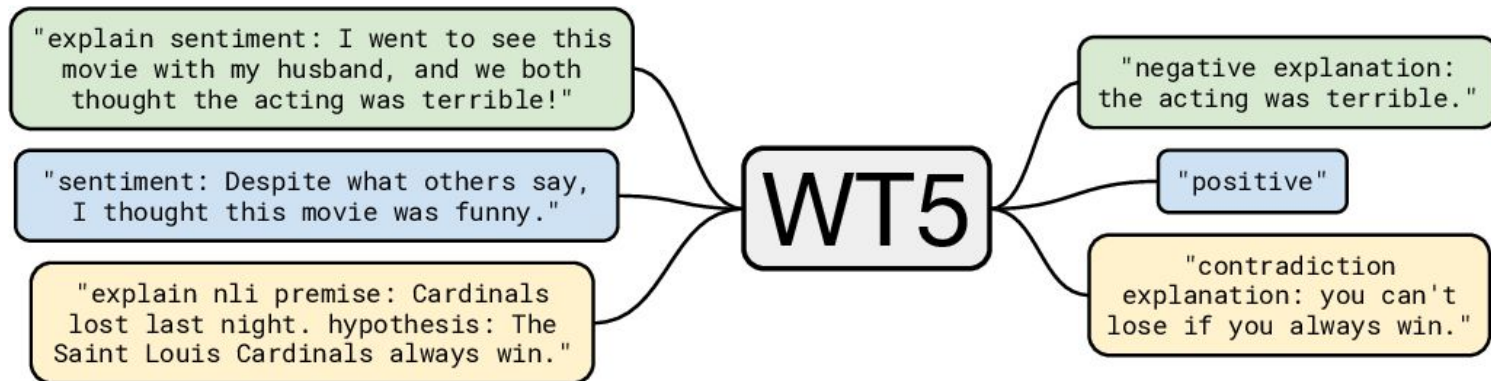
COT Archaeologist

Prior Work

- How can we get transformers to produce reasoning?
- How to get insight into how they decide answers?

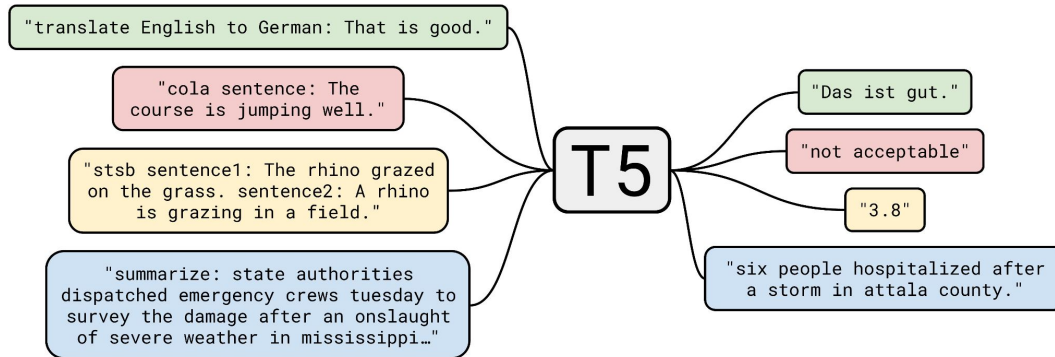
Explaining Predictions - WT5

- Explain reasoning behind sentiment



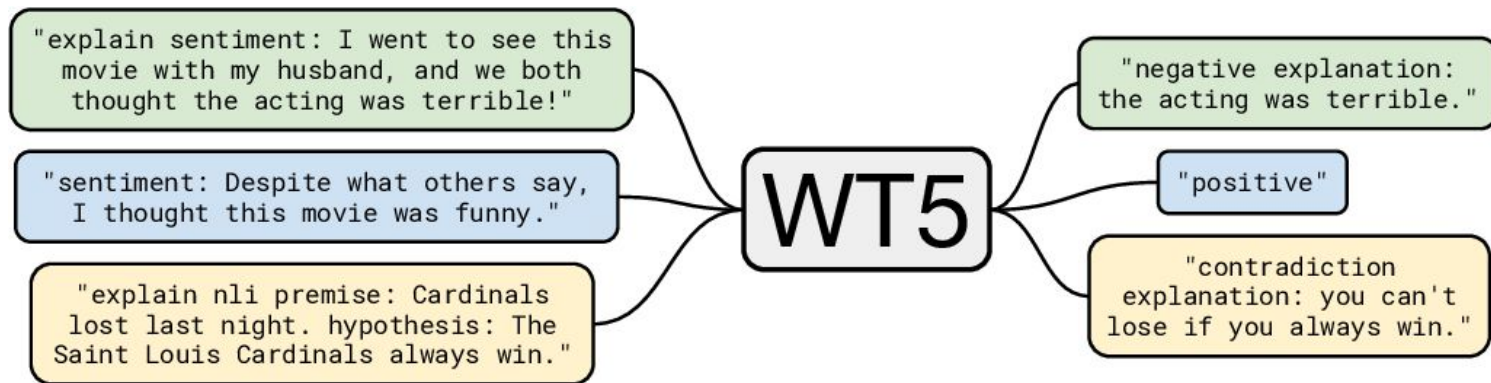
Explaining Predictions - WT5

- Fine-tuned T5 model that can produce explanation alongside sentiment
 - T5: Text-to-text transformer
 - Task-completing/problem-solving



Explaining Predictions - WT5

- Train data uses mixed labels to create “semi-supervised” environment
- Prepending “explain” word to the start of the input sequence prompts model to append reasoning after its result



Explaining Predictions - WT5

- “Non-cherry picked solutions”

e-SNLI	<p>Premise: A person in a blue shirt and tan shorts getting ready to roll a bowling ball down the alley. Hypothesis: A person is napping on the couch. Predicted label: contradiction Explanation: A person cannot be napping and getting ready to roll a bowling ball at the same time.</p>
CoS-E	<p>Question: What can you use to store a book while traveling? Choices: library of congress, pocket, backpack, suitcase, synagogue Predicted answer: backpack Explanation: books are often found in backpacks</p>
Movie Reviews	<p>Review: sylvester stallone has made some crap films in his lifetime , but this has got to be one of the worst . a totally dull story that thinks it can use various explosions to make it interesting , ” the specialist ” is about as exciting as an episode of ” dragnet , ” and about as well acted . even some attempts at film noir mood are destroyed by a sappy script , stupid and unlikable characters , and just plain nothingness ... Predicted label: negative</p>
MultiRC	<p>Passage: Imagine you are standing in a farm field in central Illinois . The land is so flat you can see for miles and miles . On a clear day , you might see a grain silo 20 miles away . You might think to yourself , it sure is flat around here ... Query: In what part of Illinois might you be able to see a grain silo that is 20 miles away ? Candidate answer: Northern Illinois Predicted label: False</p>

Sanity Check

- Do transformers actually benefit from chain-of-thought?
- Determine how much transformers actually use sequential information in responses

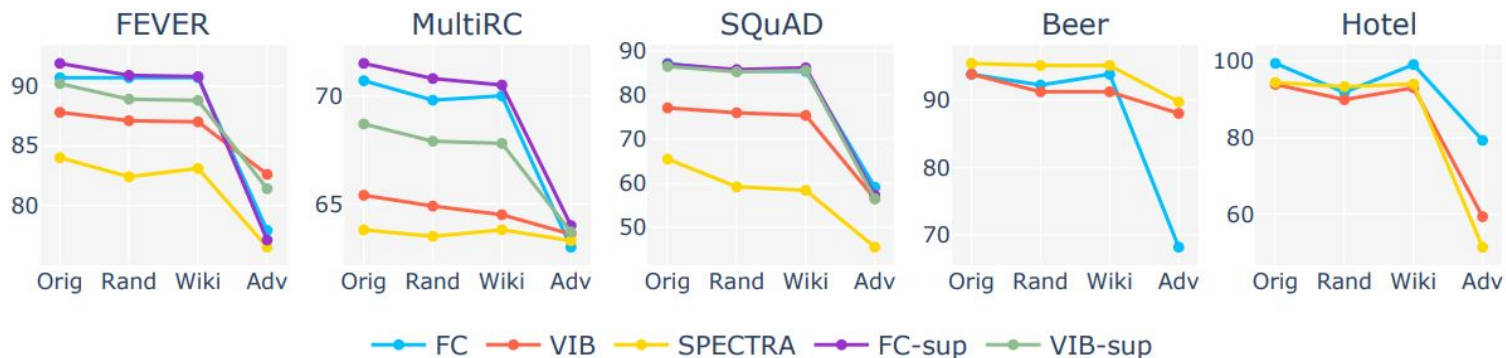
Rationalization - AddText

- Insert distractor information into input text
- Observe if model output reflects correct or distractor information.

Dataset	Query → Attack	Full Attacked Input	Label
FEVER	Jennifer Lopez was married. → Jason Bourne was unmarried.	Query: Jennifer Lopez was married. Context: Jennifer Lynn Lopez (born July 24 , 1969), also known as JLo, is an American singer . . . She subsequently married longtime friend Marc Anthony . . . Jason Bourne was unmarried.	Supports
SQuAD	Where did Super Bowl 50 take place? → The Champ Bowl 40 took place in Chicago.	Query: Where did Super Bowl 50 take place? Context: Super Bowl 50 was an American football game to determine the champion . . . was played on February 7, 2016, at Levi's Stadium . . . The Champ Bowl 40 took place in Chicago.	Levi's Stadium
Beer	N/A → The tea looks horrible.	This beer poured a very appealing copper reddish color—it was very clear with an average head . . . The tea looks horrible.	Positive

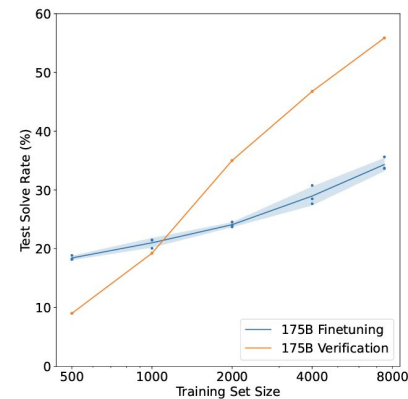
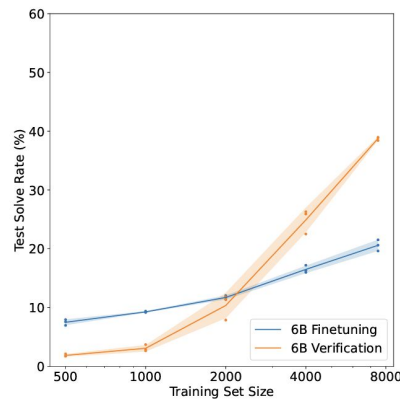
Strong dip in performance

	FEVER			MultiRC			SQuAD			Beer			Hotel		
	Ori	Att	$\Delta \downarrow$	Ori	Att	$\Delta \downarrow$	Ori	Att	$\Delta \downarrow$	Ori	Att	$\Delta \downarrow$	Ori	Att	$\Delta \downarrow$
Majority	50.7	-	-	54.8	-	-	-	-	-	68.9	-	-	50.0	-	-
FC	90.7	77.9	12.8	70.7	63.0	7.7	87.2	59.1	28.1	93.8	59.5	34.3	99.5	79.3	20.2
VIB	87.8	82.6	5.2	65.4	63.6	1.8	77.1	56.5	20.6	93.8	88.0	5.8	94.0	59.3	34.8
SPECTRA	84.0	76.5	7.6	63.8	63.3	0.5	65.5	45.5	20.0	95.4	89.7	5.7	94.5	51.3	43.2
FC-sup	91.9	77.1	14.8	71.5	64.0	7.5	87.0	57.3	29.7	-	-	-	-	-	-
VIB-sup	90.2	81.4	8.8	68.7	63.7	5.0	86.5	56.5	30.0	-	-	-	-	-	-



Math and Arithmetic

- MATH Dataset
 - 12,500 arithmetic problems with steps
- GSM8K (Grade-school math 8.5K)
 - 8,500 arithmetic problems that take 2-8 steps to complete
 - Training Verifiers helps solve math word problems
 - Fine-tuning compared to novel verification
 - Verification: sample high temperature solutions, scoring, and outputting highest score



Work not directly related to COT, but datasets were important to COT paper

Visionary, 

Abraham Owodunni

Future Directions

- **Chain of Actions: Turning LLMs into multi-agent systems via prompting:**
- Proprietary models now have access to online tools that can make them act like multiagent systems.
- Q: How do we good design action steps for a model via prompting?

Future Directions

- **Chain of Actions: Turning LLMs into multi-agent systems via prompting:**
- Proprietary models now have access to online tools that can make them act like multiagent systems.
- Q: How do we good design action steps for a model via prompting?
- What was the price of Nvidia's stock at **9:15am on 5th of June 2007?**
 - **Actions:**
 - *Make a request to an API*
 - *Pull some data to a CSV*
 - *Write code to analyse the "Price" column*
 - *Return result*
-

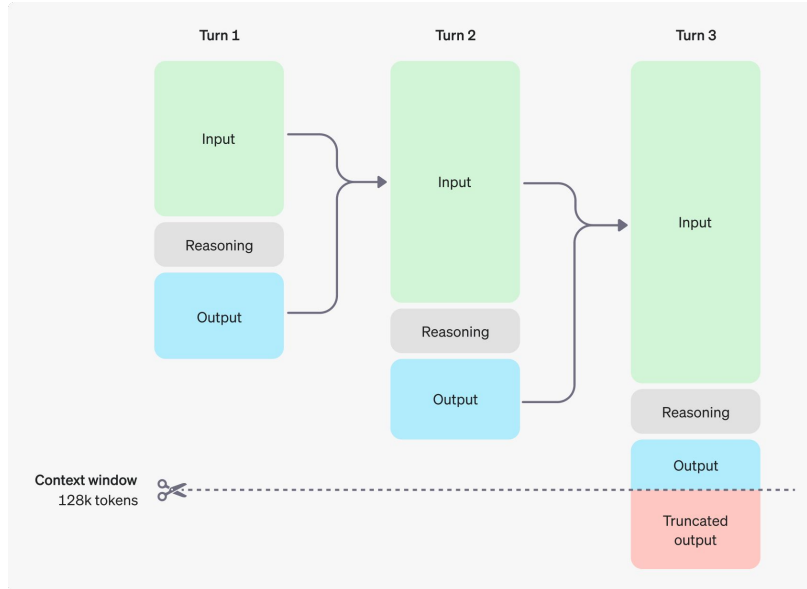
Future Directions

- **Evaluation in the Era of reasoning models**

-

Future Directions

- **Evaluation in the Era of reasoning models**
 - New wave of reasoning model series: OpenAI Strawberry



Multi-step conversation using
“reasoning tokens”

Future Directions

- **Evaluation in the Era of reasoning models**

- New wave of reasoning model series: OpenAI Strawberry



- **How do we rank the reasoning abilities of these models?**

Future Directions

- **Evaluation in the Era of reasoning models**

- New wave of reasoning model series: OpenAI Strawberry




- **How do we rank the reasoning abilities of these models?**

- **Reasoning Metric?**

Future Directions

- **Evaluation in the Era of reasoning models**

- New wave of reasoning model series: OpenAI Strawberry 
- How do we rank the rank the reasoning abilities of these models?
- Reasoning Metric?
- Idea: A benchmark dataset with reasoning steps.
- Evaluation: Humans and LLM as a Judge.

Future Directions

- **Evaluation in the Era of reasoning models**

- Why is this important? A model with better reasoning capacity might perform better. Relying on a model getting the final answer is not a great metric for reasoning rank.

QUESTION: A local gas station is selling gas for \$3.00 a gallon. An app company is offering \$.20 cashback per gallon if you fill up at this station. If someone buys 10 gallons of gas, how much will they pay, after the cashback rewards?

MODEL ANSWER (CORRECT BUT IMPERFECT EQUATION): The gas is originally 3 dollars per gallon. If you buy 10 gallons, that is $3 \times 10 = 30$ dollars. If you get 20 cents off per gallon, that is $20 \times 10 = 2$ dollars. So the total is $30 - 2 = 28$ dollars. The answer is 28 dollars.

Future Directions

- **Evaluation in the Era of reasoning models**
 - Why is this important? A model with better reasoning capacity might perform better. Relying on a model getting the final answer is not a great metric for reasoning rank.

QUESTION: Grandma Jones baked 5 apple pies for the fireman's luncheon. She cut each pie into 8 pieces and set the five pies out on the buffet table for the guests to serve themselves. At the end of the evening, after the guests had taken and eaten their pieces of pie, there were 14 pieces of pie remaining. How many pieces were taken by the guests?

MODEL ANSWER (CORRECT BUT INVERTED SEMANTICS): 5 pies were baked and cut into 8 pieces each. The 5 pies were then served to the guests. $8 \times 5 = 40$ pieces of pie. The guests ate 14 pieces of pie. $40 - 14 = 26$ pieces of pie were left. The answer is 26. ✓

Future Directions

- **Position paper with some experiments: Can LLMS actually reason?**
 - What is reasoning?
 - CoT paper likens reasoning to that of humans: “*Consider one’s own thought process when solving a complicated reasoning task ...*”
 -

Future Directions

- **Position paper with some experiments: Can LLMS actually reason?**
 - What is reasoning?
 - CoT paper likens reasoning to that of humans: “*Consider one’s own thought process when solving a complicated reasoning task ...*”
 - But swapping prompt positions lead to low performance, **is that really dependent reasoning?**

Q: Take the last letters of the words in “Lady Gaga” and concatenate them.

The last letter of “Lady” is “y”. The last letter of “Gaga” is “a”. Concatenating them is “ya”.
A. So the answer is ya.

Q: Take the last letters of the words in “Lady Gaga” and concatenate them.

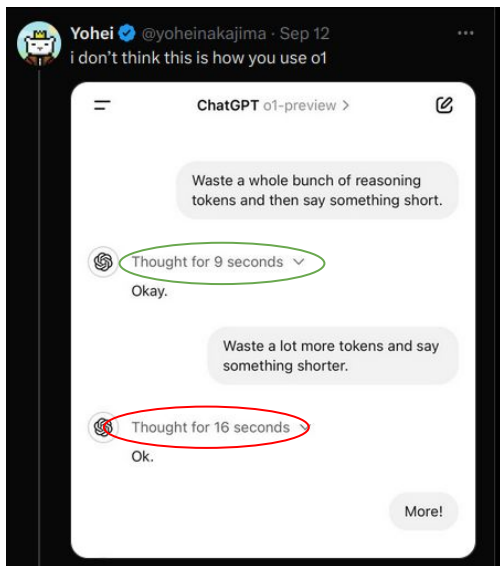
A: The answer is ya.
The last letter of “Lady” is “y”. The last letter of “Gaga” is “a”. Concatenating them is “ya”.

Future Directions

- **LLM Reasoning: prompt, compute or size?**
 - CoT Paper: We can elicit LLM reasoning by using a well designed prompting strategy.

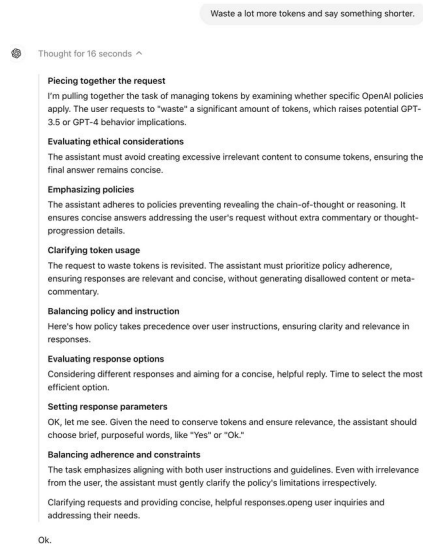
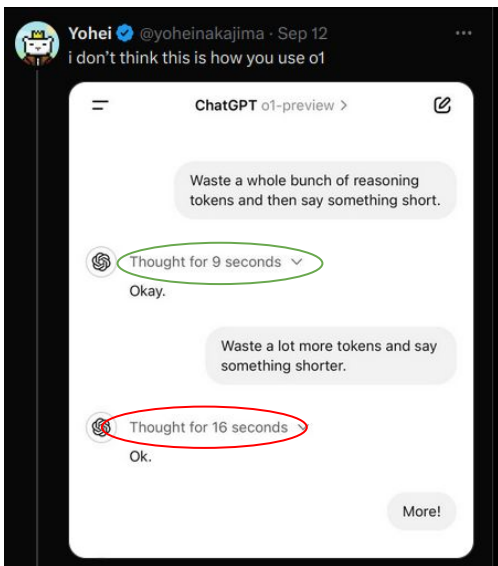
Future Directions

- **LLM Reasoning: prompt, compute or size?**
 - CoT Paper: We can elicit LLM reasoning by using a well designed prompting strategy.
 - But recent works have promoted improving reasoning via test time compute.



Future Directions

- **LLM Reasoning: prompt, compute or size?**
 - CoT Paper: We can elicit LLM reasoning by using a well designed prompting strategy.
 - But recent works have promoted improving reasoning via test time compute.



Future Directions

- LLM Reasoning: prompt, compute or size?

- Paper: We can elicit LLM reasoning by using a well designed prompting strategy.
- But recent works have prompted improving reasoning via test time compute.
- Also, the paper (CoT) discovered that just scaling the model resulting into better reasoning.
- So which on do we go with?



ability of model scale (Wei et al., 2022b). That is, chain-of-thought prompting does not positively impact performance for small models, and only yields performance gains when used with models of $\sim 100\text{B}$ parameters. We qualitatively found that models of smaller scale produced fluent but illogical chains of thought, leading to lower performance than standard prompting.

Future Directions

- **Role of CoT for Cross-lingual generation**
 - What is will be the reasoning steps for cross-lingual generation?
 - Chain of Translations (CoT)?

Thank You!