



University of
Sheffield



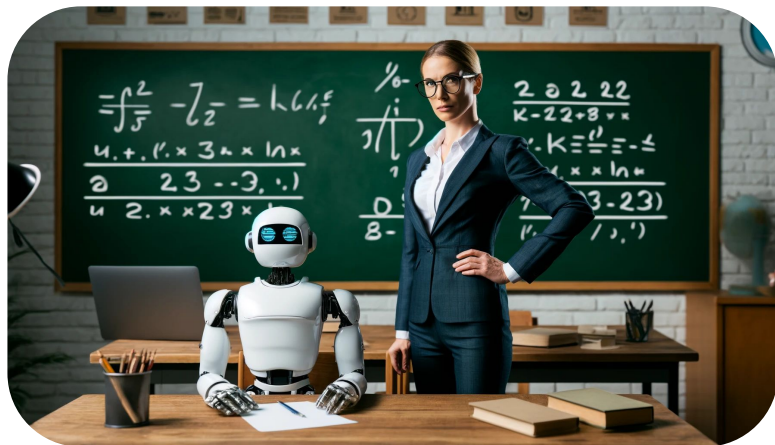
Numbers in the Mist: Towards Accurate Numerical Understanding in LMs

Nafise Sadat Moosavi

Department of Computer Science

LMs and Numerical Reasoning

A LM should be able to do
 $(5464327865 * 7685439872) / 76$



A LM doesn't need to do any math,
give them a calculator



LMs and Numerical Reasoning

Claim:

The average salary of software engineers in New York increased by 30% from 2020 to 2023.

Evidence 1:

The average salary of software engineers in 2020 in New York is \$100,000.

Evidence 2:

The average salary of software engineers in New York in 2023 is \$120,000.

$$((120,000 - 100,000) / 100,000) * 100 = 20\%$$



LMs and Numerical Reasoning

A recent survey conducted in 2023 found that 65% of the 1,000 respondents preferred remote work over in-office work. In a similar survey from 2020, 50% of the 800 respondents expressed a preference for remote work. The increase in preference for remote work can be attributed to several factors, including the increased flexibility and the reduction in commuting time. **The survey also revealed that the average weekly commuting time saved by remote workers was 10 hours in 2023, compared to 8 hours in 2020.**



Between 2020 and 2023, the preference for remote work increased from 50% to 65%. Additionally, remote workers saved **2 more hours on average** in weekly commuting time in 2023 compared to 2020.

LMs and Numerical Reasoning

A company has sales data for the first quarter (Q1) of 2023 for its three main products:

- **Product A:** 3,000 units sold, \$150,000 revenue
- **Product B:** 2,500 units sold, \$200,000 revenue
- **Product C:** 1,500 units sold, \$90,000 revenue

Product B with \$80 per unit

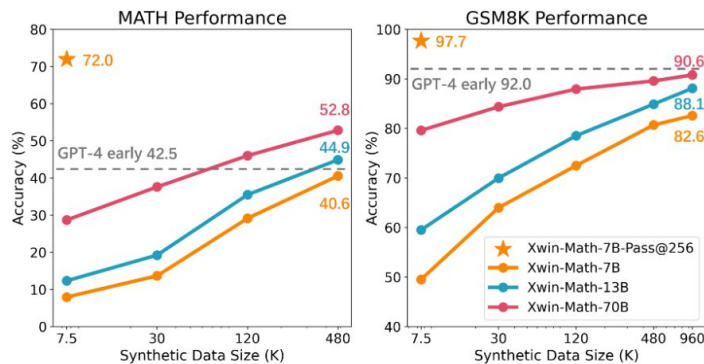
Which product performed the best in terms of revenue per unit?



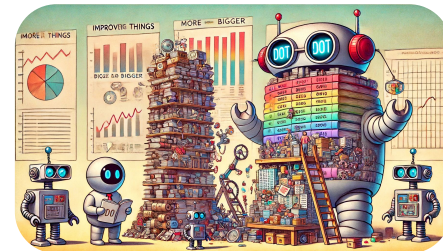
Improving Numerical Reasoning

More data

Larger models



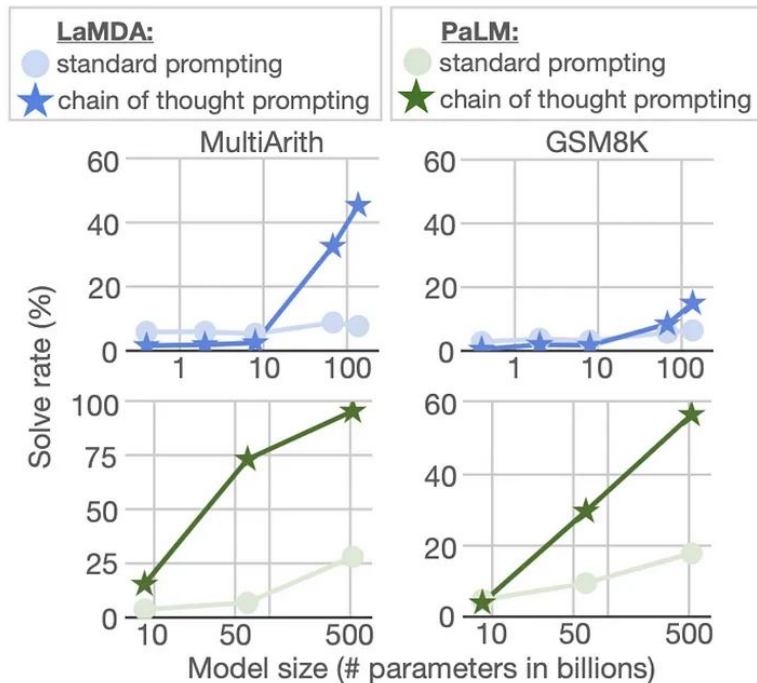
Li et al., 2024. [Common 7B Language Models Already Possess Strong Math Capabilities](#)



Improving Numerical Reasoning

More data

Larger models



Improving Numerical Reasoning



Hugging Face

How about smaller models??

Models 30,460

Filter by name

Full-text search

11 Sort: Most downloads

google/flan-t5-large

Text2Text Generation · Updated Jul 17, 2023 · ± 6.82M · ♥ 479

vennify/t5-base-grammar-correction

Text2Text Generation · Updated Jan 14, 2022 · ± 1.46M · ♥ 138

unikei/t5-base-split-and-rephrase

Text2Text Generation · Updated Jan 26 · ± 1.43M · ♥ 14

prithivida/parrot_paraphraser_on_T5

Text2Text Generation · Updated May 18, 2021 · ± 1.34M · ♥ 135

google/flan-t5-base

Text2Text Generation · Updated Jul 17, 2023 · ± 1.29M · ♥ 734

facebook/m2m100_418M

Text2Text Generation · Updated Feb 29 · ± 1.14M · ♥ 221

Rostlab/prot_t5_xl_uniref50

Text2Text Generation · Updated Jan 31, 2023 · ± 730k · ♥ 38

ybelkada/tiny-random-T5ForConditionalGeneration-calibrated

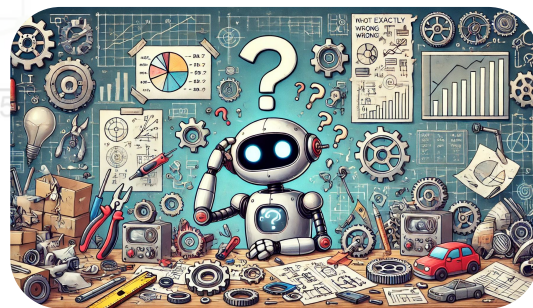
Text2Text Generation · Updated Apr 5, 2023 · ± 710k

lmsys/fastchat-t5-3b-v1.0

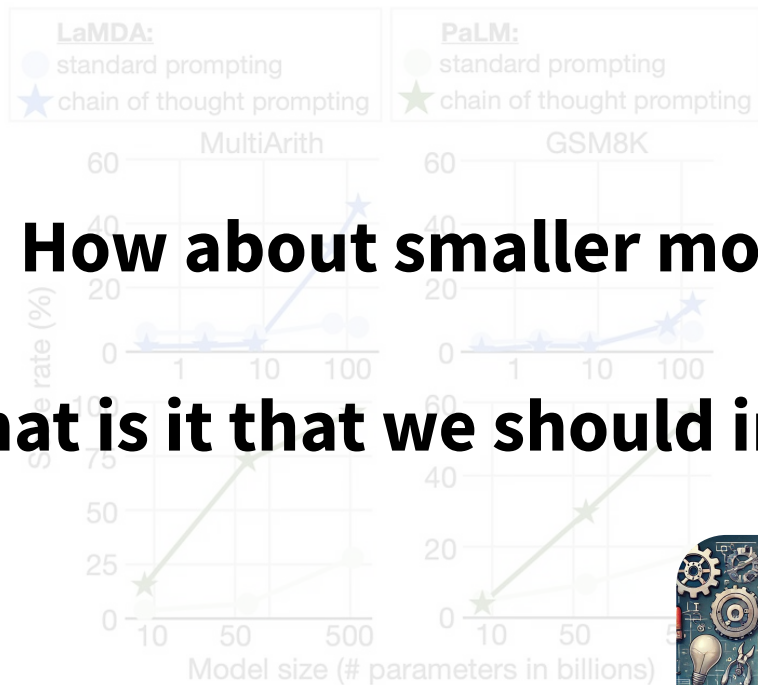
Text2Text Generation · Updated Jun 29, 2023 · ± 692k · ♥ 351

google/byt5-small

Text2Text Generation · Updated Jan 24, 2023 · ± 500k · ♥ 44

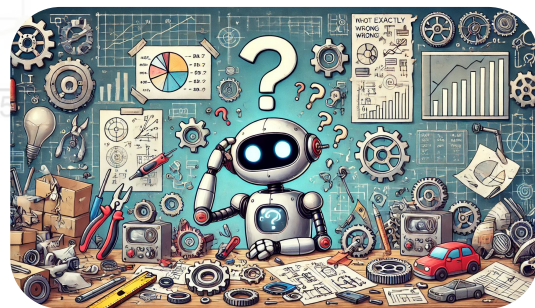


Improving Numerical Reasoning

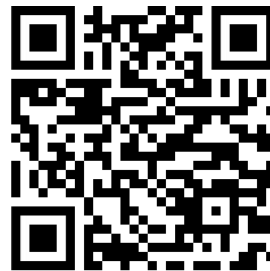


How about smaller models??

What is it that we should improve??




FERMAT: An Alternative to Accuracy for Numerical Reasoning



FERMAT

Flexible Evaluation set for
Representing Multi-views
of Arithmetic Types

- 
- Understanding numbers
 - Inferring the underlying equation
 - Performing math operations



Jasivan Sivakumar

Number Understanding

Keep the context the same, change numbers

A Euro is **5** yens. How much is **25** Euros?

Number Understanding

A Euro is **5** yens. How much is **25** Euros?

- Same numbers different formatting
 - A Euro is **five** yens. How much is **twenty five** Euros?
 - A Euro is **5.0** yens. How much is **25.0** Euros?
- Commuted
 - A Euro is **25** yens. How much is **5** Euros?

Number Understanding

A Euro is **5** yens. How much is **25** Euros?

- Same digits different numbers
 - A Euro is **0.5** yens. How much is **2.5** Euros?
 - A Euro is **5000** yens. How much is **2500** Euros?

Number Understanding

A Euro is **5** yens. How much is **25** Euros?

- Different number ranges
 - 2, 3, or 4 digit integers
 - A Euro is 886 yens. How much is 621 Euros?
 - Integers less than 1000
 - A Euro is **319** yens. How much is **26** Euros?
 - Integers greater than 1000
 - A Euro is **2132** yens. How much is **8146** Euros?
 - Decimals
 - A Euro is **73.9** yens. How much is **9.4** Euros?

Training Dependency

- **Exact:** all the numbers and operations are seen during finetuning
 - A Euro is 5 yens. How much is 25 Euros?
 - If you have 5 packs of cookies and each pack contains 25 cookies, how many cookies do you have in total?
- **All Numbers:** all the numbers are seen
- **Number & Operation:** at least one number and operation
- **One Number**

Training Dependency

- **Exact:** all the numbers and operations are seen during finetuning
 - A Euro is 5 yens. How much is 25 Euros?
 - If you have 5 packs of cookies and each pack contains 25 cookies, how many cookies do you have in total?



Inferring the underlying equation

Mathematical Operations

Hops	Expression	Frequency
One-hop	$a + b$	154
	$a - b$	162
	$a \times b$	113
	$a \div b$	102
Two-hop	$(a + b) - c$	190
	$a \times (b + c)$	100
	$(a + b) \div c$	90
	$a \times (b - c)$	100
	$(a - b) \div c$	100
Total		1111

Understanding Numbers

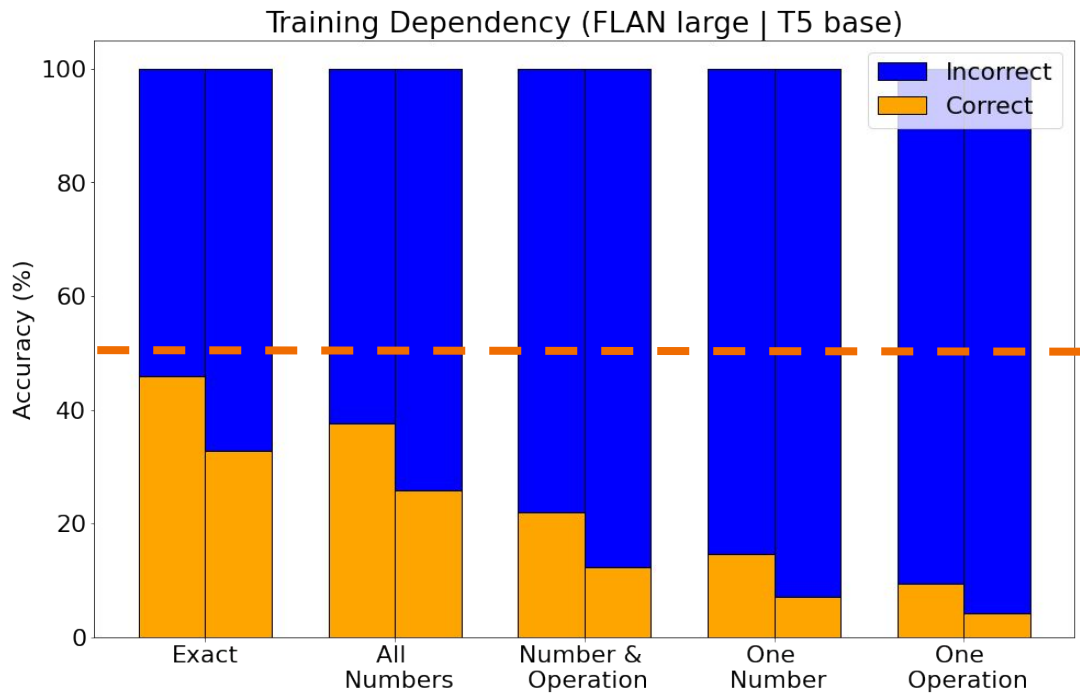
Models (size)		Number Understanding																		
		Alternate Representations											Range of numbers							
		Same numbers					Same digits					Grouping		Integers					Decimals	
		Original	Fixed 1dp	Fixed 2dp	Worded	Commuted	Original 1dp	Original 2dp	Original 1dp no 0	Original 2dp no 0	Original 1000+	1000+ comma	1000+ space	1000+ random	Integers 0 to 1000	2 digit	3 digit	4 digit	1dp random	2dp random
Zero-shot	T0 (3B)	2.88	1.98	2.79	0.39	3.49	3.99	1.29	1.47	3.33	0.93	0.00	0.09	0.09	0.18	0.75	0.12	0.06	2.04	0.27
	FLAN XL (3B)	22.86	10.44	14.52	20.13	18.28	6.66	3.57	3.69	5.79	5.28	0.00	0.00	0.00	0.45	4.83	0.33	0.00	4.08	0.33
	Bhaskara (2.7B)	23.18	21.60	20.88	18.23	18.49	5.31	3.65	3.87	4.55	4.05	0.00	0.00	0.00	0.18	3.56	0.18	0.18	1.31	0.14
	FLAN large (770M)	11.79	4.71	6.27	10.26	11.24	3.99	1.65	3.51	2.07	2.46	0.00	0.00	0.03	0.12	1.56	0.24	0.03	2.04	0.54
	FLAN base (220M)	4.98	1.95	3.90	3.69	3.98	2.88	1.83	3.48	2.22	0.93	0.00	0.00	0.00	0.18	0.90	0.33	0.00	0.54	0.12
	T5 base (220M)	1.71	2.70	1.62	0.00	2.13	2.34	0.99	2.07	1.62	0.99	0.00	0.00	0.00	0.09	0.36	0.00	0.00	0.81	0.27
	BART base (140M)	2.79	2.79	2.88	0.00	2.46	2.25	0.99	2.88	1.89	0.81	0.00	0.09	0.09	0.00	0.27	0.00	0.00	0.81	0.18
	NT5 (3M)	8.19	8.10	8.10	2.84	5.97	6.71	4.95	4.64	2.48	7.25	0.95	0.00	0.00	6.39	7.52	6.89	6.21	5.00	2.21

Understanding Numbers

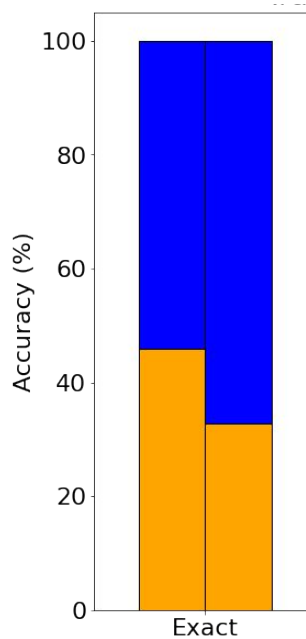
Models (size)		Number Understanding																		
		Alternate Representations											Range of numbers							
		Same numbers					Same digits					Grouping		Integers				Decimals		
		Original	Fixed 1dp	Fixed 2dp	Worded	Commuted	Original 1dp	Original 2dp	Original 1dp no 0	Original 2dp no 0	Original 1000+	1000+ comma	1000+ space	1000+ random	Integers 0 to 1000	2 digit	3 digit	4 digit	1dp random	2dp random
Zero-shot	T0 (3B)	2.88	1.98	2.79	0.39	3.49	3.99	1.29	1.47	3.33	0.93	0.00	0.09	0.09	0.18	0.75	0.12	0.06	2.04	0.27
	FLAN XL (3B)	22.86	10.44	14.52	20.13	18.28	6.66	3.57	3.69	5.79	5.28	0.00	0.00	0.00	0.45	4.83	0.33	0.00	4.08	0.33
	Bhaskara (2.7B)	23.18	21.60	20.88	18.23	18.49	5.31	3.65	3.87	4.55	4.05	0.00	0.00	0.00	0.18	3.56	0.18	0.18	1.31	0.14
	FLAN large (770M)	11.79	4.71	6.27	10.26	11.24	3.99	1.65	3.51	2.07	2.46	0.00	0.00	0.03	0.12	1.56	0.24	0.03	2.04	0.54
	FLAN base (220M)	4.98	1.95	3.90	3.69	3.98	2.88	1.83	3.48	2.22	0.93	0.00	0.00	0.00	0.18	0.90	0.33	0.00	0.54	0.12
	T5 base (220M)	1.71	2.70	1.62	0.00	2.13	2.34	0.99	2.07	1.62	0.99	0.00	0.00	0.00	0.09	0.36	0.00	0.00	0.81	0.27
	BART base (140M)	2.79	2.79	2.88	0.00	2.46	2.25	0.99	2.88	1.89	0.81	0.00	0.09	0.09	0.00	0.27	0.00	0.00	0.81	0.18
	NT5 (3M)	8.19	8.10	8.10	2.84	5.97	6.71	4.95	4.64	2.48	7.25	0.95	0.00	0.00	6.39	7.52	6.89	6.21	5.00	2.21
Fine-tuned	FLAN large (770M)	28.80	29.79	30.33	8.91	26.02	33.93	29.70	25.20	32.13	18.90	0.00	0.00	5.76	17.01	24.12	15.57	10.98	25.65	13.86
	FLAN base (220M)	26.55	27.63	27.09	6.84	19.64	29.79	27.18	19.44	26.55	15.39	0.00	0.09	6.30	15.48	21.87	15.39	11.43	24.84	15.75
	T5 base (220M)	19.44	21.24	20.34	6.39	16.53	20.88	14.31	10.17	16.02	7.65	0.00	1.17	1.89	7.29	14.76	8.91	4.23	15.84	6.84
	BART base (140M)	18.63	21.24	21.24	0.90	14.89	23.04	18.18	17.28	3.51	10.35	0.00	0.00	5.76	13.68	15.57	12.69	9.18	17.64	10.98
	NT5 (3M)	14.04	15.12	14.49	3.06	12.44	16.11	13.41	13.59	8.73	8.55	0.63	5.04	5.04	13.77	14.85	13.68	8.73	15.03	10.71

200K examples from 100 templates written by math teachers

Training Dependency



Inferring the Underlying Equation



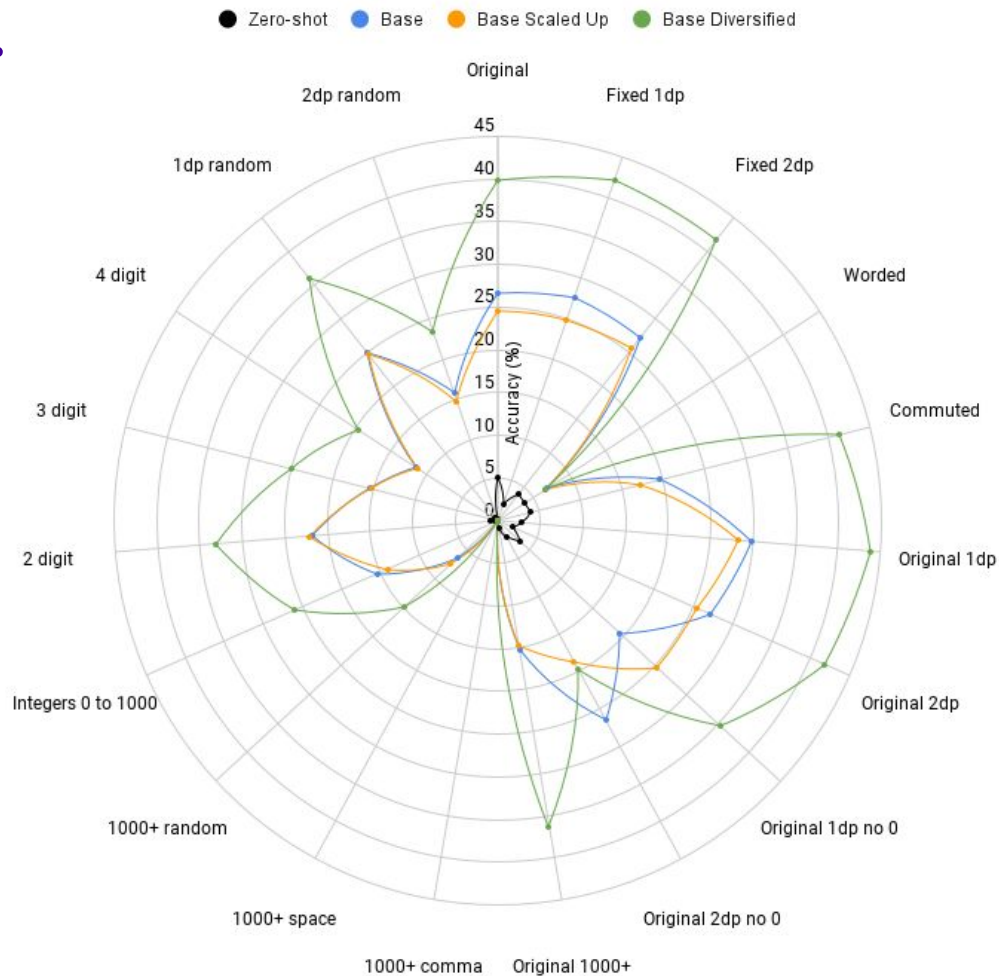
Performing Math Operations

Models (size)		Mathematical Operations									
		One-hop				Two-hop					
		a+b	a-b	a*b	a/b	(a+b)-c	a*(b+c)	(a+b)/c	a*(b-c)	(a-b)/c	
Zero-shot	T0 (3B)	1.18	0.66	0.70	1.90	2.14	1.33	2.43	0.37	1.49	
	FLAN XL (3B)	8.61	7.66	11.65	7.92	2.73	3.42	4.25	3.63	4.98	
	Bhaskara (2.7B)	8.84	7.24	8.72	9.61	2.10	5.00	8.55	2.89	8.31	
	FLAN large (770M)	3.88	1.81	6.91	3.83	1.80	1.77	4.51	1.63	2.89	
	FLAN base (220M)	0.88	1.18	1.94	3.10	1.41	1.25	3.17	0.54	2.15	
	T5 base (220M)	0.03	1.23	0.38	1.81	1.23	0.57	1.81	0.35	1.45	
	BART base (140M)	0.03	1.70	0.38	2.35	1.06	0.33	2.93	0.50	2.65	
	NT5 (3M)	11.17	17.85	0.63	2.45	1.19	0.68	1.93	0.31	1.08	
Fine-tuned	FLAN large (770M)	22.85	23.40	23.57	22.60	13.89	14.34	17.76	10.10	20.05	
	FLAN base (220M)	30.22	29.75	18.76	21.47	7.83	5.81	7.54	5.45	19.10	
	T5 base (220M)	12.53	13.27	23.41	14.46	7.13	4.56	3.13	7.35	9.45	
	BART base (140M)	26.00	17.84	13.98	11.67	2.41	2.56	9.92	5.90	11.25	
	NT5 (3M)	34.39	31.24	0.84	1.09	7.13	0.53	0.06	0.72	0.22	

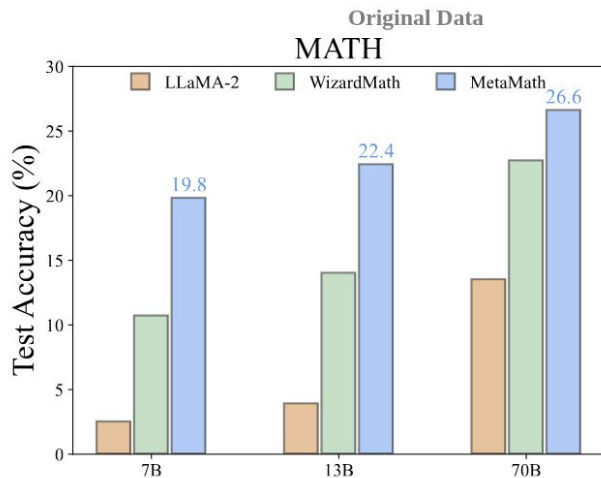
Enhancing Inference of

Data Augmentation

- Zero-shot
- Base (200k)
- Base scaled (200k+100k)
- Base diversified (200k+100k)



Enhancing Inference of Equations



Meta-Question: James buys 5 packs of beef that are 4 pounds each. The price of beef is \$5.50 per pound. How much did he pay?

Answer: He bought $5 \times 4 = 20$ pounds of beef. So he paid $20 \times 5.5 = \$110$. The answer is: 110



Question Bootstrapping

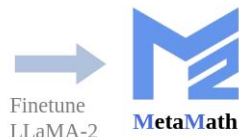
Rephrasing Question: What is the total amount that James paid when he purchased 5 packs of beef, each weighing 4 pounds, at a price of \$5.50 per pound? **Answer:**

Self-Verification Question: James buys x packs of beef that are 4 pounds each. The price of beef is \$5.50 per pound. He paid 110. *What is the value of unknown variable x ?* **Answer:**

FOBAR Question: James buys x packs of beef that are 4 pounds each. The price of beef is \$5.50 per pound. How much did he pay? *If we know the answer to the above question is 110, what is the value of unknown variable x ?* **Answer:**

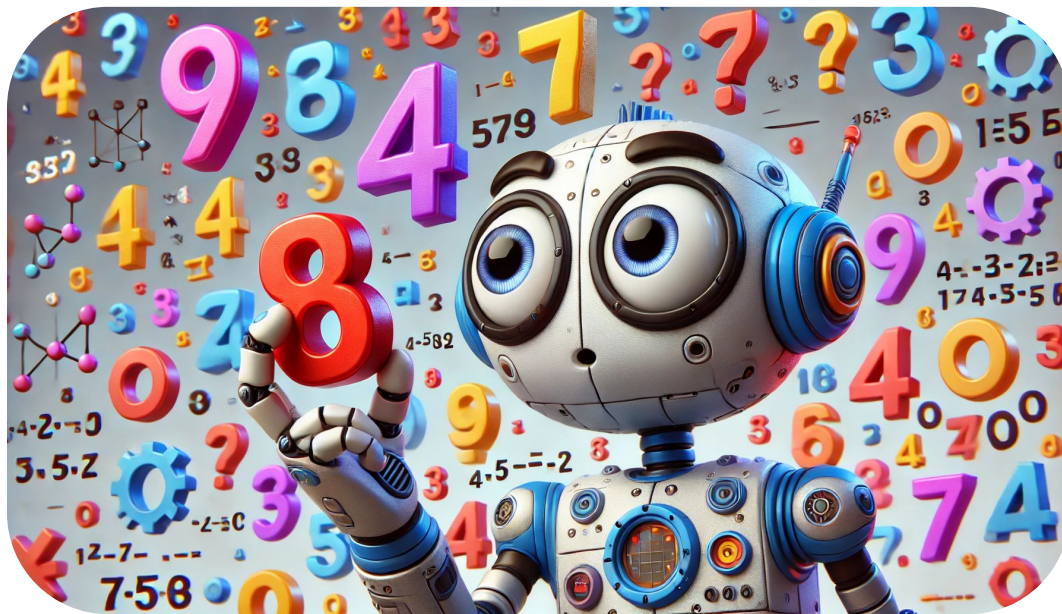
Answer Augment: James buys 5 packs of beef that are 4 pounds each, so he buys a total of $5 \times 4 = 20$ pounds of beef. The price of beef is \$5.50 per pound, so he pays $20 \times \$5.50 = \110 . The answer is: 110

MetaMathQA



Yu et al., 2024. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models

Enhancing Number Understanding



Tokenization

- Traditional tokenization algorithms are optimized for common words
- Numbers are not processed the same way as common words

1234  1234 infeasible, infinite vocab

1234  12, 34 meaningless

1235  1, 235

Tokenization

Digit tokenization

1234  1234 infeasible, infinite vocab

1234  12, 34 meaningless

1234  1, 2, 3, 4

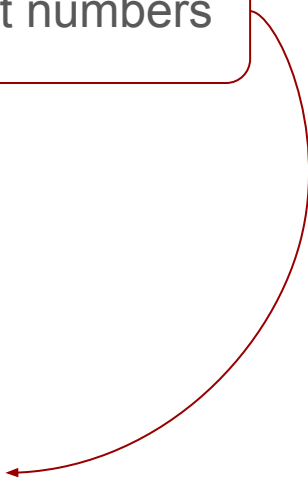
Digit Tokenization

- ✖ Model may have already learned good embeddings for frequent numbers
- ✖ Aggregating the overall embedding of a number given its digits
 - Infinite combination of digits into integers, unlike common words

Digit Tokenization

- ✗ Model may have already learned good embeddings for frequent numbers
- ✗ Aggregating the overall embedding of a number given its digits
 - Infinite combination of digits into integers, unlike common words

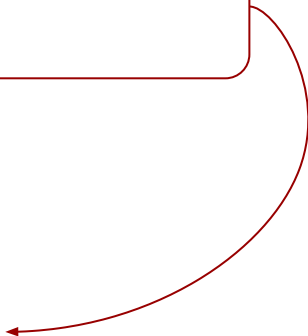
Can we benefit from combining BPE and digit tokenization?



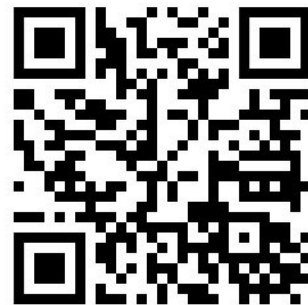
Digit Tokenization

- ✖ Model may have already learned good embeddings for frequent numbers
- ✖ Aggregating the overall embedding of a number given its digits
 - Infinite combination of digits into integers, unlike common words

Would explicitly incorporating a mathematically sound aggregation method benefit numerical reasoning?



Arithmetic-Based Pretraining: Improving Numeracy of Pretrained LMs



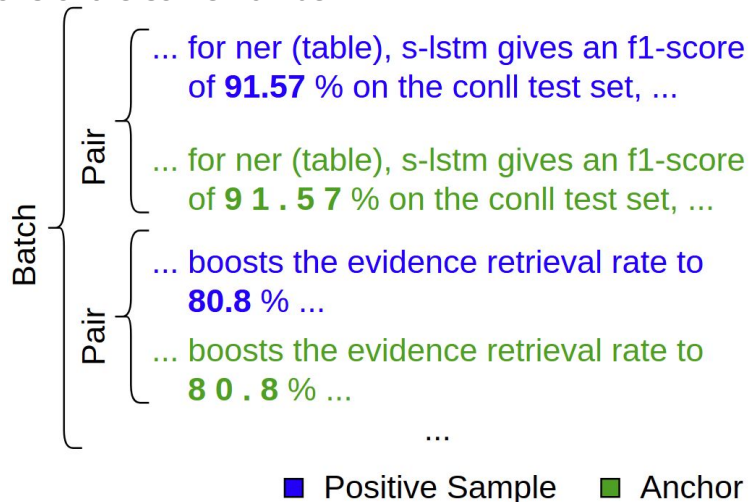
Dominic Petrak



Iryna Gurevych

Contrastive Loss

- Using different tokenization algorithms
 - Byte-pair encoding, Character-level embeddings
- Using contrastive learning
 - Learning a similar representation for different tokenizations of the same number



Extended Pretraining

- ✓ An extended pretraining step focusing on arithmetic reasoning
 - The Inferable Number Prediction Task

Inferable Number Prediction

DROP

<s> He lied on the ground, motionless, for about 7 minutes before he was taken off the field on a cart. Dallas lead 12-10 with under 2 minutes to go. Dallas tried to come back, but Seattle forced a turnover on downs to end the game. </s> With less than <mask> minutes to go, how many points ahead was Dallas? </s>

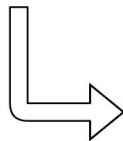


With less than **2** minutes to go, how many points ahead was Dallas?

Model	F1 Score	Accuracy
Our Approach	76.58	88.55
Baseline	65.78	74.32

SciGen

<s> <R> <C> Model <C> F1 Score <C> Accuracy <R> <C> Our Approach <C> 76.58 <C> 88.55 <R> <C> Their Approach <C> 65.78 <C> 74.32 <CAP> Comparison between us and them. </s> Our approach achieves an F1 score <mask> points higher than their approach. </s>



Our approach achieves an F1 score **10.8** points higher than their approach.

Extended Pretraining

Combining the contrastive loss and the Inferable Number Prediction Task

$$\mathcal{L} = \frac{\mathcal{L}_C}{2} + \frac{\mathcal{L}_{INP}}{2}$$

Evaluation

- Tasks
 - Reading comprehension (DROP)

Reasoning	Passage (some parts shortened)	Question	Answer
Subtraction (28.8%)	That year, his Untitled (1981) , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In 1517, the seventeen-year-old King sailed to Castile. There, his Flemish court In May 1518, Charles traveled to Barcelona in Aragon.	Where did Charles travel to first, Castile or Barcelona?	Castile
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on 2 March 1992. The JNA formed a battlegroup to counterattack the next day.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal , yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. ... Carolina closed out the half with Kasay nailing a 44-yard field goal. ... In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.	Which kicker kicked the most field goals?	John Kasay
Coreference Resolution (3.7%)	James Douglas was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before 1543 he married Elizabeth , daughter of James Douglas, 3rd Earl of Morton. In 1553 James Douglas succeeded to the title and estates of his father-in-law.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10
Other Arithmetic (3.2%)	Although the movement initially gathered some 60,000 adherents , the subsequent establishment of the Bulgarian Exarchate reduced their number by some 75%.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000

Evaluation

- Tasks
 - Inference-On-Tables (InfoTabs)

Dressage	
Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

H1: Dressage was introduced in the Olympic games in 1912.

H2: Both men and women compete in the equestrian sport of Dressage.

H3: A dressage athlete can participate in both individual and team events.

H4: FEI governs dressage only in the U.S.

Figure 1: A semi-structured premise (the table). Two hypotheses (H1, H2) are entailed by it, H3 is neither entailed nor contradictory, and H4 is a contradiction.

Evaluation

- Tasks
 - Data-to-text (SciGen, WikiBIO)

	in-domain	out-of-domain		
	MultiNLI	SNLI	Glockner	SICK
MQAN	72.30	60.91	41.82	53.95
+ coverage	73.84	65.38	78.69	54.55
ESIM (ELMO)	80.04	68.70	60.21	51.37
+ coverage	80.38	70.05	67.47	52.65

Table 2: Impact of using coverage for improving generalization across different datasets of the same task (NLI). All models are trained on MultiNLI.

Table 2 shows the performance for both systems for in-domain (the MultiNLI development set) as well as out-of-domain evaluations on SNLI, Glockner, and SICK datasets.

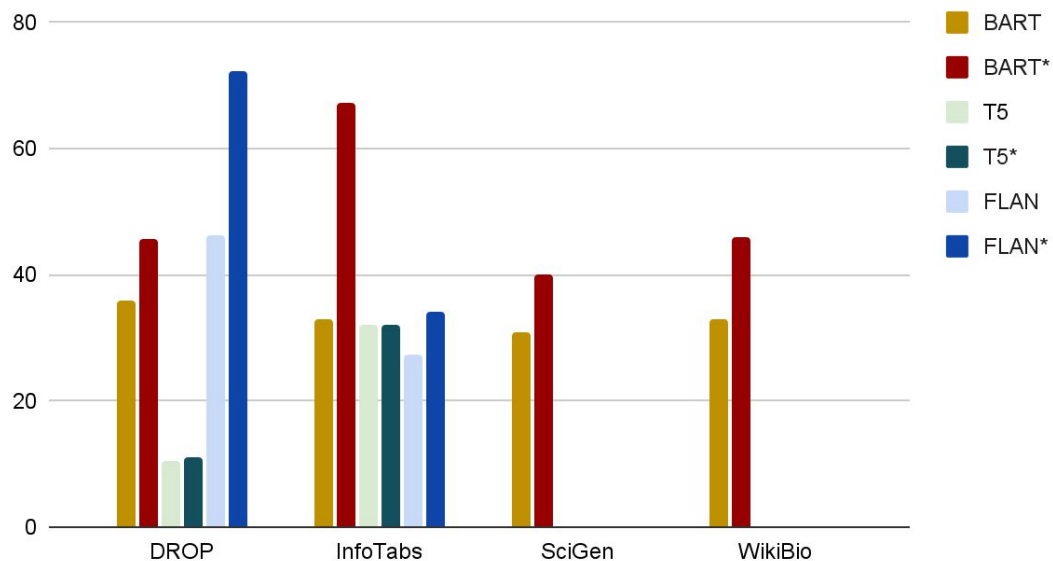
The results show that coverage information considerably improves the generalization of both examined models across various NLI datasets. The resulting cross-dataset improvements on the SNLI and Glockner datasets are larger than those on the SICK dataset. The reason is that the dataset creation process and therefore, the task formulation is similar in SNLI and MultiNLI, but they are different from SICK. In particular, in the neutral pairs

Evaluation

- Models
 - BART-large (406M)
 - T5-base (220M)
 - FLAN-T5 base (220M)

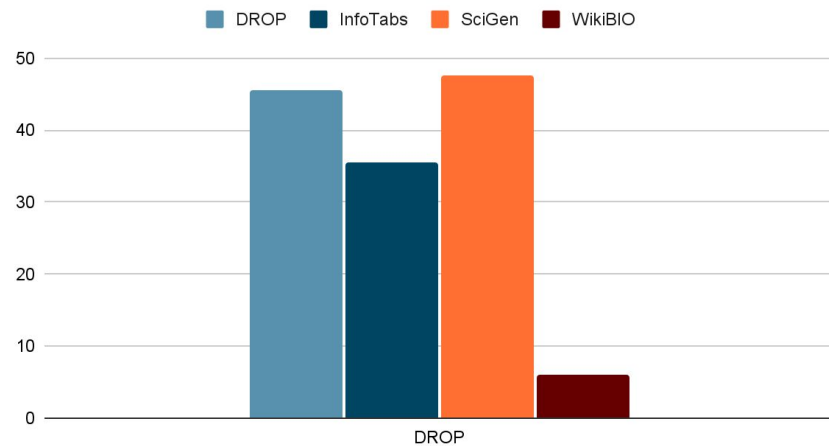
Results

Downstream performance

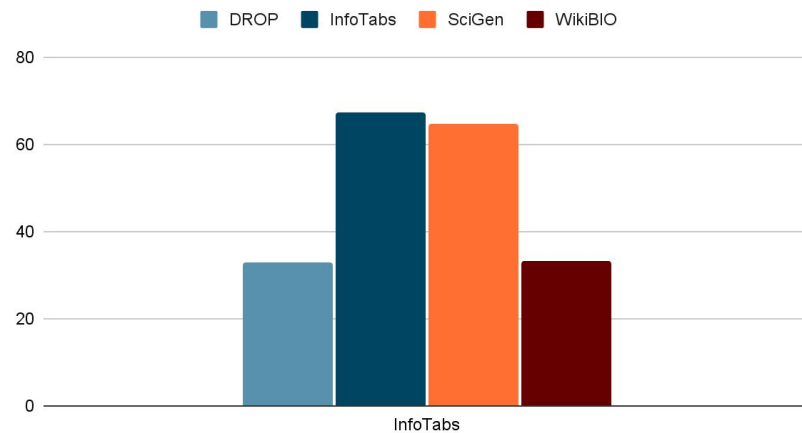


Out-of-domain Pretraining

EM



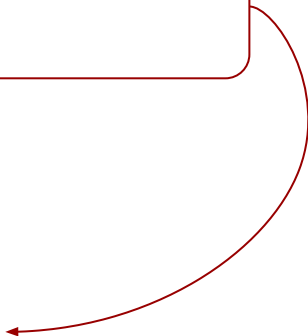
EM



Digit Tokenization

- ✗ Model may have already learned good embeddings for frequent numbers
- ✗ Aggregating the overall embedding of a number given its digits
 - Infinite combination of digits into integers, unlike common words

Would explicitly incorporating a mathematically sound aggregation method benefit numerical reasoning?



How to Leverage Digit Embeddings to Represent Numbers?

247 [F] 2 4 7 [/F]

247 [F] AGG 2 4 7 [/F]



Jasivan Sivakumar

Aggregation

Weighted sum of digits embeddings $\sum w_i \cdot d_i$

1. Reserve embeddings of single-digit numbers
2. Increasing weights based on digit positions

$$2^2 * 2 \quad 2^1 * 4 \quad 2^0 * 7 \quad \sim \quad 100*2 + 10*4 + 7$$

Aggregation

Weighted sum of digits embeddings $\sum w_i \cdot d_i$

1. Reserve embeddings of single-digit numbers
2. Increasing weights based on digit positions
3. Regularization

$$w_i = 2^{N-i} \times \frac{3(N+1-i)(N+2-i)}{N(N+1)(N+2)}$$

Normalised triangular number sequence

$$w_i - w_{i-1} = w_0 \times i$$

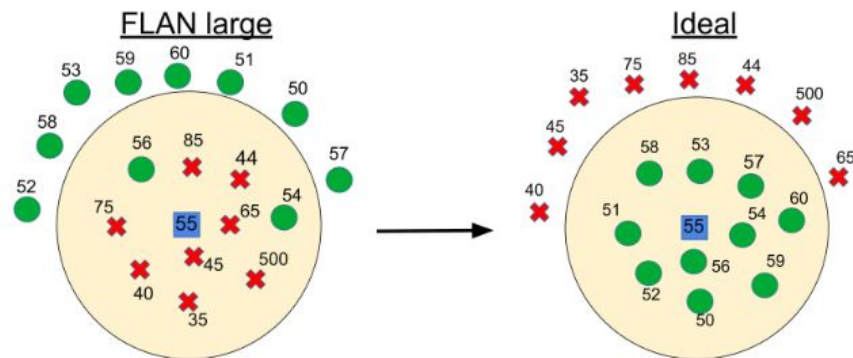
Intrinsic Evaluation

- Distinguishing between distinct numbers
- Mirroring the natural numerical order

Intrinsic Evaluation

F_1 for natural k-Nearest Neighbours vs embedding k-Nearest Neighbours

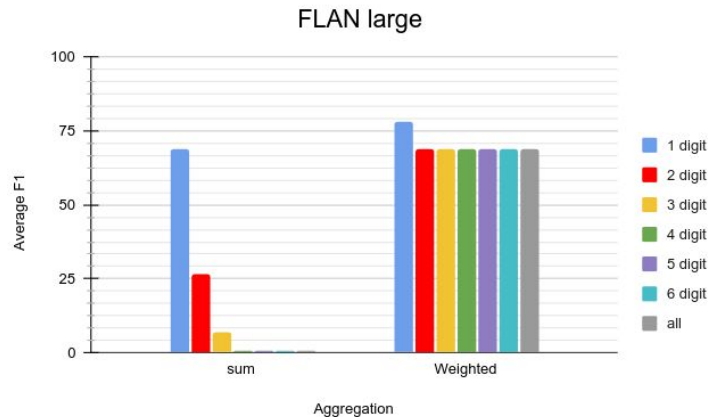
- Distinguishing between distinct numbers
- Mirroring the natural numerical order



Intrinsic Evaluation

F_1 for natural k-Nearest Neighbours vs embedding k-Nearest Neighbours

- Distinguishing between distinct numbers
- Mirroring the natural numerical order



Incorporating Aggregation in LMs

Input embedding

247 [F] 2 4 7 [/F]

247 [F] AGG 2 4 7 [/F]

Output loss

$$\mathcal{L}_{AUX} = \log_2 (\|W(p) - W(l)\|_2)$$

$$\mathcal{L} = \lambda \times \mathcal{L}_{CE} + (1 - \lambda) \times \mathcal{L}_{AUX}$$

Incorporating Aggregation in LMs

Input embedding

247

[F] 2 4 7 [/F]

247

[F] AGG 2 4 7 [/F]

Model- and task-agnostic

Output loss

$$\mathcal{L}_{AUX} = \log_2 (\|W(p) - W(l)\|_2)$$

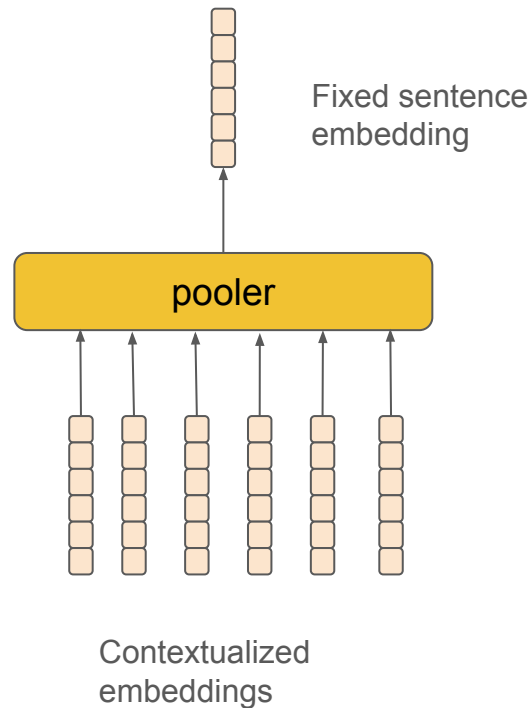
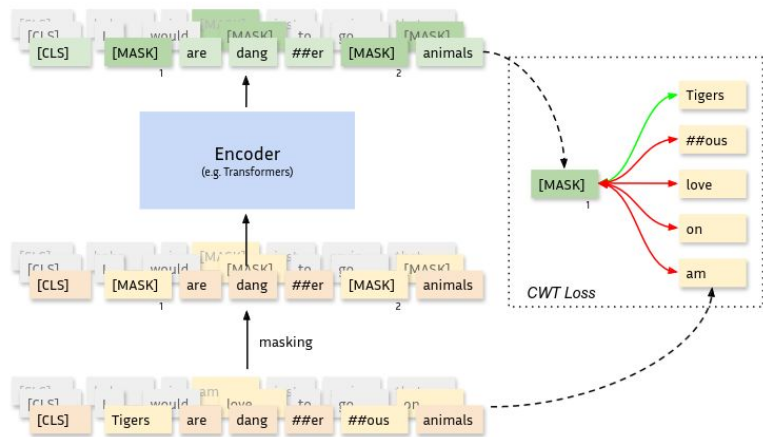
$$\mathcal{L} = \lambda \times \mathcal{L}_{CE} + (1 - \lambda) \times \mathcal{L}_{AUX}$$

Model-agnostic, task-specific

Incorporating Aggregation in LMs

Incorporating Weights (Accuracy %)		MAWPS	FERMAT													
			Original	Commutd	Integers 0 to 1000	2-digit integers	3-digit integers	4-digit integers	1000+	1000+ same	1dp random	2dp random	a+b	a-b	a*b	a/b
BART base (140M)	Digits	19.20	16.65	8.73	10.26	13.41	10.89	7.74	5.58	10.89	17.82	8.37	40.91	10.62	9.56	11.76
	[AGG] + Digits	+2.00	+0.63	+1.53	-1.17	-0.90	-2.16	-0.27	+0.09	+0.09	+1.08	-0.27	-3.90	-0.74	+1.77	0.00
	Digits + Aux Loss	+1.40	+1.89	+1.80	+0.54	+0.81	0.00	+0.81	+1.17	-1.26	+0.18	+0.63	+2.01	+0.19	+4.25	-1.27
FLAN base (250M)	Digits	23.00	28.35	17.82	17.10	22.86	17.37	13.77	10.35	18.72	25.83	18.45	63.38	19.57	12.92	11.27
	[AGG] + Digits	+0.80	+2.79	+0.27	+2.52	+0.81	+1.80	+2.79	+1.80	+0.90	+0.45	-0.09	+4.48	+3.21	-0.27	+1.08
	Digits + Aux Loss	+1.80	+2.25	+0.36	+3.15	+2.16	+1.71	+2.79	+0.81	+3.87	+1.89	-0.18	+3.90	+5.80	+0.27	+1.57
FLAN large (780M)	Digits	28.80	42.39	21.06	25.65	31.32	24.30	21.87	16.47	23.31	36.36	25.83	63.12	39.88	18.23	18.14
	[AGG] + Digits	+1.20	+0.45	+0.45	+0.81	+2.07	+2.79	+0.99	+1.35	+2.88	+0.27	+0.54	+6.17	+3.83	+0.53	+1.47
	Digits + Aux Loss	+1.00	+0.99	-0.18	+1.62	+2.88	+2.79	+0.72	+1.53	+1.26	+1.26	+0.63	-0.39	+1.79	+0.18	-1.08

Other Potential Applications



Godey et al., 2023. [Headless Language Models: Learning without Predicting with Contrastive Weight Tying](#)

Conclusions

- Numerical reasoning remains an open challenge for LMs
- Fine-grained evaluations can better identify future directions
- Improving numeric understanding: open challenge and promising direction
 - Designing better aggregation methods
 - Using more direct signals to enhance the learning of aggregations

Questions?

