Scaling Instruction-Finetuned Language Models

Flan-PaLM
540B

Paper: H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

Motivation

Pretraining has driven NLP progress GPT-3

instruction finetuning brings further gains

InstructGPT FLAN T0

This work: scaling up instruction finetuning

Key findings:

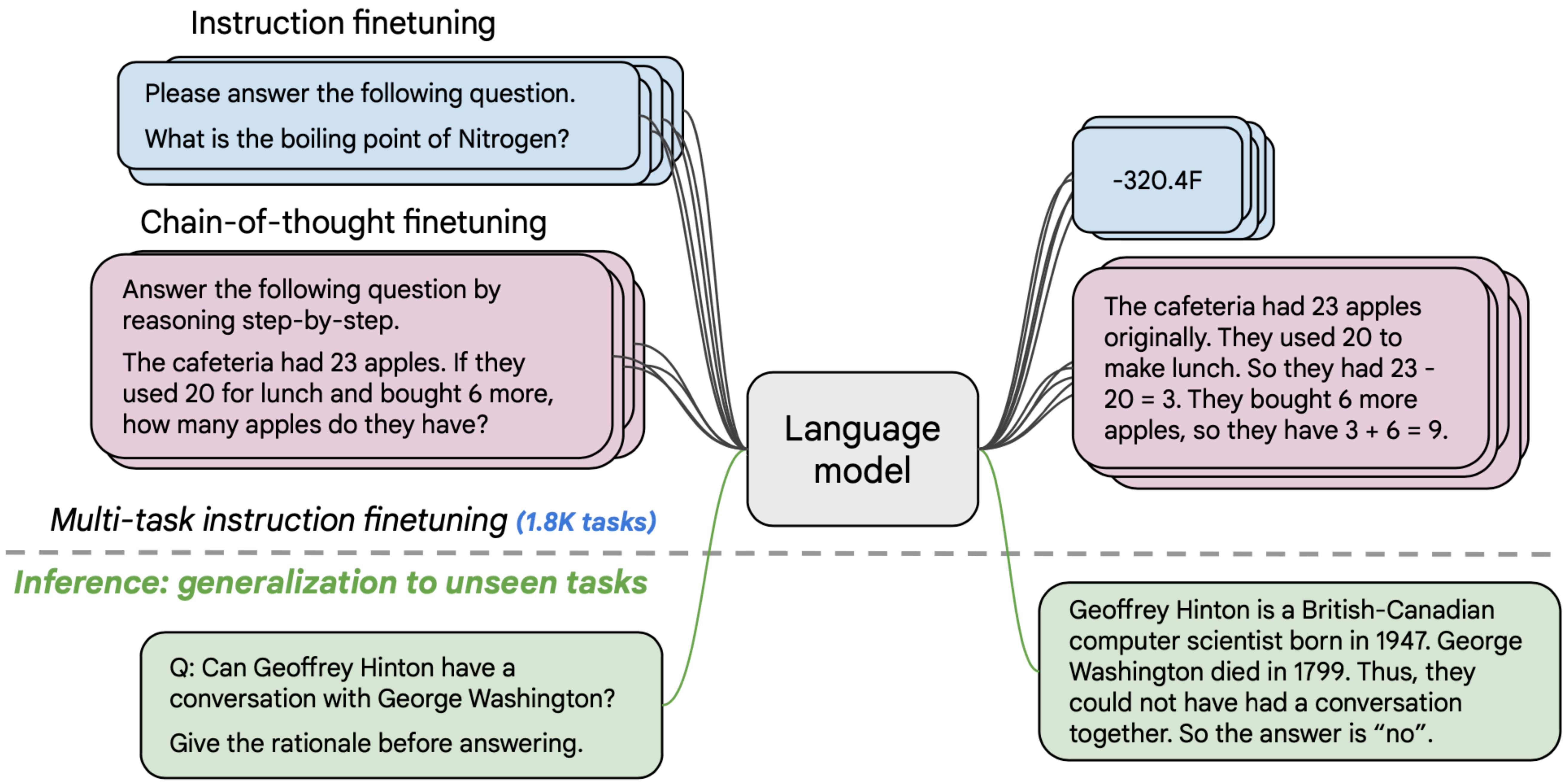
- scaling up the model
- scaling up the number of tasks

both yield performance gains

chain-of-thought (CoT) in finetuning mix: useful

Highlight: Flan-PaLM 540B finetuned on 1.8K tasks

beats SotA (PaLM 540B) +9.4% across 4 benchmarks



Instruction finetuning

Please answer the following question.
What is the boiling point of Nitrogen?

-320.4F

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Language model

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.

Multi-task instruction finetuning **(1.8K tasks)**

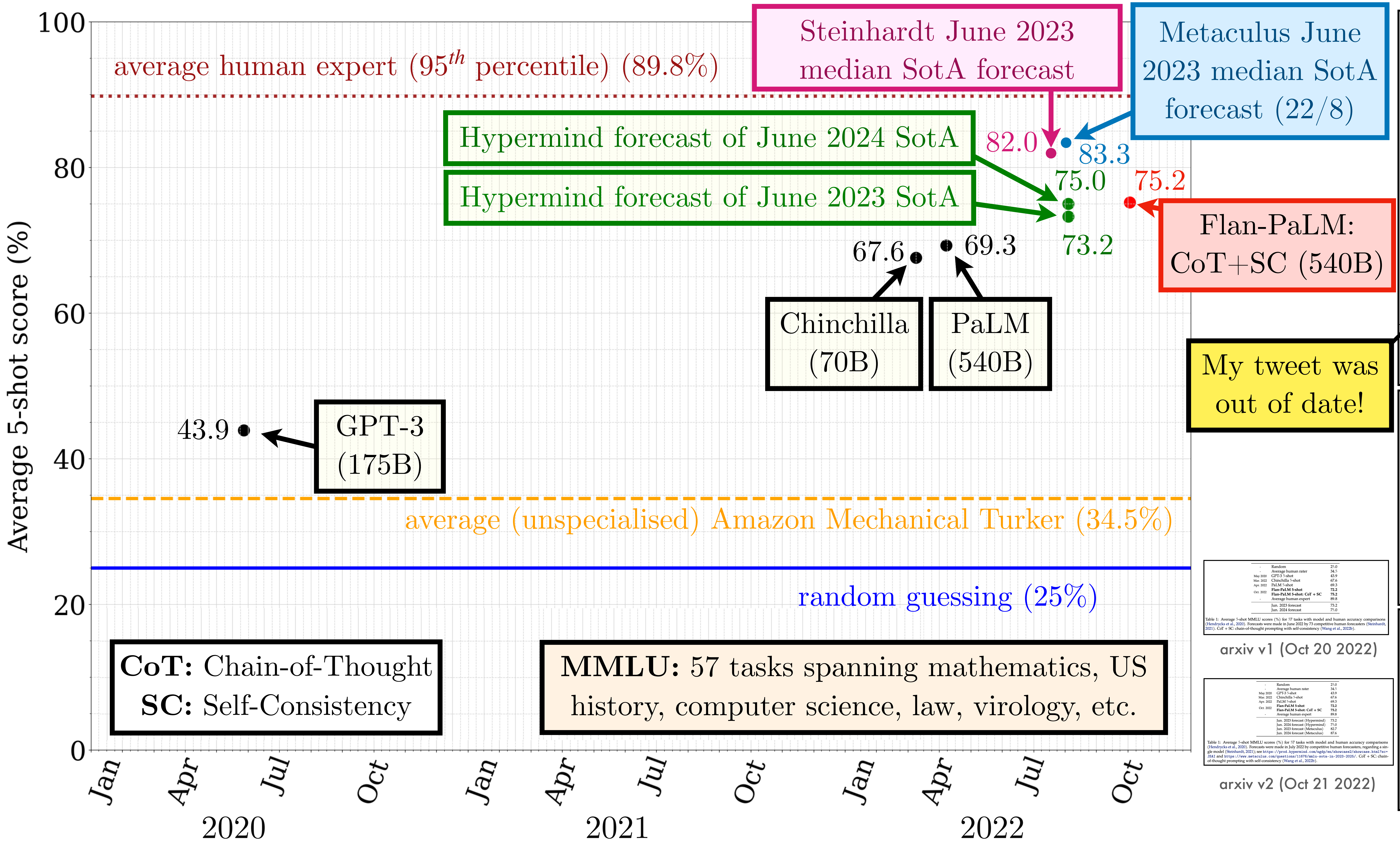
Inference: generalization to unseen tasks

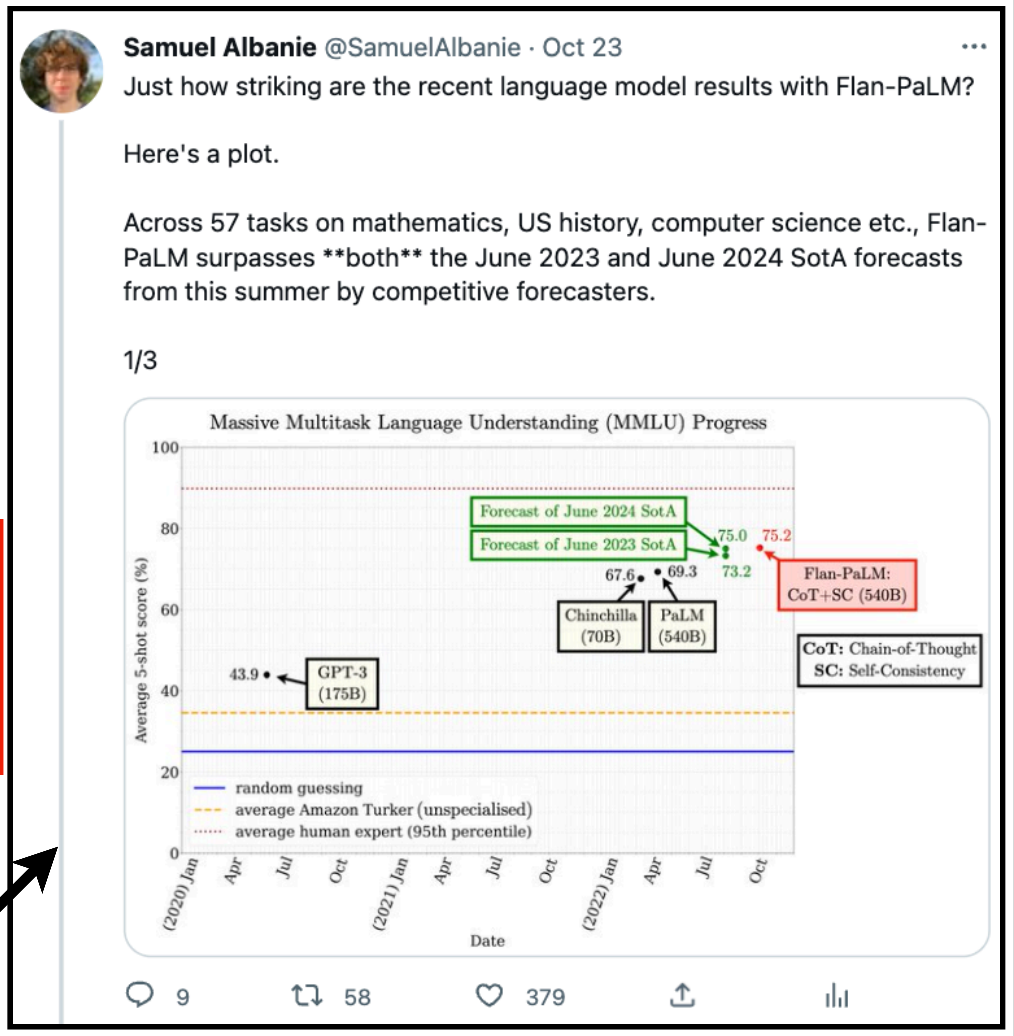
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

References/image credits:
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(InstructGPT) L. Ouyang et al., "Training language models to follow instructions with human feedback", arxiv (2022)
(FLAN) J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)

Progress on Massive Multitask Language Understanding (MMLU)





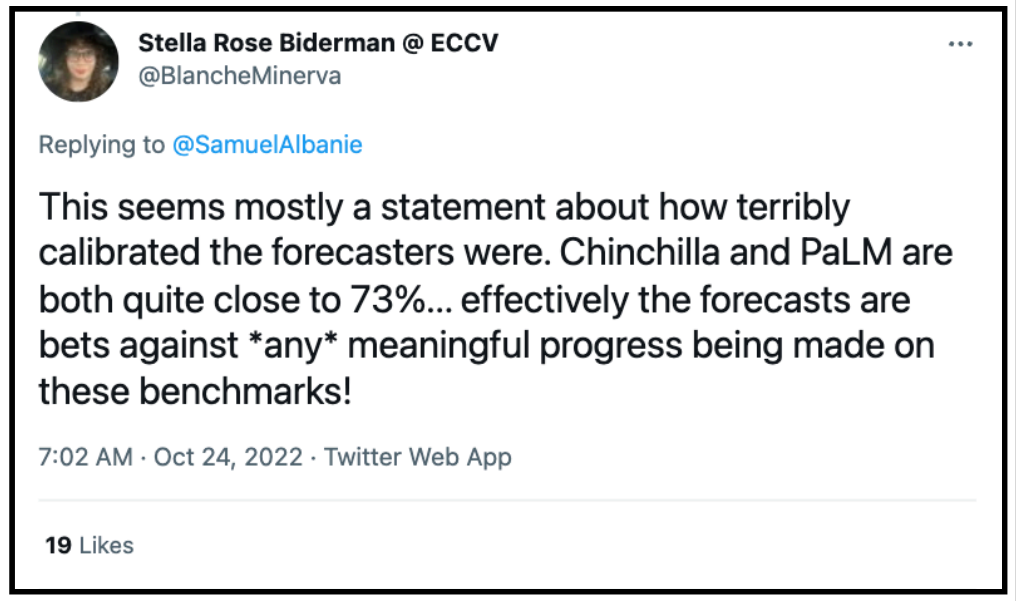
Samuel Albanie @SamuelAlbanie · Oct 23

Just how striking are the recent language model results with Flan-PaLM?

Here's a plot.

Across 57 tasks on mathematics, US history, computer science etc., Flan-PaLM surpasses **both** the June 2023 and June 2024 SotA forecasts from this summer by competitive forecasters.

1/3



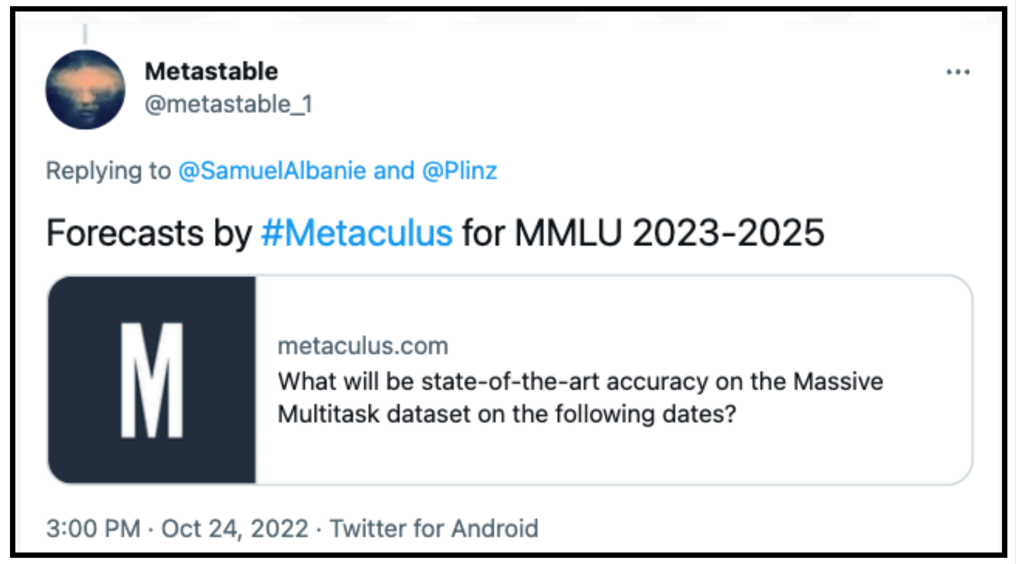
Stella Rose Biderman @ ECCV
@BlancheMinerva

Replying to @SamuelAlbanie

This seems mostly a statement about how terribly calibrated the forecasters were. Chinchilla and PaLM are both quite close to 73%... effectively the forecasts are bets against *any* meaningful progress being made on these benchmarks!

7:02 AM · Oct 24, 2022 · Twitter Web App

19 Likes



Metastable
@metastable_1

Replying to @SamuelAlbanie and @Plinz

Forecasts by #Metaculus for MMLU 2023-2025

metaculus.com
What will be state-of-the-art accuracy on the Massive Multitask dataset on the following dates?

3:00 PM · Oct 24, 2022 · Twitter for Android

Twitter discussion

References
(MMLU) D. Hendrycks et al., "Measuring Massive Multitask Language Understanding", ICLR (2020)
(Code to reproduce plot) https://colab.research.google.com/drive/1plhkdUhQB-IijHAQTXlQaGolIgcBGrfe?usp=sharing
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(Chinchilla) J. Hoffmann et al., "Training Compute-Optimal Large Language Models", arxiv (2022)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)
(Forecasts) https://prod.hypermind.com/ngdp/en/showcase2/showcase.html?sc=JSAI (challenges updated 15th August 2022)
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(SC) X. Wang et al., "Self-consistency improves chain of thought reasoning in language models", arxiv (2022)
(Twitter thread for discussion of forecasts) https://twitter.com/SamuelAlbanie/status/1584257440891416576

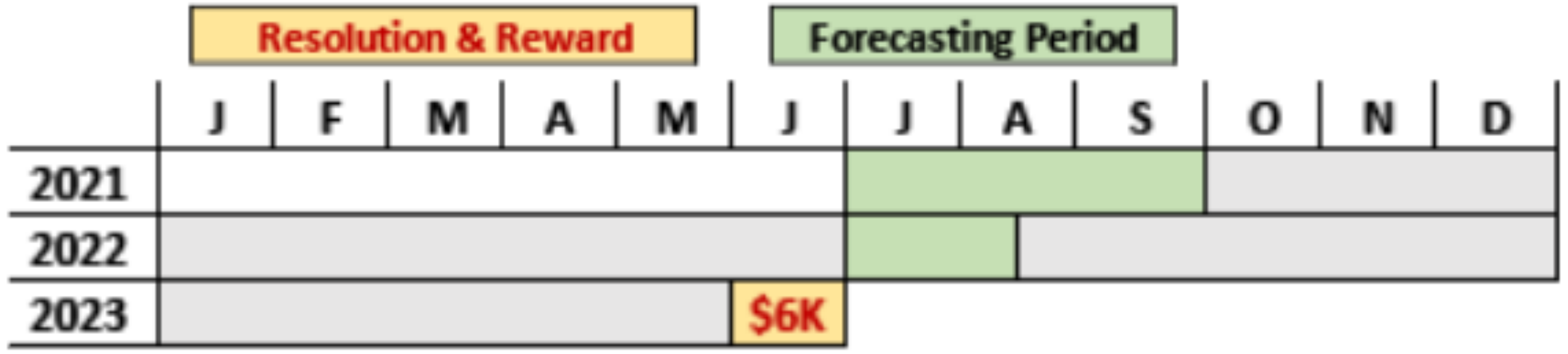
Background on the June 2023 MMLU Hypermind Forecast

Commissioned by J. Steinhardt and his team as part of an effort to forecast AI progress

Topics: geopolitics & future capabilities (5K USD prize/benchmark for 2022, 2023, 2024, 2025)

Funded by Open Philanthropy and run on the Hypermind prediction market platform

Structure of competition:



Recommended reading

AI Forecasting: One Year In

JUL 3, 2022 • 7 MIN READ

How were results one year in? Not very good (2/4 forecasts outside 90% credible intervals)

Question: "Was progress surprising, or were the forecasters bad?"

Possible limitations: $5K prize per benchmark (4 questions) - not enough incentive

Interface does not allow arbitrary prob. distributions

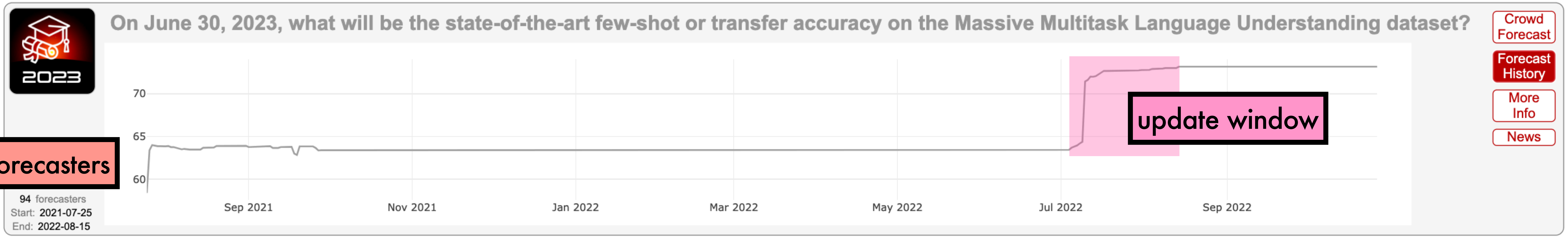
Forecasters just not skilled enough

Subjective assessment from Steinhardt: progress until June 30th 2022 was still "surprising"

References/Notes/Image credits
(Forecasts) https://prod.hypermind.com/ngdp/en/showcase2/showcase.html?sc=JSAI (relevant challenges updated 15th August 2022)
(Hypermind prediction market) https://predict.hypermind.com/hypermind/app.html#welcome
(AI Forecasting: One Year In) https://bounded-regret.ghost.io/ai-forecasting-one-year-in/

Steinhardt credits the question design to Alex Wei, Collin Burns, Jean-Stanislas Denain, and Dan Hendrycks https://bounded-regret.ghost.io/ai-forecasting/

The MMLU June 2023 Forecast (updated August 2022)



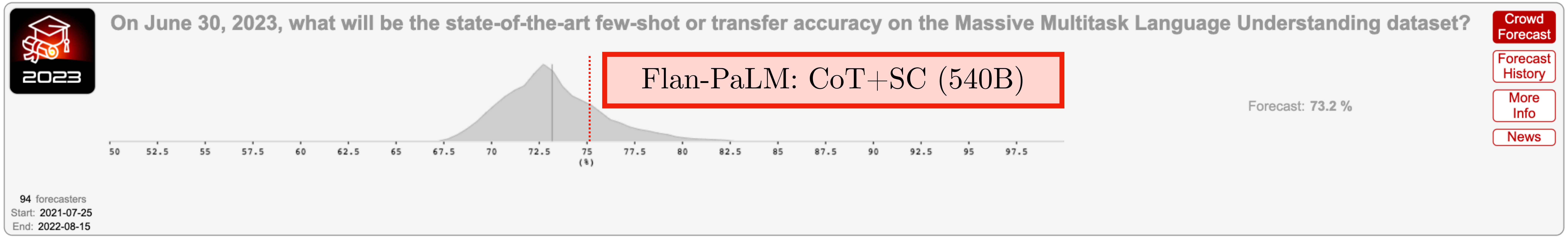
On June 30, 2023, what will be the state-of-the-art few-shot or transfer accuracy on the Massive Multitask Language Understanding dataset?

94 forecasters
Start: 2021-07-25
End: 2022-08-15

update window

94 forecasters

Note: "the hypermind aggregate combines all crowd forecasts but places higher weight on forecasters with a good track record"



On June 30, 2023, what will be the state-of-the-art few-shot or transfer accuracy on the Massive Multitask Language Understanding dataset?

Flan-PaLM: CoT+SC (540B)

94 forecasters
Start: 2021-07-25
End: 2022-08-15

Other forecasts: Metaculus and Steinhardt

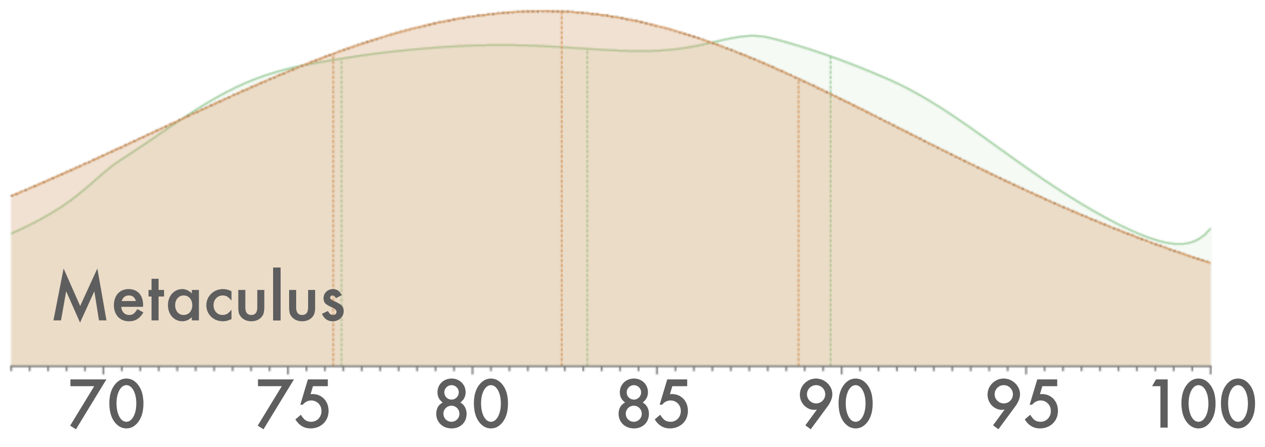
Forecasting ML Benchmarks in 2023
JUL 17, 2022 • 15 MIN READ

Recommended reading

"Interestingly, the Hypermind median is only at 72.5% right now. Given the ability to combine Minerva + Chinchilla, this intuitively seems too low to me." (Steinhardt, July 2022)

Steinhardt median estimate: 82%

Probability density   Closes 29th June 2023



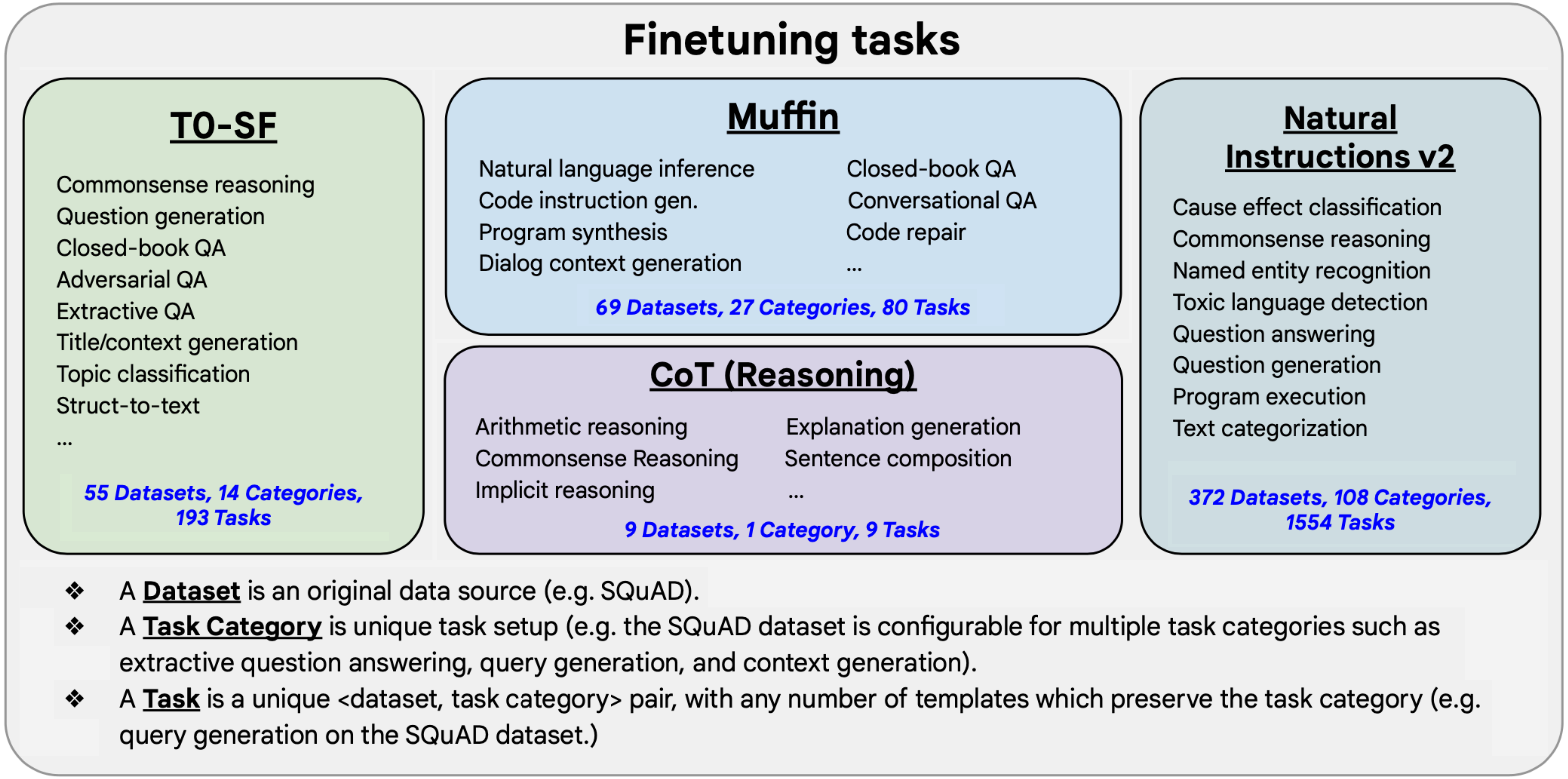
Takeaway: the hypermind forecasts were conservative relative to alternatives

Recommendation: contribute to future forecasts!

References
(MMLU) D. Hendrycks et al., "Measuring Massive Multitask Language Understanding", ICLR (2020)
(Reference for hypermind aggregate) J. Steinhardt, https://bounded-regret.ghost.io/ai-forecasting/ (2021)
(Forecast for 2023 and quote about hypermind median) https://bounded-regret.ghost.io/forecasting-math-and-mmlu-in-2023/

Flan Finetuning: Data and Language Models

Data



Finetuning tasks

T0-SF

- Commonsense reasoning
- Question generation
- Closed-book QA
- Adversarial QA
- Extractive QA
- Title/context generation
- Topic classification
- Struct-to-text
- ...

55 Datasets, 14 Categories, 193 Tasks

Muffin

- Natural language inference
- Code instruction gen.
- Program synthesis
- Dialog context generation
- Closed-book QA
- Conversational QA
- Code repair
- ...

69 Datasets, 27 Categories, 80 Tasks

CoT (Reasoning)

- Arithmetic reasoning
- Commonsense Reasoning
- Implicit reasoning
- Explanation generation
- Sentence composition
- ...

9 Datasets, 1 Category, 9 Tasks

Natural Instructions v2

- Cause effect classification
- Commonsense reasoning
- Named entity recognition
- Toxic language detection
- Question answering
- Question generation
- Program execution
- Text categorization

372 Datasets, 108 Categories, 1554 Tasks

- A **Dataset** is an original data source (e.g. SQuAD).
- A **Task Category** is unique task setup (e.g. the SQuAD dataset is configurable for multiple task categories such as extractive question answering, query generation, and context generation).
- A **Task** is a unique <dataset, task category> pair, with any number of templates which preserve the task category (e.g. query generation on the SQuAD dataset.)

Language models

T5 (2020)

PaLM (2022)

U-PaLM (2022)

Model sizes studied

80 million parameters

↓ *up to*

540 billion parameters

Note: finetuning uses at most 1.6% of compute (relative to pretraining)

References/image credits:

H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

(Muffin) based on J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022) and other datasets

(T0-SF) based on V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)

(Natural Instructions v2) Y. Wang et al., "Benchmarking generalization via in-context instructions on 1,600+ language tasks", arxiv (2022)

(T5) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2020)

(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)

(U-PaLM) Y. Tay et al., "Transcending Scaling Laws with 0.1% Extra Compute", arxiv (2022)

Evaluation data

Objective: assess Flan-PaLM on world knowledge and reasoning tasks

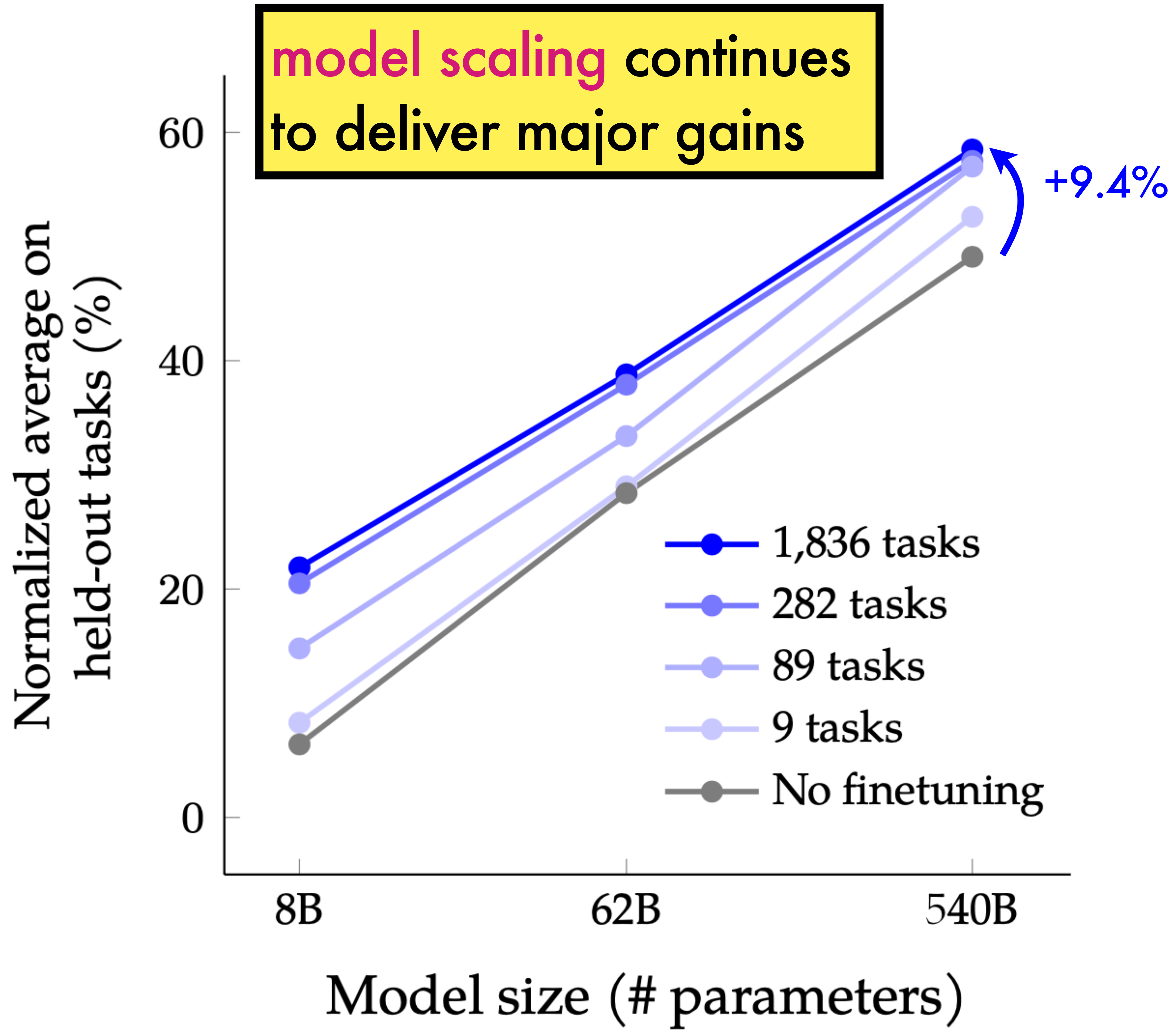
Do not evaluate on GPT-3 test suite (training sets are included in Flan-PaLM finetuning mixture)

MMLU	exam questions across 57 tasks (mathematics, law, medicine etc.)
BBH	23 tasks from BIG-Bench (where PaLM is worse than average human rater)
TyDiQA	question answering across 8 typologically diverse languages
MGSM	multilingual benchmark of mathematics problems translated to 10 languages

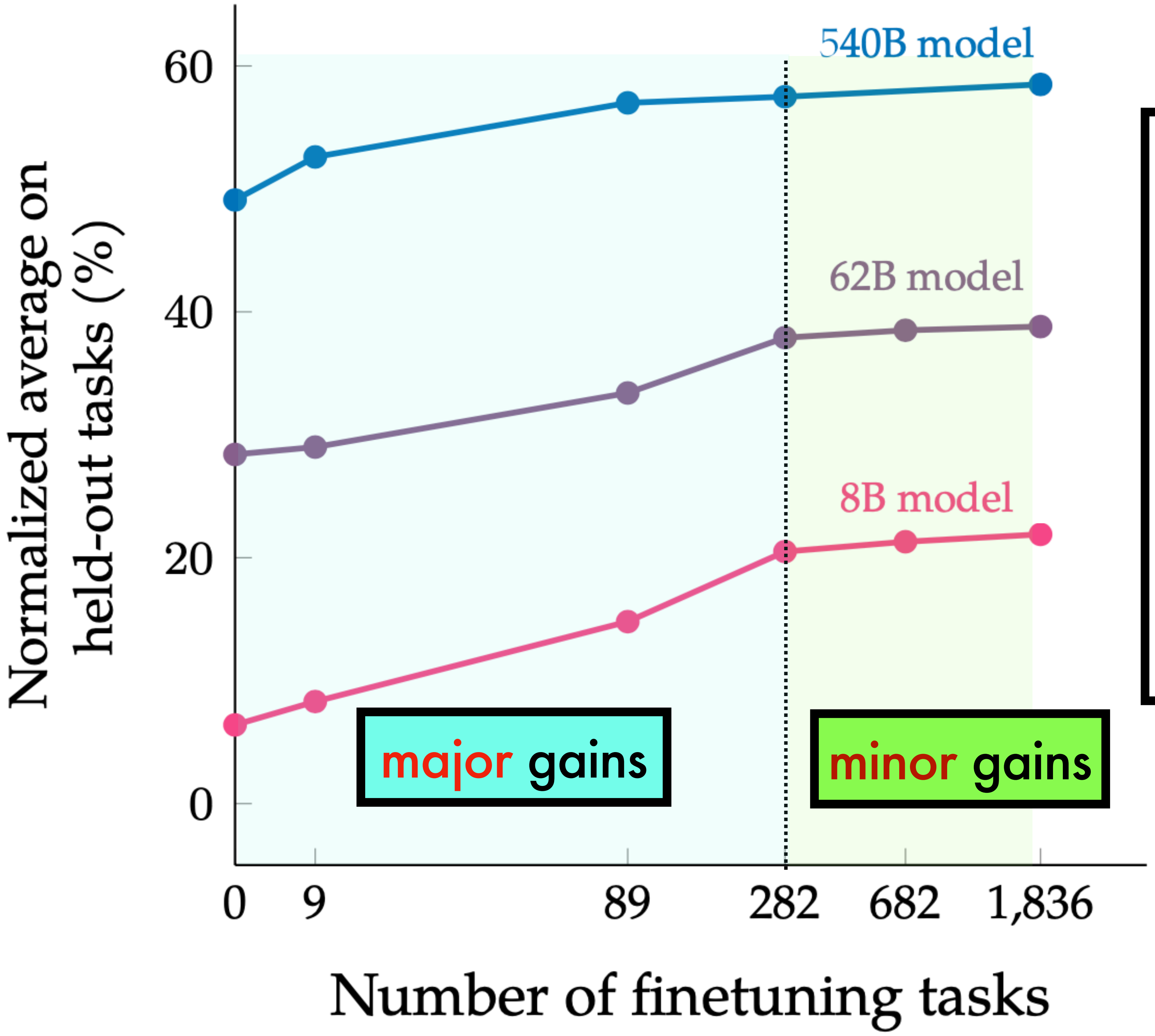
Note: benchmarks also used by PaLM (did not find data contamination)

References:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(GPT-3) T. Brown et al., "Language models are few-shot learners", NeurIPS (2020)
(MMLU) D. Hendrycks et al., "Measuring Massive Multitask Language Understanding", ICLR (2020)
(BBH) M. Suzgun et al., "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them", arxiv (2022)
(TyDiQA) J. Clark et al., "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages", ACL (2020)
(MGSM) F. Shi et al., "Language Models are Multilingual Chain-of-Thought Reasoners", arxiv (2022)

Scaling up to 540B parameters and 1.8K tasks



model scaling continues to deliver major gains



major gains

minor gains

Why do gains flatten?

Possible explanations:

1. Extra tasks not diverse
2. Gains come from eliciting existing knowledge

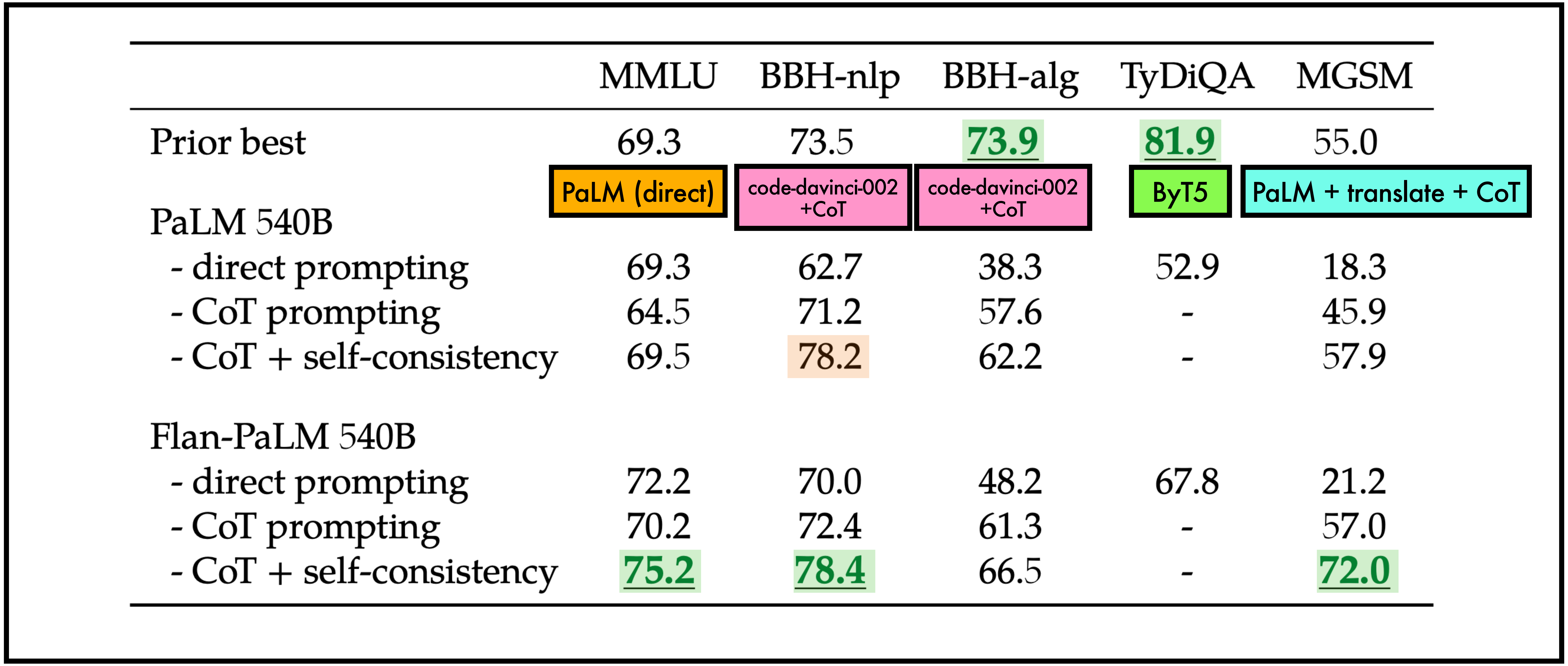
Note: finetuning contributes only 0.2% of total tokens

Experiments on PaLM architectures few-shot prompted accuracy (exact match) on held-out tasks

Metric: normalised average on MMLU, BBH, TyDiQA, MGSM (various direct/CoT combinations)

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

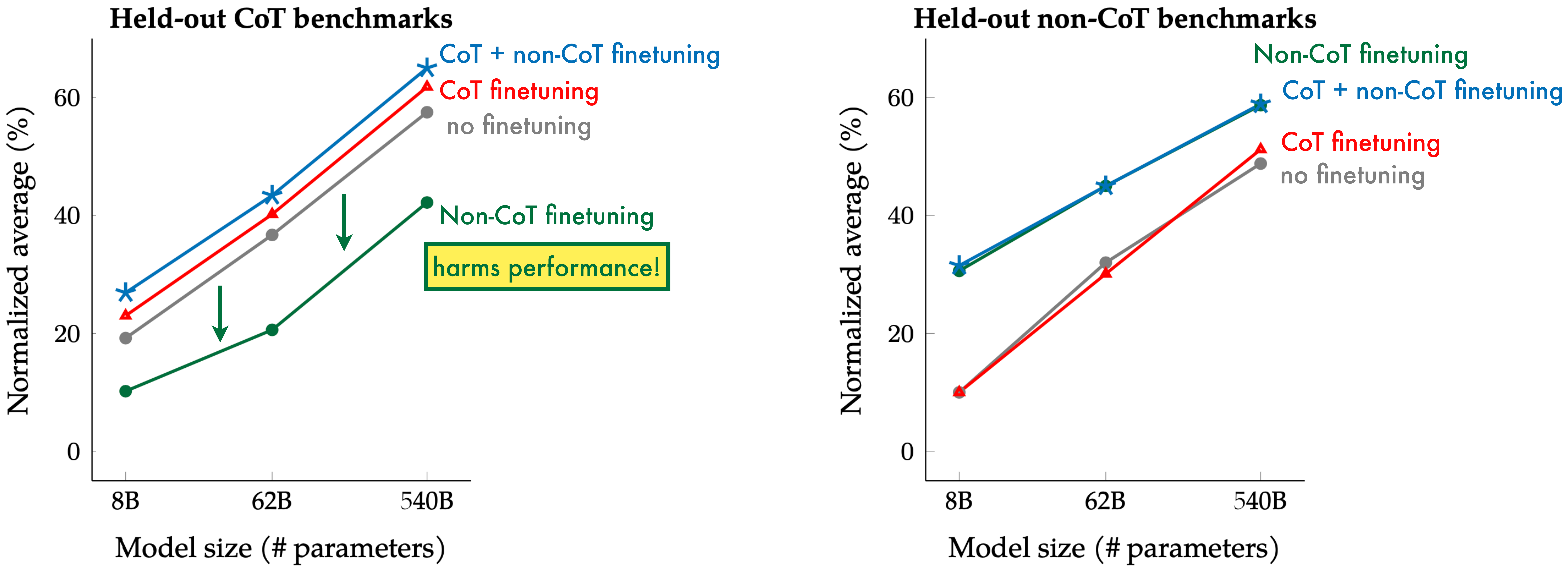
Flan-PaLM CoT finetuning (comparison to SotA)



	MMLU	BBH-nlp	BBH-alg	TyDiQA	MGSM
Prior best	69.3	73.5	**73.9**	**81.9**	55.0
	PaLM (direct)	code-davinci-002 +CoT	code-davinci-002 +CoT	ByT5	PaLM + translate + CoT
PaLM 540B					
- direct prompting	69.3	62.7	38.3	52.9	18.3
- CoT prompting	64.5	71.2	57.6	-	45.9
- CoT + self-consistency	69.5	78.2	62.2	-	57.9
Flan-PaLM 540B					
- direct prompting	72.2	70.0	48.2	67.8	21.2
- CoT prompting	70.2	72.4	61.3	-	57.0
- CoT + self-consistency	**75.2**	**78.4**	66.5	-	**72.0**

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)
(code-davinci-002) M. Chen et al. "Evaluating large language models trained on code", arxiv (2021)
(ByT5) L. Xue et al., "Byt5: Towards a token-free future with pre-trained byte-to-byte models", TACL (2022)
(PaLM+translate+CoT) F. Shi et al., "Language models are multilingual chain-of-thought reasoners", arxiv (2022)
(SC) X. Wang et al., "Self-consistency improves chain of thought reasoning in language models", arxiv (2022)

Including chain-of-thought data in finetuning



Note: there are only 9 CoT finetuning datasets, but 496 non-CoT finetuning datasets

Takeaway: instruction finetuning improves unseen tasks in the same prompting paradigm

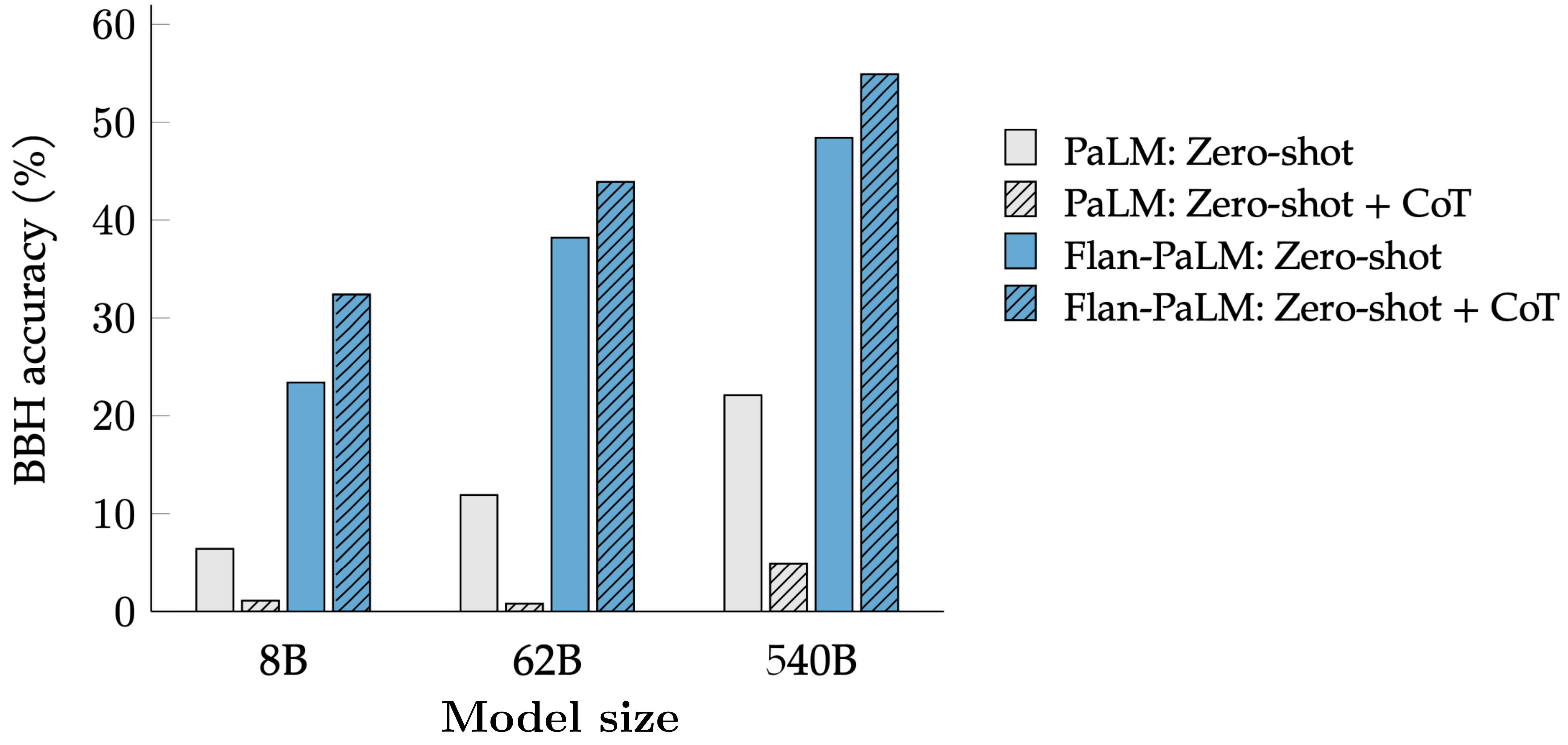
Both CoT and non-CoT required for good results in both paradigms

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

Zero-shot chain-of-thought reasoning

Zero-shot CoT reasoning useful since few-shot CoT exemplars require prompt engineering

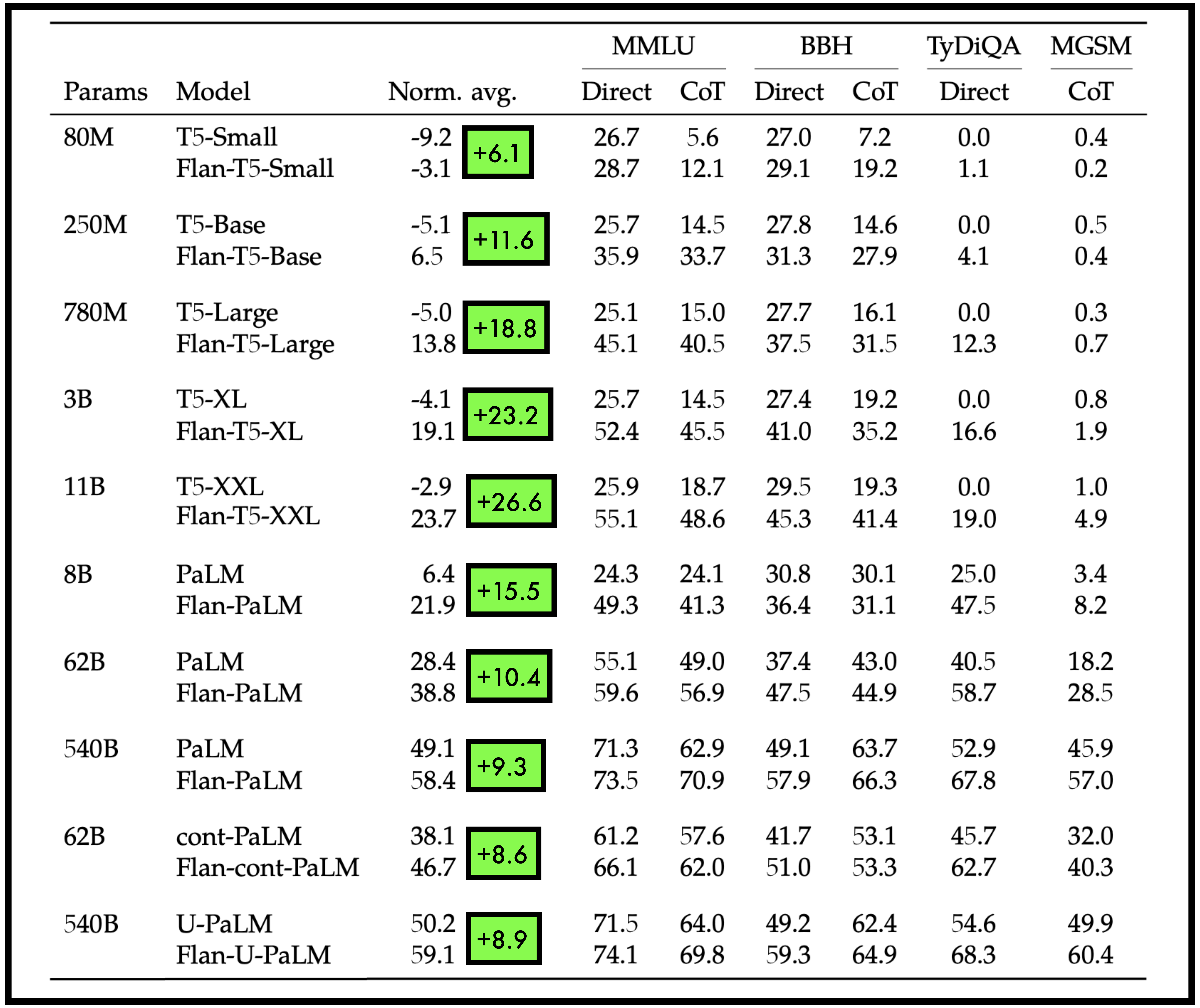
The phrase "let's think step by step" is used for zero-shot CoT prompting



Takeaway: Finetuning on CoT data enables zero-shot CoT reasoning

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
("Let's think step by step") T. Kojima et al., "Large Language Models are Zero-Shot Reasoners", arxiv (2022)

Results: instruction finetuning for different models



				MMLU		BBH		TyDiQA	MGSM
Params	Model	Norm. avg.		Direct	CoT	Direct	CoT	Direct	CoT
80M	T5-Small	-9.2	+6.1	26.7	5.6	27.0	7.2	0.0	0.4
	Flan-T5-Small	-3.1		28.7	12.1	29.1	19.2	1.1	0.2
250M	T5-Base	-5.1	+11.6	25.7	14.5	27.8	14.6	0.0	0.5
	Flan-T5-Base	6.5		35.9	33.7	31.3	27.9	4.1	0.4
780M	T5-Large	-5.0	+18.8	25.1	15.0	27.7	16.1	0.0	0.3
	Flan-T5-Large	13.8		45.1	40.5	37.5	31.5	12.3	0.7
3B	T5-XL	-4.1	+23.2	25.7	14.5	27.4	19.2	0.0	0.8
	Flan-T5-XL	19.1		52.4	45.5	41.0	35.2	16.6	1.9
11B	T5-XXL	-2.9	+26.6	25.9	18.7	29.5	19.3	0.0	1.0
	Flan-T5-XXL	23.7		55.1	48.6	45.3	41.4	19.0	4.9
8B	PaLM	6.4	+15.5	24.3	24.1	30.8	30.1	25.0	3.4
	Flan-PaLM	21.9		49.3	41.3	36.4	31.1	47.5	8.2
62B	PaLM	28.4	+10.4	55.1	49.0	37.4	43.0	40.5	18.2
	Flan-PaLM	38.8		59.6	56.9	47.5	44.9	58.7	28.5
540B	PaLM	49.1	+9.3	71.3	62.9	49.1	63.7	52.9	45.9
	Flan-PaLM	58.4		73.5	70.9	57.9	66.3	67.8	57.0
62B	cont-PaLM	38.1	+8.6	61.2	57.6	41.7	53.1	45.7	32.0
	Flan-cont-PaLM	46.7		66.1	62.0	51.0	53.3	62.7	40.3
540B	U-PaLM	50.2	+8.9	71.5	64.0	49.2	62.4	54.6	49.9
	Flan-U-PaLM	59.1		74.1	69.8	59.3	64.9	68.3	60.4

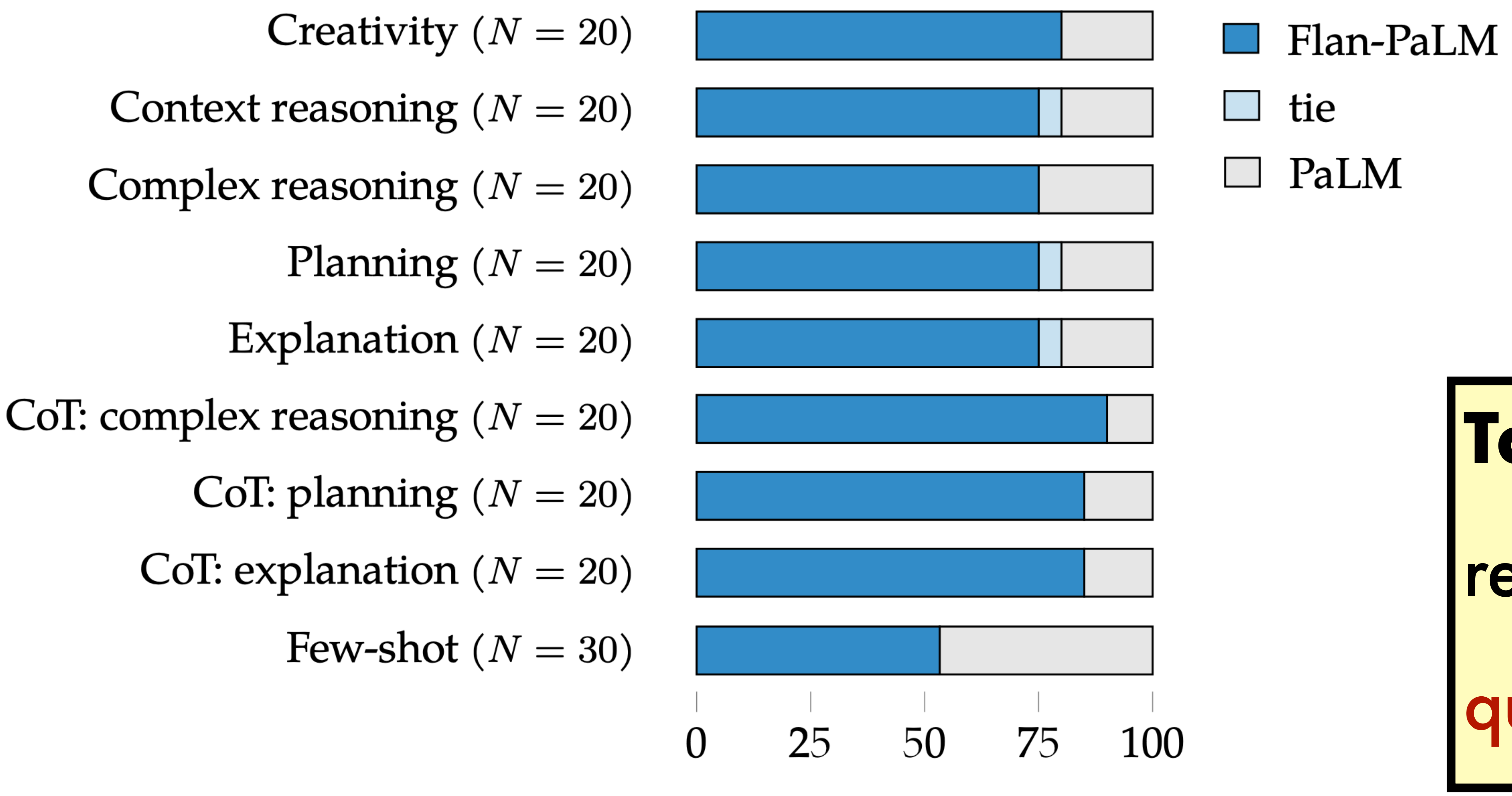
References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(T5) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2020)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)
(U-PaLM) Y. Tay et al., "Transcending Scaling Laws with 0.1% Extra Compute", arxiv (2022)

Evaluating open-ended generation

Standard benchmarks may not correlate with human preferences about model outputs

Human study is run on outputs on 190 open-ended questions spanning creativity, reasoning etc.

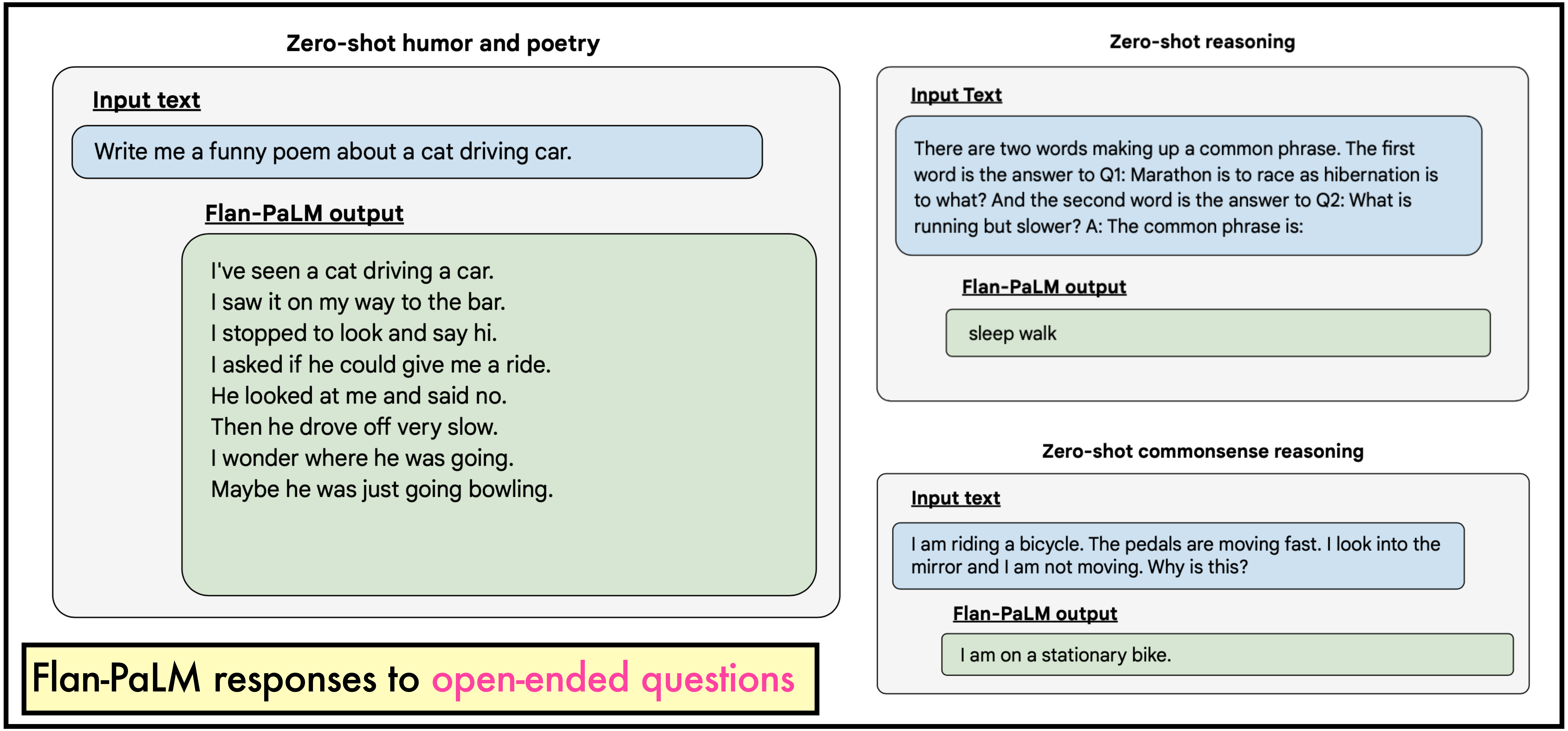
Compare PaLM 540B and Flan-PaLM 540B (sample 5 responses with $\tau = 0.7$, rank and filter)



Takeaway: instruction-finetuning produces responses to open-ended zero-shot questions that are preferred by humans

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(observation that human preferences can differ from benchmark rankings) L. Ouyang et al., "Training language models to follow instructions with human feedback", arxiv (2022)

Qualitative Examples



Zero-shot humor and poetry

Input text

Write me a funny poem about a cat driving car.

Flan-PaLM output

I've seen a cat driving a car.
I saw it on my way to the bar.
I stopped to look and say hi.
I asked if he could give me a ride.
He looked at me and said no.
Then he drove off very slow.
I wonder where he was going.
Maybe he was just going bowling.

Zero-shot reasoning

Input Text

There are two words making up a common phrase. The first word is the answer to Q1: Marathon is to race as hibernation is to what? And the second word is the answer to Q2: What is running but slower? A: The common phrase is:

Flan-PaLM output

sleep walk

Zero-shot commonsense reasoning

Input text

I am riding a bicycle. The pedals are moving fast. I look into the mirror and I am not moving. Why is this?

Flan-PaLM output

I am on a stationary bike.

Flan-PaLM responses to open-ended questions

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

Related Work

Instruction finetuning

InstructGPT (2022) Flan (2022) T0 (2022)

Differences to prior work:

- Build on mixtures of prior datasets (and add new CoT data etc.)
- Explore larger model scaling (up to 540B params vs 137B)
- Finetune on a mixture of zero-shot and few-shot instruction formats

Reasoning via finetuning

Prior work has either:

- finetuned on a single reasoning dataset Ling et al. (2017)
- explored models of smaller scale e-SNLI (2018)

Flan-PaLM: finetuning mixture (inc. CoT) helps unseen reasoning tasks

Also related, finetuning on self-generated CoT datasets LMSI (2022)

Flan-PaLM: finetune jointly on both CoT and non-CoT data

Compute-efficient LM improvements

Scaling LMs brings gains but is costly Scaling laws (2020)

Improvements via compute-efficient alternatives

Chinchilla (2022) Task Transferability (2022) UL2R (2022)

Flan-U-PaLM shows that UL2R is complementary

Other improvements:

- architectures Primer (2021)
- training objectives UL2 (2022)
- better data GLAM (2022)

References:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(InstructGPT) L. Ouyang et al., "Training language models to follow instructions with human feedback", arxiv (2022)
(FLAN) J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022)
(T0) V. Sanh et al., "Multitask Prompted Training Enables Zero-Shot Task Generalization", ICLR (2022)
L. Ling et al., "Program induction by rationale generation: Learning to solve and explain algebraic word problems", ACL (2017)
(e-SNLI) O-M. Camburu et al., "e-snli: Natural language inference with natural language explanations", NeurIPS (2018)
(LMSI) J. Huang et al., "Large Language Models Can Self-Improve", arxiv (2022)
J. Kaplan et al., "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020)
(Chinchilla) J. Hoffmann et al., "Training Compute-Optimal Large Language Models", arxiv (2022)
V. Padmakumar et al., "Exploring the Role of Task Transferability in Large-Scale Multi-Task Learning", arxiv (2022)
(UL2R) Y. Tay et al., "Transcending Scaling Laws with 0.1% Extra Compute", arxiv (2022)
(Primer) D. So et al., "Searching for Efficient Transformers for Language Modeling", NeurIPS (2021)
(UL2) Y. Tay et al., "Unifying Language Learning Paradigms", arxiv (2022)
(GLAM) N. Du et al., "Glam: Efficient scaling of language models with mixture-of-experts", ICML (2022)

Key findings

Instruction finetuning benefits from model scale and increasing the number of tasks

- Scaling number of tasks brings diminishing gains, but finetuning (vs not) is effective at all scales

Instruction finetuning on joint CoT and non-CoT data brings substantial benefits for reasoning tasks

Instruction finetuning generalises across scales, model families (e.g. T5, PaLM) & objectives (UL2R)

Instruction finetuning improves usability (as assessed by human annotators)

Instruction finetuning is efficient (e.g. 0.2% of total compute for PaLM 540B for +9.4% gain)

⟹It is likely that instruction finetuning will be broadly useful for pretrained language models

References:

H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(T5) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2020)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)
(UL2R) Y. Tay et al., "Transcending Scaling Laws with 0.1% Extra Compute", arxiv (2022)

Flan finetuning: nuts and bolts

Each model uses the same hyperparameters except learning rate | batch size | dropout | finetuning steps

Learning rate schedule is constant

Finetuning is performed with Adafactor

Packing (like T5) combines examples into a single sequence

<EOS> token used to separate inputs from targets

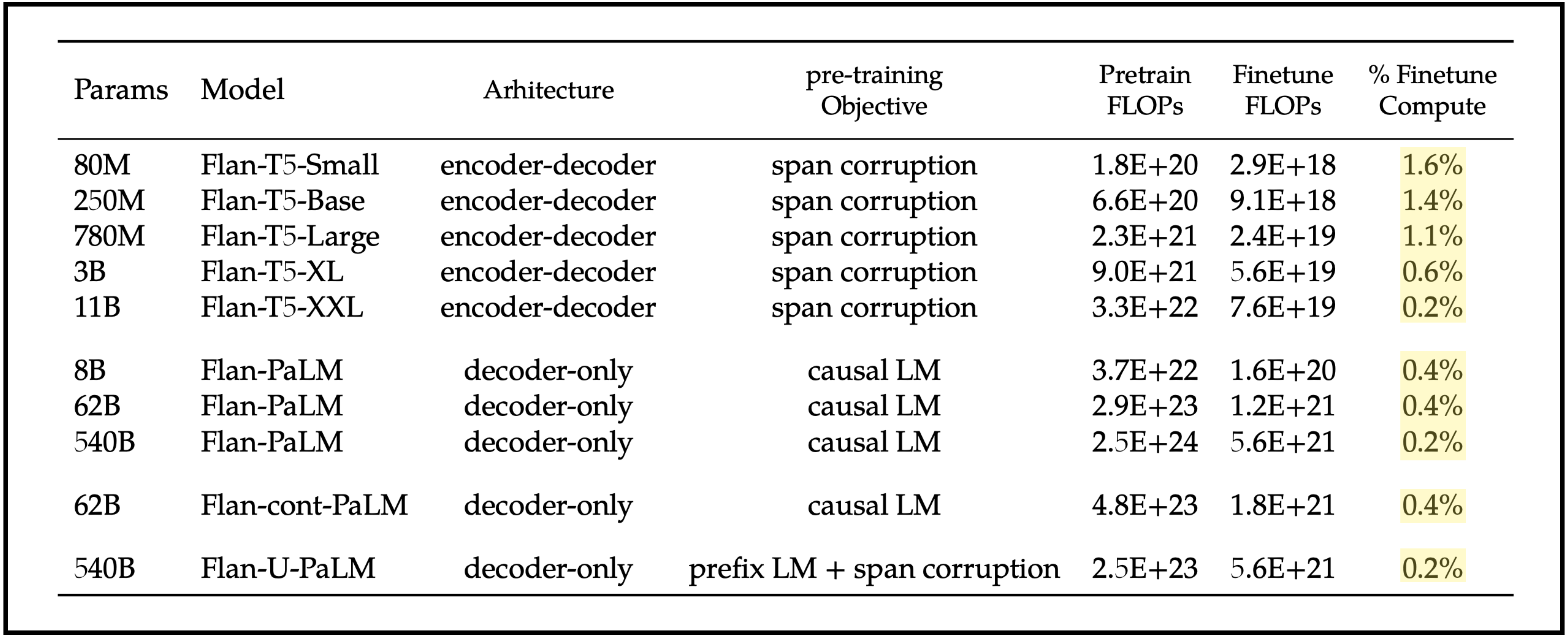
Masking used to prevent tokens attending across boundaries in the pack

For each model, one checkpoint is used for all evaluations

The implementation uses the JAX-based T5X framework

References:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(Adafactor) N. Shazeer et al., "Adafactor: Adaptive learning rates with sublinear memory cost." ICML (2018)
(T5) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2020)
(T5X) A. Roberts, "Scaling Up Models and Data with t5x and seqio", arxiv (2022)

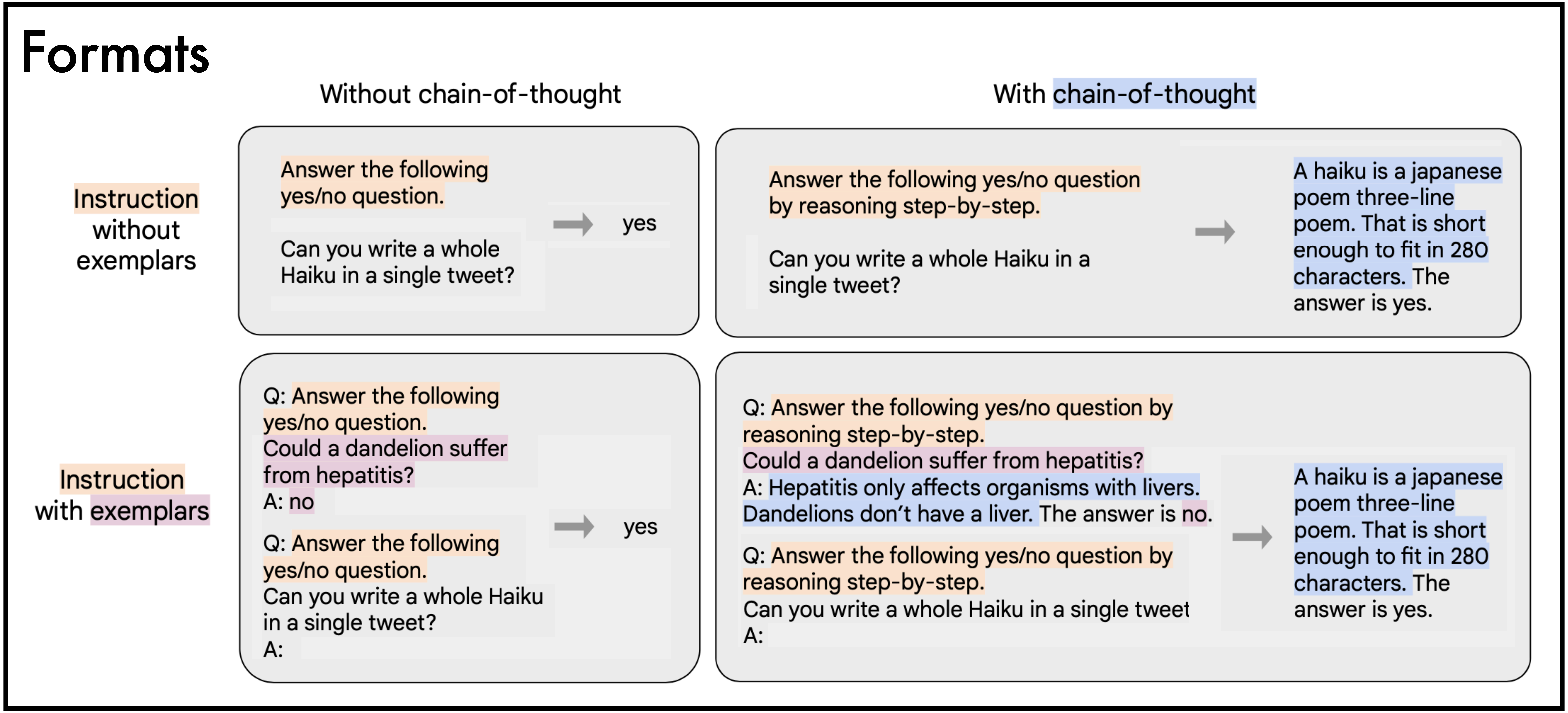
Flan finetuning architecture configurations



Params	Model	Arhitecture	pre-training Objective	Pretrain FLOPs	Finetune FLOPs	% Finetune Compute
80M	Flan-T5-Small	encoder-decoder	span corruption	1.8E+20	2.9E+18	1.6%
250M	Flan-T5-Base	encoder-decoder	span corruption	6.6E+20	9.1E+18	1.4%
780M	Flan-T5-Large	encoder-decoder	span corruption	2.3E+21	2.4E+19	1.1%
3B	Flan-T5-XL	encoder-decoder	span corruption	9.0E+21	5.6E+19	0.6%
11B	Flan-T5-XXL	encoder-decoder	span corruption	3.3E+22	7.6E+19	0.2%
8B	Flan-PaLM	decoder-only	causal LM	3.7E+22	1.6E+20	0.4%
62B	Flan-PaLM	decoder-only	causal LM	2.9E+23	1.2E+21	0.4%
540B	Flan-PaLM	decoder-only	causal LM	2.5E+24	5.6E+21	0.2%
62B	Flan-cont-PaLM	decoder-only	causal LM	4.8E+23	1.8E+21	0.4%
540B	Flan-U-PaLM	decoder-only	prefix LM + span corruption	2.5E+23	5.6E+21	0.2%

References:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(T5) C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer", JMLR (2020)
(Flan) J. Wei et al., "Finetuned Language Models are Zero-Shot Learners", ICLR (2022)
(PaLM) A. Chowdhery et al., "Palm: Scaling language modeling with pathways", arxiv (2022)
(U-PaLM) Y. Tay et al., "Transcending Scaling Laws with 0.1% Extra Compute", arxiv (2022)

Finetuning data formats



Formats

	Without chain-of-thought	With chain-of-thought
Instruction without exemplars	Answer the following yes/no question. Can you write a whole Haiku in a single tweet? → yes	Answer the following yes/no question by reasoning step-by-step. Can you write a whole Haiku in a single tweet? → A haiku is a japanese poem three-line poem. That is short enough to fit in 280 characters. The answer is yes.
Instruction with exemplars	Q: Answer the following yes/no question. Could a dandelion suffer from hepatitis? A: no Q: Answer the following yes/no question. Can you write a whole Haiku in a single tweet? A: → yes	Q: Answer the following yes/no question by reasoning step-by-step. Could a dandelion suffer from hepatitis? A: Hepatitis only affects organisms with livers. Dandelions don't have a liver. The answer is no. Q: Answer the following yes/no question by reasoning step-by-step. Can you write a whole Haiku in a single tweet A: → A haiku is a japanese poem three-line poem. That is short enough to fit in 280 characters. The answer is yes.

References/image credits:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)

Evaluation methods and metrics

MMLU BBH evaluate under both direct and Chain-of-Thought (CoT) prompting

TyDiQA only measure direct prompting exact-match score (not passage highlighting)

MGSM only measure CoT prompting (direct prompting scores poorly)

Few-shot configurations follow prior work:

MMLU 5-shot
BBH 3-shot
TyDiQA 1-shot
MGSM 8-shot

A "normalised average" is reported for each model (in the style of BIG-Bench)

The normalised average is the macro-average over six normalised scores:

MMLU-Direct
MMLU-CoT
BBH-Direct
BBH-CoT
TyDiQA-Direct
MGSM-CoT

References:
H. Chung et al., "Scaling Instruction-Finetuned Language Models", arxiv (Oct. 2022)
(MMLU) D. Hendrycks et al., "Measuring Massive Multitask Language Understanding", ICLR (2020)
(BBH) M. Suzgun et al., "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them", arxiv (2022)
(TyDiQA) J. Clark et al., "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages", ACL (2020)
(MGSM) F. Shi et al., "Language Models are Multilingual Chain-of-Thought Reasoners", arxiv (2022)
(Big-Bench) A. Srivastava et al., "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models", arxiv (2022)