

# **Numerical Reasoning in NLP**

**Nafise Sadat Moosavi**

**Department of Computer Science**

# Reasoning in NLP

- Understanding human language requires different reasoning skills
  - Commonsense reasoning, arithmetic reasoning, temporal reasoning, etc

# Reasoning in NLP

- Understanding human language requires different reasoning skills
  - Commonsense reasoning, arithmetic reasoning, temporal reasoning, etc

The XMT model improves the state-of-the-art results on the MNLI dataset by 20 points. The LSTM and ESIM models were the previous top-performing systems on MNLI with the accuracy of 56% and 74%, respectively. This improvement is the result of using an additional pretraining step.

- What is the accuracy of the XMT model on the MNLI dataset?

# Reasoning in NLP

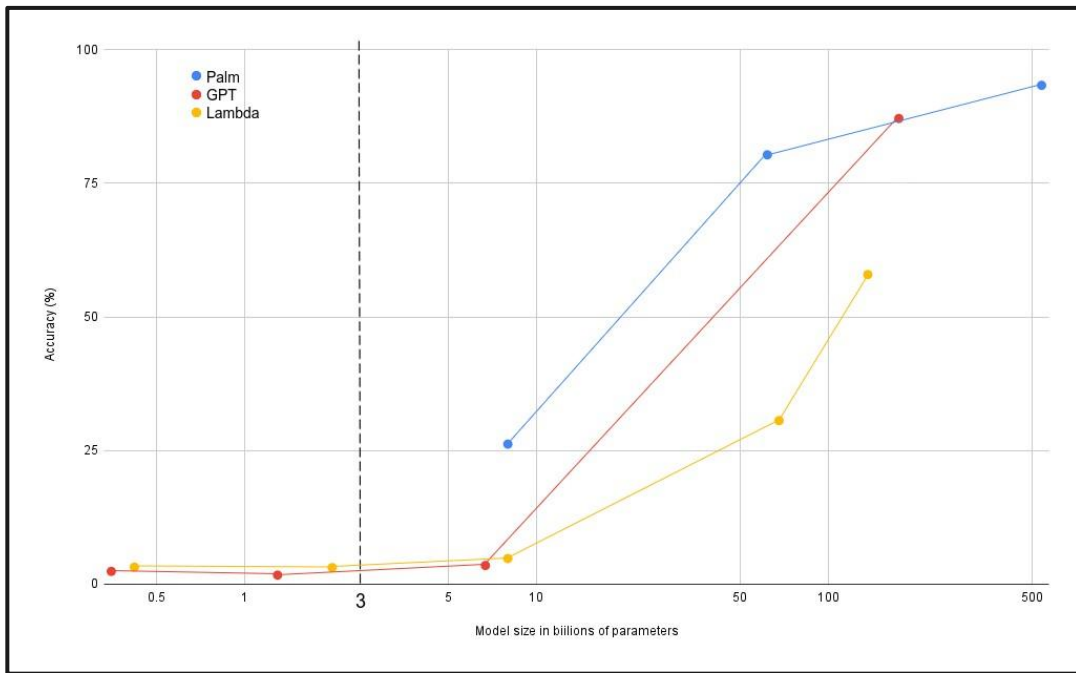
- Understanding human language requires different reasoning skills
  - Commonsense reasoning, arithmetic reasoning, temporal reasoning, etc

The XMT model improves the state-of-the-art results on the MNLI dataset by 20 points. The LSTM and ESIM models were the previous top-performing systems on MNLI with the accuracy of 56% and 74%, respectively. This improvement is the result of using an additional pretraining step.

- What is the accuracy of the XMT model on the MNLI dataset?

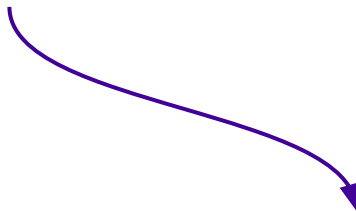
# Numerical Reasoning

- Scaling



# End-to-End Reasoning in Downstream Applications

- Dataset Creation
- Evaluation
- Improvement



with less than enormous models



Creating a dataset for end-to-end reasoning

# SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables



Nafise Sadat Moosavi



Andreas Rücklé



Dan Roth



Iryna Gurevych

# SciGen: Task Definition

	<b>ellipsis (Inflection)</b>	<b>ellipsis (VP)</b>
Baseline	53.0	28.4
concat	<b>76.2</b>	76.6
CADec	72.2	<b>80.0</b>

Caption: Accuracy on ellipsis test set.

**Input:** scientific tables

**Task:** describing findings of the table by performing arithmetic reasoning over its content

# SciGen: Task Definition

	<b>ellipsis (Inflection)</b>	<b>ellipsis (VP)</b>
Baseline	53.0	28.4
concat	<b>76.2</b>	76.6
CADec	72.2	<b>80.0</b>

Caption: Accuracy on ellipsis test set.

For ellipsis, both models improve substantially over the baseline (by 19-51 percentage points), with concat stronger for inflection tasks and CADec stronger for VP ellipsis

# SciGen: Task Definition

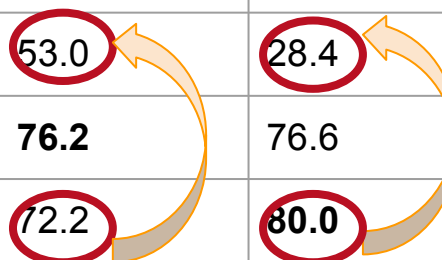
	ellipsis (Inflection)	ellipsis (VP)
Baseline	53.0	28.4
concat	<b>76.2</b>	76.6
CADec	72.2	<b>80.0</b>

For ellipsis, both models improve substantially over the baseline (by 19-51 percentage points), with concat stronger for inflection tasks and CADec stronger for VP ellipsis

Caption: Accuracy on ellipsis test set.

# SciGen: Task Definition

	ellipsis (Inflection)	ellipsis (VP)
Baseline	53.0	28.4
concat	76.2	76.6
CADec	72.2	80.0

A diagram with two curved orange arrows. One arrow starts at the 'Baseline' value of 53.0 in the 'ellipsis (Inflection)' column and points to the 'concat' value of 76.2. The other arrow starts at the 'Baseline' value of 28.4 in the 'ellipsis (VP)' column and points to the 'CADec' value of 80.0. The baseline values 53.0 and 28.4 are circled in red.

For ellipsis, both models improve substantially over the baseline (by 19-51 percentage points), with concat stronger for inflection tasks and CADec stronger for VP ellipsis

Caption: Accuracy on ellipsis test set.

# SciGen: Task Definition

	<b>ellipsis (Inflection)</b>	<b>ellipsis (VP)</b>
Baseline	53.0	28.4
concat	<b>76.2</b>	76.6
CADec	72.2	<b>80.0</b>

Caption: Accuracy on ellipsis test set.

For ellipsis, both models improve substantially over the baseline (by 19-51 percentage points), with concat stronger for inflection tasks and CADec stronger for VP ellipsis

# SciGen: Task Definition

	<b>ellipsis (Inflection)</b>	<b>ellipsis (VP)</b>
Baseline	53.0	28.4
concat	<b>76.2</b>	76.6
CADec	72.2	<b>80.0</b>

Caption: Accuracy on ellipsis test set.

For ellipsis, both models improve substantially over the baseline (by 19-51 percentage points), with concat stronger for inflection tasks and CADec stronger for VP ellipsis

# Data Collection

	CoNLL			LEA		
	max	MINA	head	max	MINA	head
CoNLL-2012 test set						
Stanford rule-based	55.60 (8)	57.55 (8)	57.38 (8)	47.31 (8)	49.65 (8)	49.44 (8)
cort	63.03 (7)	64.60 (6)	64.51 (6)	56.10 (6)	58.05 (6)	57.93 (6)
Peng et al.	63.05 (6)	63.50 (7)	63.54 (7)	55.22 (7)	55.76 (7)	55.80 (7)
deep-coref ranking	65.59 (5)	67.29 (5)	67.09 (5)	59.58 (5)	61.70 (5)	61.43 (5)
deep-coref RL	65.81 (4)	67.50 (4)	67.36 (4)	59.76 (4)	61.84 (4)	61.64 (4)
Lee et al. 2017 single	67.23 (3)	68.55 (3)	68.53 (3)	61.24 (3)	62.87 (3)	62.82 (3)
Lee et al. 2017 ensemble	68.87 (2)	70.12 (2)	70.05 (2)	63.19 (2)	64.76 (2)	64.64 (2)
Lee et al. 2018	72.96 (1)	74.26 (1)	75.29 (1)	67.73 (1)	69.32 (1)	70.40 (1)
WikiCoref						
Stanford rule-based	51.78 (4)	53.79 (5)	57.10 (4)	43.28 (5)	45.48 (6)	49.28 (4)
deep-coref ranking	52.90 (3)	55.16 (2)	57.13 (3)	44.40 (3)	46.98 (3)	49.05 (5)
deep-coref RL	50.73 (5)	54.26 (4)	57.16 (2)	41.98 (6)	46.02 (4)	49.29 (3)
Lee et al. 2017 single	50.38 (6)	52.16 (6)	54.02 (6)	43.86 (4)	45.75 (5)	47.69 (6)
Lee et al. 2017 ensemble	53.63 (2)	55.03 (3)	56.80 (5)	47.50 (2)	48.98 (2)	50.87 (2)
Lee et al. 2018	57.89 (1)	59.90 (1)	61.33 (1)	52.42 (1)	54.63 (1)	56.19 (1)

Table 4: Evaluations based on maximum span, MINA, and head spans on the CoNLL-2012 test set and WikiCoref. The ranking of corresponding scores is specified in parentheses. Rankings which are different based on maximum vs. MINA spans are highlighted.

CoNLL-2012 contains the newswire, broadcast news, broadcast conversation, telephone conversation, magazine, weblogs, and Bible genres while the annotated documents in WikiCoref are selected from Wikipedia.

## 6.2 Results

Table 4 shows the maximum vs. minimum span evaluations of several recent coreference resolvers on the CoNLL-2012 test set and the WikiCoref dataset. The examined coreference resolvers are as follows: the Stanford rule-based system (Lee et al., 2013), the coreference resolver of Peng et al. (2015), the ranking model of cort (Matschat and Strube, 2015), the ranking and reinforcement learning models of deep-coref (Clark and Manning, 2016a), the single and ensemble models of Lee et al. (2017), and the current state-of-the-art system by Lee et al. (2018).

We make the following observations based on the results of Table 4:

Using minimum spans in coreference evaluation strongly affects the comparisons in the cross-dataset setting. The results on the WikiCoref dataset show that mention boundary detection errors specifically affect coreference scores in cross-dataset evaluations. The ranking of systems is very different by using maximum vs. min-

imum spans. The reinforcement learning model of deep-coref, i.e., deep-coref-RL, has the most significant difference when it is evaluated based on maximum vs. minimum spans (about 4 points). The ensemble model of deep-coref, on the other hand, has the least difference between maximum and minimum span scores (1.4 points), which indicates it better recognizes maximum span boundaries in out-of-domain data.

Using minimum spans in coreference evaluation reduces the gap between the performance on gold vs. system mentions. It is shown that there is a large gap between the performance of a coreference resolver on gold vs. system mentions, see e.g., Peng et al. (2015). The use of minimum spans in coreference evaluation reduces this gap by about two points. The comparison of the results of different systems on gold and system mentions using both maximum and minimum spans are included in Appendix A.

Evaluation based on minimum spans reduces the differences that are merely due to better maximum boundary detection. The coreference resolver of Peng et al. (2015) has the smallest difference between its maximum and minimum span evaluation scores. This result indicates the superiority of Peng et al. (2015)'s mention

- Annotation by authors
  - Computer Science articles from arXiv.org
- Data cleaning

# Data Collection

	CoNLL				LEA		
	max	MINA	head		max	MINA	head
CoNLL-2012 test set							
Stanford rule-based	55.60 (8)	57.55 (8)	57.38 (8)		47.31 (8)	49.65 (8)	49.44 (8)
cort	63.03 (7)	64.60 (6)	64.51 (6)		56.10 (6)	58.05 (6)	57.93 (6)
Peng et al.	63.05 (6)	63.50 (7)	63.54 (7)		55.22 (7)	55.76 (7)	55.80 (7)
deep-coref ranking	65.59 (5)	67.29 (5)	67.09 (5)		59.58 (5)	61.70 (5)	61.43 (5)
deep-coref RL	65.81 (4)	67.50 (4)	67.36 (4)		59.76 (4)	61.84 (4)	61.64 (4)
Lee et al. 2017 single	67.23 (3)	68.55 (3)	68.53 (3)		61.24 (3)	62.87 (3)	62.82 (3)
Lee et al. 2017 ensemble	68.87 (2)	70.12 (2)	70.05 (2)		63.19 (2)	64.76 (2)	64.64 (2)
Lee et al. 2018	72.96 (1)	74.26 (1)	75.29 (1)		67.73 (1)	69.32 (1)	70.40 (1)
WikiCoref							
Stanford rule-based	51.78 (4)	53.79 (5)	57.10 (4)		43.28 (5)	45.48 (6)	49.28 (4)
deep-coref ranking	52.90 (3)	55.16 (2)	57.13 (3)		44.40 (3)	46.98 (3)	49.05 (5)
deep-coref RL	50.73 (5)	54.26 (4)	57.16 (2)		41.98 (6)	46.02 (4)	49.29 (3)
Lee et al. 2017 single	50.38 (6)	52.16 (6)	54.02 (6)		43.86 (4)	45.75 (5)	47.69 (6)
Lee et al. 2017 ensemble	53.63 (2)	55.03 (3)	56.80 (5)		47.50 (2)	48.98 (2)	50.87 (2)
Lee et al. 2018	57.89 (1)	59.90 (1)	61.33 (1)		52.42 (1)	54.63 (1)	56.19 (1)

Table 4: Evaluations based on maximum span, MINA, and head spans on the CoNLL-2012 test set and WikiCoref. The ranking of corresponding scores is specified in parentheses. Rankings which are different based on maximum vs. MINA spans are highlighted.

CoNLL-2012 contains the newswire, broadcast news, broadcast conversation, telephone conversation, magazine, weblogs, and Bible genres while the annotated documents in WikiCoref are selected from Wikipedia.

## 6.2 Results

Table 4 shows the maximum vs. minimum span evaluations of several recent coreference resolvers on the CoNLL-2012 test set and the WikiCoref dataset. The examined coreference resolvers are as follows: the Stanford rule-based system (Lee et al., 2013), the coreference resolver of Peng et al. (2015), the ranking model of cort (Martschat and Strube, 2015), the ranking and reinforcement learning models of deep-coref (Clark and Manning, 2016a), the single and ensemble models of Lee et al. (2017), and the current state-of-the-art system by Lee et al. (2018).

We make the following observations based on the results of Table 4:

Using minimum spans in coreference evaluation strongly affects the comparisons in the cross-dataset setting. The results on the WikiCoref dataset show that mention boundary detection errors specifically affect coreference scores in cross-dataset evaluations. The ranking of systems is very different by using maximum vs. min-

imum spans. The reinforcement learning model of deep-coref, i.e., deep-coref-RL, has the most significant difference when it is evaluated based on maximum vs. minimum spans (about 4 points). The ensemble model of deep-coref, on the other hand, has the least difference between maximum and minimum span scores (1.4 points), which indicates it better recognizes maximum span boundaries in out-of-domain data.

Using minimum spans in coreference evaluation reduces the gap between the performance on gold vs. system mentions. It is shown that there is a large gap between the performance of a coreference resolver on gold vs. system mentions, see e.g., Peng et al. (2015). The use of minimum spans in coreference evaluation reduces this gap by about two points. The comparison of the results of different systems on gold and system mentions using both maximum and minimum spans are included in Appendix A.

Evaluation based on minimum spans reduces the differences that are merely due to better maximum boundary detection. The coreference resolver of Peng et al. (2015) has the smallest difference between its maximum and minimum span evaluation scores. This result indicates the superiority of Peng et al. (2015)'s mention



High quality



Does not scale to large training data sizes



Using LaTeX sources to automatically extract table-description pairs

# Data Collection

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

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



“**xx**table**xx**anchor**S3T2** Table 2: Impact of using coverage for improving generalization across ...”

“Table **xx**ref**S3T2** shows the performance for both systems for in-domain ...”



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

# Data Collection

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

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

“~~xx~~table~~xx~~anchor~~S3T2~~ Table 2: Impact of using coverage for improving generalization across ...”



+ Rule-based post-pruning

“Table ~~xx~~ref~~S3T2~~ shows the performance for both systems for in-domain ...”

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

# SciGen

Dataset	Pairs	Cell	Num.	Text	Vocab	Domain	Annotation	Reasoning
WikiBIO	400K	17	3	97	400K	Biography	Automated	No
Rotowire	11K	649	429	337	11.3K	Basketball	Automated	Few
ToTTo	136K	3	1	17	136K	Open (Wikipedia)	Human	Few
LogicNLG	37K	91	35	14	122K	Open (Wikipedia)	Human/Automated	Yes
SciGen (few-shot)	1.3K	54	35	116	11K	Scientific	Expert	Yes
SciGen (medium)	18K	51	34	124	54K	Scientific	Expert/Automated	Yes
SciGen (Large)	53K	55	38	133	127K	Scientific	Expert/Automated	Yes

# Experiments

Baselines

BART-large, T5-large

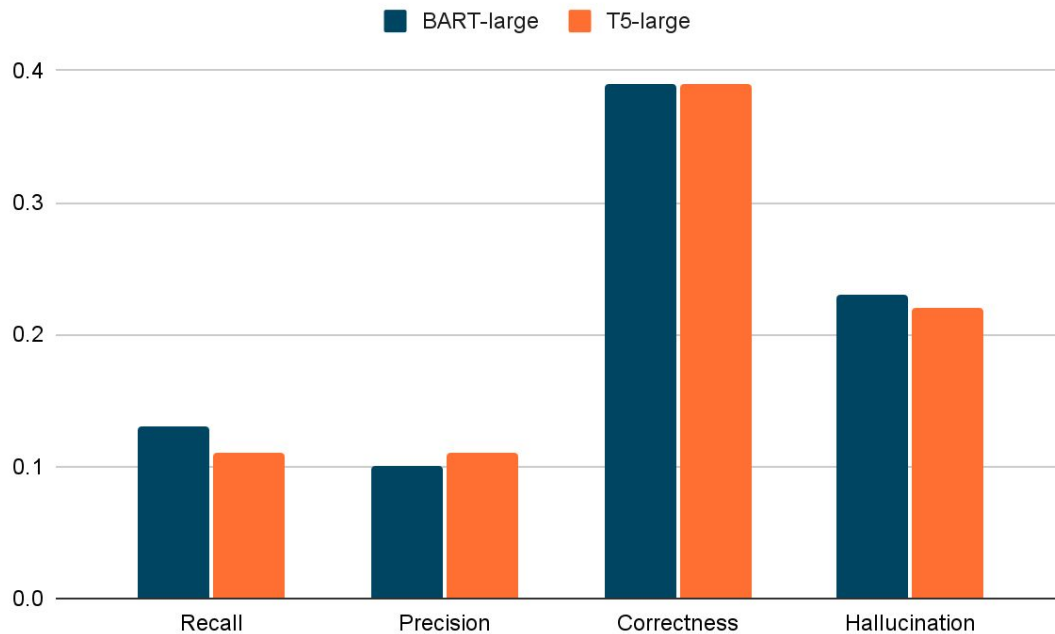
# Experiments

- Automatic Evaluation
  - BLEU, METEOR, BertScore, MoverScore, BLEURT
- Human Evaluation
  - Recall, Precision, Correctness, Hallucination

## Results: Automatic Metrics

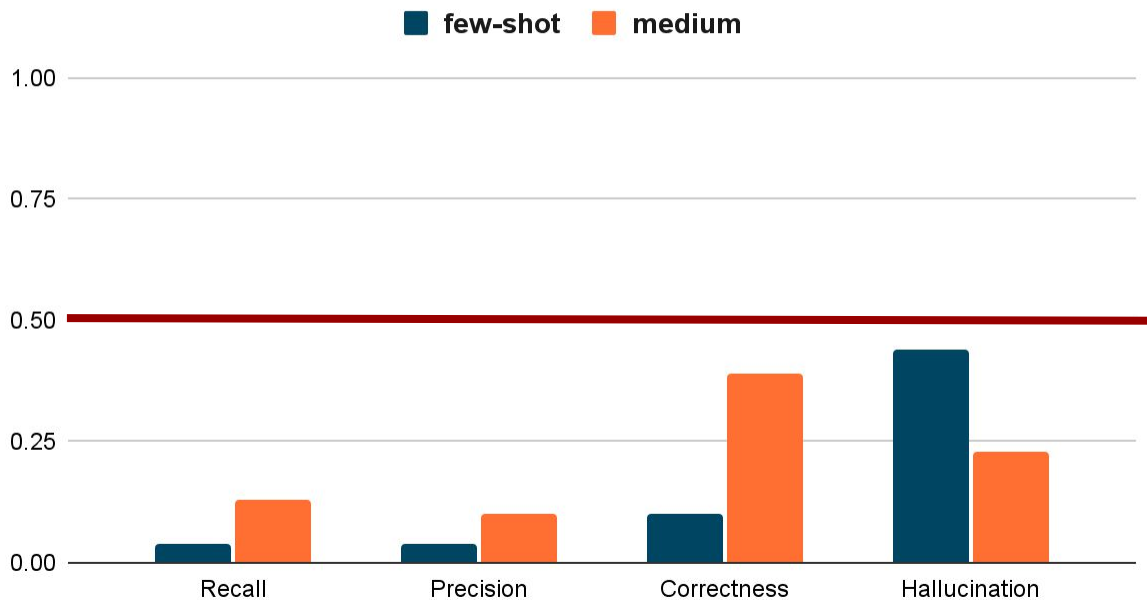
Setting	Model	BLEU	METEOR	MoverS	BertS	BLEURT
Test(C&L)						
Zero	BART	1.11	0.11	-0.02	0.74	-1.23
	T5-base	0.69	0.04	-0.05	0.76	-1.31
	T5-large	1.16	0.06	-0.06	0.76	-1.28
Few	BART	4.73	0.22	<b>0.14</b>	<b>0.84</b>	<b>-0.66</b>
	T5-base	2.59	0.13	0.02	0.79	-1.09
	T5-large	3.16	0.16	0.06	0.81	-0.95
Medium	BART	<b>5.30</b>	<b>0.23</b>	0.13	<b>0.84</b>	-0.72
	T5-base	3.32	0.15	0.05	0.82	-0.89
	T5-large	3.65	0.17	0.10	0.83	-0.77
Large	BART	5.04	0.22	<b>0.14</b>	<b>0.84</b>	-0.71
	T5-base	3.38	0.15	0.06	0.82	-0.85
	T5-large	3.84	0.18	0.10	0.83	-0.79

# Results: Human Evaluation



# Results: Human Evaluation

BART-large



# Results

	ellipsis (infl.)	ellipsis (VP)
baseline	53.0	28.4
concat	<b>76.2</b>	76.6
s-hier-to-2.tied	66.4	65.6
CADec	72.2	<b>80.0</b>

Table 8: Accuracy on ellipsis test set.

Few	table 8 shows the performance of our model on the ellipsis test set. <b>our concat model outperforms all the other models that do not use concat embeddings.</b>
Med.	the results show that both concat and cadec significantly outperform the baseline and s-hier-to-2.tied on the ellipsis test set. the concat model is better than the baseline on both inflected and verb-inflated forms, <b>while</b> cadec is better on the vp version. on the inflated version, cadec outperforms the baseline by a large margin and <b>is within 3 points of the best model.</b> the model with the best accuracy on the verb portion of the test set (80.0%) <b>is very close to the best performance on the noun portion (75.6%). the results indicate that cadec has better generalizability than the other models.</b>
Large	the concatenation model concatenates the output of the concatenated encoder and decoder, and <b>the s-hier-to-2.tied model performs slightly better than concat on both ellipsis (infl.) and vice versa.</b> cadec outperforms the other two models by a large margin, achieving 80.0% accuracy on the vp ellipsoids, which is the new state of the art.

# Results

Generated descriptions

Fluent

Look like valid descriptions

**But** contain factually incorrect or irrelevant facts wrt table contents

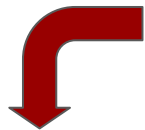
# Challenges

Generated descriptions

Fluent

Look like valid descriptions

**But** contain factually incorrect or irrelevant facts wrt table contents



Requires better evaluation  
metrics



Requires generation models  
with better reasoning skills

# Questions?

- SciGen: a new dataset to enable end-to-end arithmetic reasoning in text generation
- Challenges
  - Evaluation metrics
  - Reasoning-aware models

<https://github.com/UKPLab/SciGen>



Improving end-to-end arithmetic reasoning

# Arithmetic-Based Pretraining Improving Numeracy of Pretrained Language Models



Dominic Petrak



Nafise Sadat Moosavi



Iryna Gurevych

# Numerical Reasoning

- Specialized architectures
- Pretraining from scratch

## Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension

Daniel Andor, Luheng He, Kenton Lee, Emily Pitler

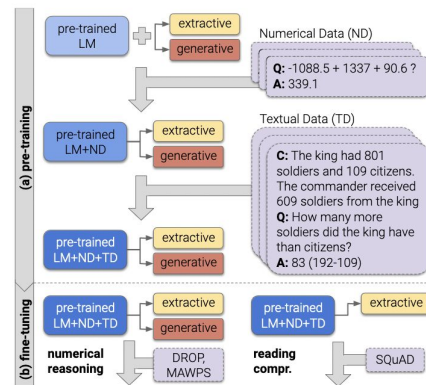


Figure 1: An overview of our approach for injecting numerical skills into a pre-trained LM. (a) We add two pre-training steps over large amounts of synthetic numerical data (ND) and textual data (TD); (b) we further fine-tune the model over either numerical reasoning datasets (DROP, MAWPS) or reading comprehension datasets (SQUAD).

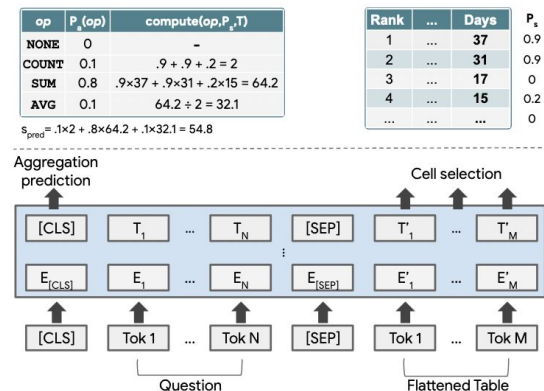


Figure 1: TAPAS model (bottom) with example model outputs for the question: "Total number of days for the top two". Cell prediction (top right) is given for the selected column's table cells in bold (zero for others) along with aggregation prediction (top left).

# Our Approach

- Improved number representation
- Specialized extended pretraining step

# Number Representation

- Commonly used tokenizations are based on the frequency of patterns
  - Byte Pair Encoding (Sennrich et al., 2016) or WordPiece (Wu et al., 2016)
  - 0.72 and 0.73
    - $[0, ., 72]$  and  $[0, ., 7, 3]$
- This is not suitable for numbers!

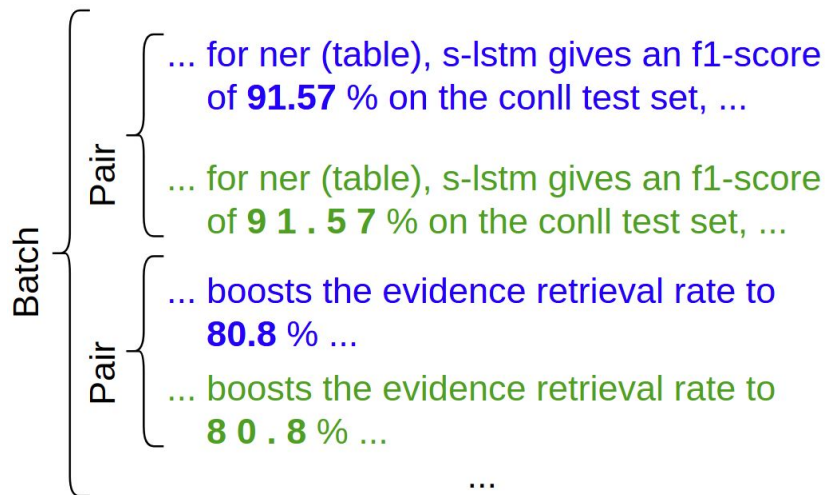
# Number Representation

- ✓ Making the semantic representation of numbers independent of the underlying tokenization
  - Using different tokenization algorithms
    - Byte-pair encoding
    - Character-level embeddings
  - Using contrastive learning
    - Learning a similar representation for different tokenizations of the same number

# Number Representation

## ✓ Making the semantic representation of numbers independent of the underlying tokenization

- Using different tokenization algorithms
- Using contrastive learning



■ Positive Sample   ■ Anchor

# Arithmetic Reasoning

- ✓ An extended pretraining step focusing on arithmetic reasoning
  - Masked word prediction pretraining does not target arithmetic reasoning

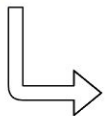
# Arithmetic Reasoning

- ✓ An extended pretraining step focusing on arithmetic reasoning
  - Masked word prediction pretraining does not target 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>

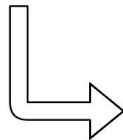


With less than **2** minutes to go, how many points ahead was Dallas?

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 <b>Untitled (1981)</b> , a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was <b>sold by Robert Lehrman for \$16.3 million, well above its \$12 million high estimate.</b>	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000
Comparison (18.2%)	In <b>1517, the seventeen-year-old King sailed to Castile.</b> There, his Flemish court .... <b>In May 1518, Charles traveled to Barcelona in Aragon.</b>	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, <b>Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack</b> 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 <b>2 March 1992.</b> The JNA formed a battlegroup to counterattack the <b>next day.</b>	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 <b>Matt Prater nailing a 43-yard field goal</b> , yet Carolina answered as kicker <b>John Kasay ties the game with a 39-yard field goal.</b> ... Carolina closed out the half with <b>Kasay nailing a 44-yard field goal.</b> ... In the fourth quarter, Carolina sealed the win with <b>Kasay's 42-yard field goal.</b>	Which kicker kicked the most field goals?	John Kasay
Coreference Resolution (3.7%)	<b>James Douglas</b> was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before <b>1543 he married Elizabeth</b> , daughter of James Douglas, 3rd Earl of Morton. <b>In 1553 James Douglas succeeded to the title and estates of his father-in-law.</b>	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 <b>60,000 adherents</b> , the subsequent establishment of the Bulgarian Exarchate <b>reduced their number by some 75%.</b>	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000

# Evaluation

- Tasks
  - Inference-On-Tables (InfoTabs)

Dressage	
<b>Highest governing body</b>	International Federation for Equestrian Sports (FEI)
<i>Characteristics</i>	
<b>Contact</b>	No
<b>Team members</b>	Individual and team at international levels
<b>Mixed gender</b>	Yes
<b>Equipment</b>	Horse, horse tack
<b>Venue</b>	Arena, indoor or outdoor
<i>Presence</i>	
<b>Country or region</b>	Worldwide
<b>Olympic</b>	1912
<b>Paralympic</b>	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	<b>73.84</b>	<b>65.38</b>	<b>78.69</b>	<b>54.55</b>
ESIM (ELMO)	80.04	68.70	60.21	51.37
+ coverage	<b>80.38</b>	<b>70.05</b>	<b>67.47</b>	<b>52.65</b>

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.

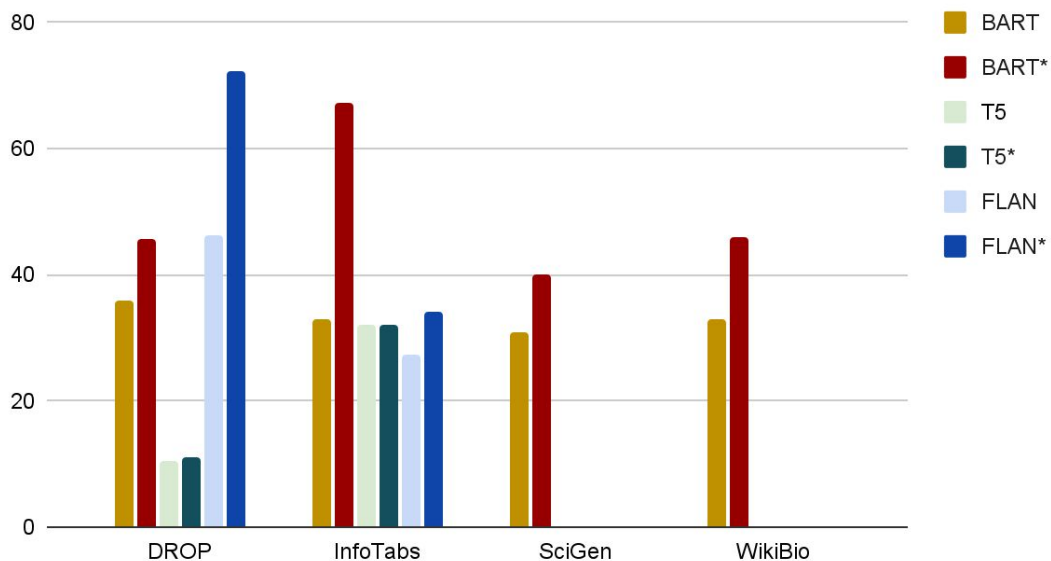
The results show that coverage information considerably improves the generalization of both examined models across various NLI datasets. The resulting cross-dataset improvements on the SNLI and Glockner datasets are larger than those on the SICK dataset. The reason is that the dataset creation process 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)

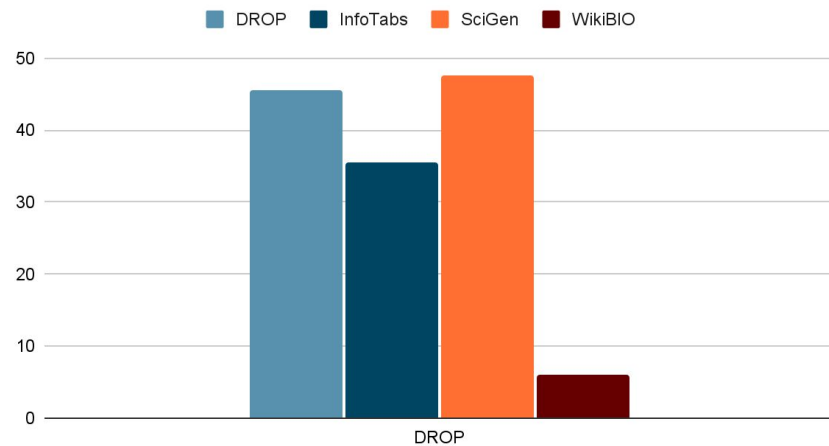
# Results

Downstream performance

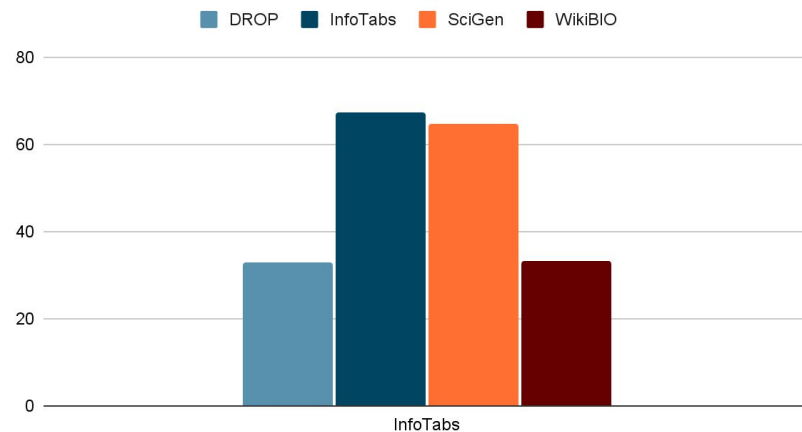


# Out-of-domain Pretraining

EM



EM



# Questions?

- Promising results in various downstream tasks
  - Using an extended pretraining step
  - No need to an architecture change
  - No scaling
  - No additional annotated data



Better evaluation

# FERMAT: An Alternative to Accuracy for Numerical Reasoning



Jasivan Sivakumar

# Problem

Measuring the performance using a single score

- What are the shortcomings and strengths?
- Where to go from here?

# Problem

Measuring the performance using a single score

- What are the shortcomings and strengths?
- Where to go from here?

## FERMAT

Flexible Evaluation set for  
Representing Multi-views  
of Arithmetic Types



Evaluates models on:

- Number Understanding
- Mathematical Operations
- Training Dependency

# Number Understanding

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?

# 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

# Training Dependencies

- **Exact:** all the numbers and operations are seen during finetuning
  - A Euro is 5 yens. How much is 25 Euros?
  - Each apple costs 5 cents. How much do 25 apples cost?
- **All Numbers:** all the numbers are seen
- **Number & Operation:** at least one number and operation
- **One Number**

## Zero-shot Evaluation

Models (size)		Number Understanding																		
		Alternate Representations												Range of numbers						
		Same numbers					Same digits					Grouping		Integers				Decimals		
		Original	Fixed 1dp	Fixed 2dp	Worded	Commutted	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

# Zero-shot Evaluation

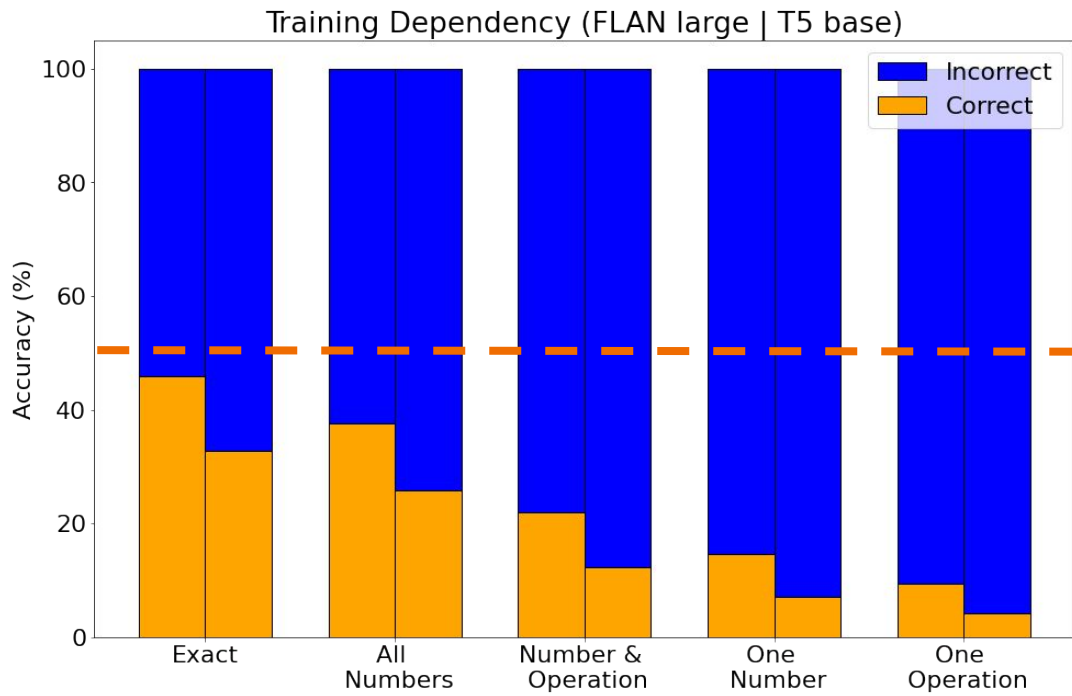
Models (size)		Number Understanding																		
		Alternate Representations												Range of numbers						
		Same numbers					Same digits					Grouping		Integers				Decimals		
		Original	Fixed 1dp	Fixed 2dp	Worded	Commutd	Original 1dp	Original 2dp	Original 1dp no 0	Original 2dp no 0	Original 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

# Finetuning

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

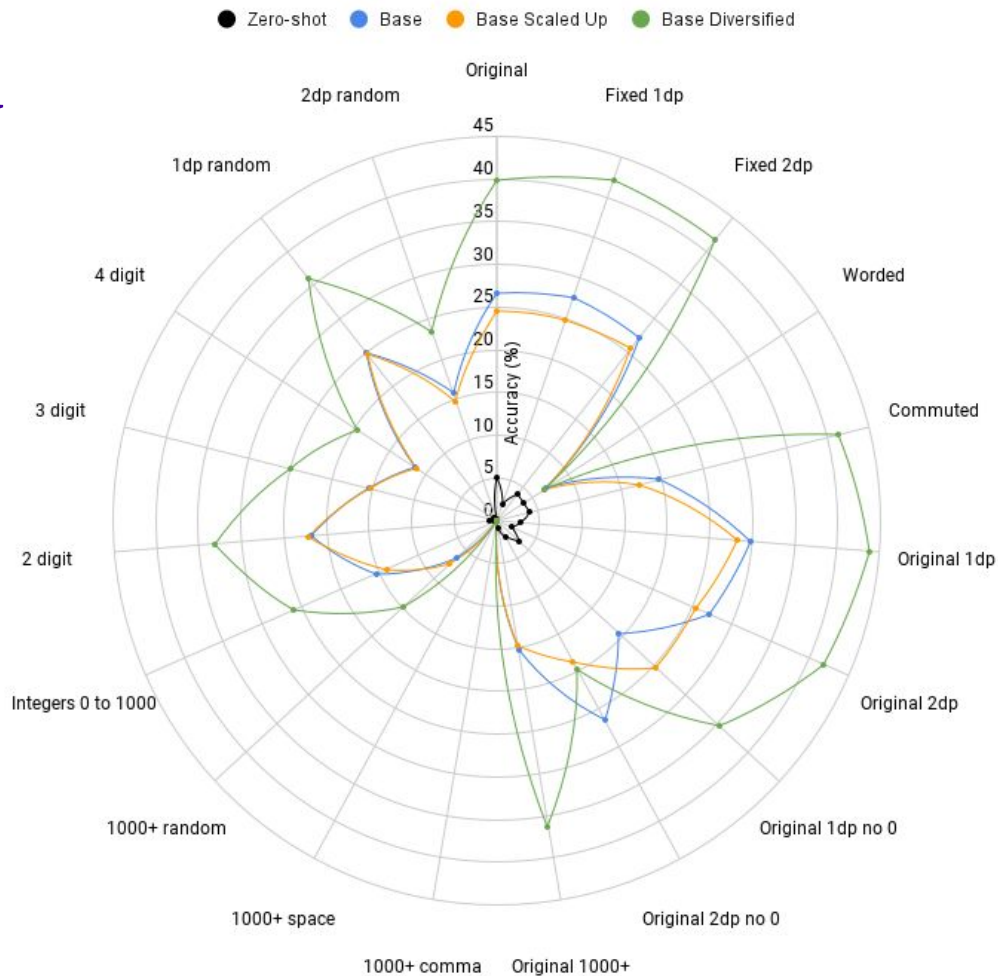
200K examples from 100 templates written by math teachers

# Training Dependency



# Impact of training data

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



# Conclusions

- Enable learning & evaluation
  - Creating datasets for end-to-end reasoning
  - Designing proper evaluation metrics
- Improving end-to-end arithmetic reasoning
  - Better number understanding
  - Specialized (extended) pretraining objectives
  - Language diversity

# Questions?

