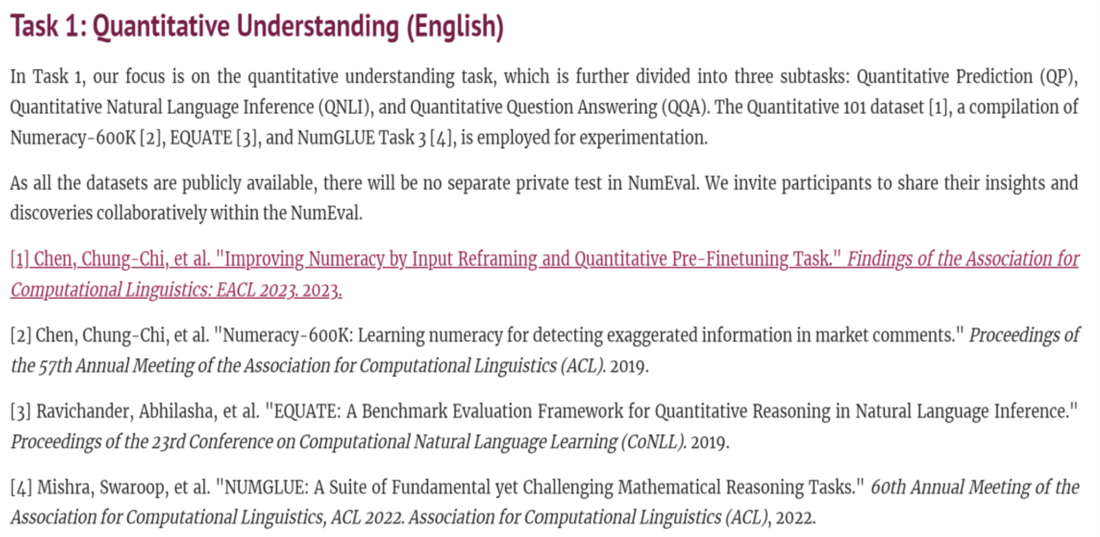
NumEval Past Papers , Findings and Potential Approaches

Presentation One (Seon)

Our Strategy

1



Task 1: Quantitative Understanding (English)

In Task 1, our focus is on the quantitative understanding task, which is further divided into three subtasks: Quantitative Prediction (QP), Quantitative Natural Language Inference (QNLI), and Quantitative Question Answering (QQA). The Quantitative 101 dataset [1], a compilation of Numeracy-600K [2], EQUATE [3], and NumGLUE Task 3 [4], is employed for experimentation.

As all the datasets are publicly available, there will be no separate private test in NumEval. We invite participants to share their insights and discoveries collaboratively within the NumEval.

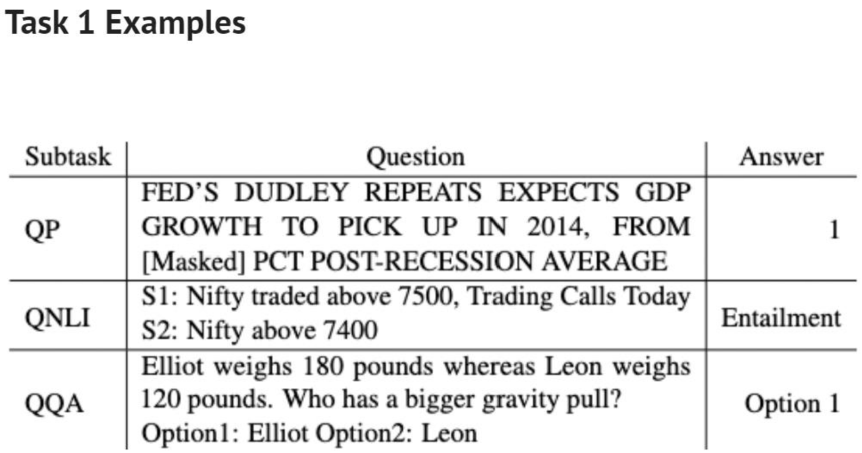
[1] Chen, Chung-Chi, et al. "Improving Numeracy by Input Reframing and Quantitative Pre-Finetuning Task." *Findings of the Association for Computational Linguistics: EACL 2023*. 2023.

[2] Chen, Chung-Chi, et al. "Numeracy-600K: Learning numeracy for detecting exaggerated information in market comments." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

[3] Ravichander, Abhilasha, et al. "EQUATE: A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference." *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 2019.

[4] Mishra, Swaroop, et al. "NUMGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks." *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022. Association for Computational Linguistics (ACL)*, 2022.

2



Task 1 Examples

Subtask	Question	Answer
QP	FED'S DUDLEY REPEATS EXPECTS GDP GROWTH TO PICK UP IN 2014, FROM [Masked] PCT POST-RECESSION AVERAGE	1
QNLI	S1: Nifty traded above 7500, Trading Calls Today S2: Nifty above 7400	Entailment
QQA	Elliot weighs 180 pounds whereas Leon weighs 120 pounds. Who has a bigger gravity pull? Option1: Elliot Option2: Leon	Option 1

1. We Started reading and researching old papers and findings and come up with news ideas and inspire from them

2. From what we learned from these papers , figure out what we can do to further improve Quantitative understanding task! How will we approach our task at hand! What can we do different than 2023 Numeval paper!

Biggest Takeaway from 2023 findings paper

1

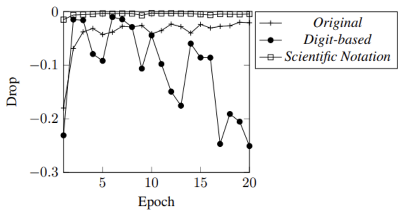


Figure 1: BERT's innumeracy phenomenon. (Performance Drop between CND-T1 and CND-T2.)

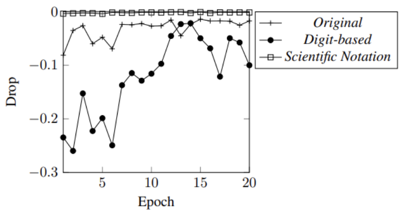


Figure 3: LinkBERT's innumeracy phenomenon. (Performance Drop between CND-T1 and CND-T2.)

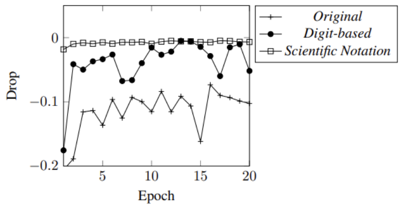


Figure 2: RoBERTa's innumeracy phenomenon. (Performance Drop between CND-T1 and CND-T2.)

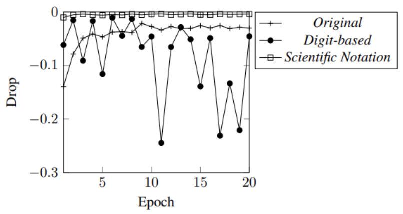
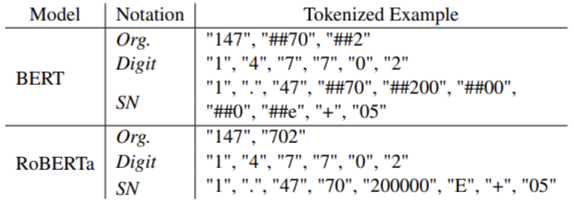


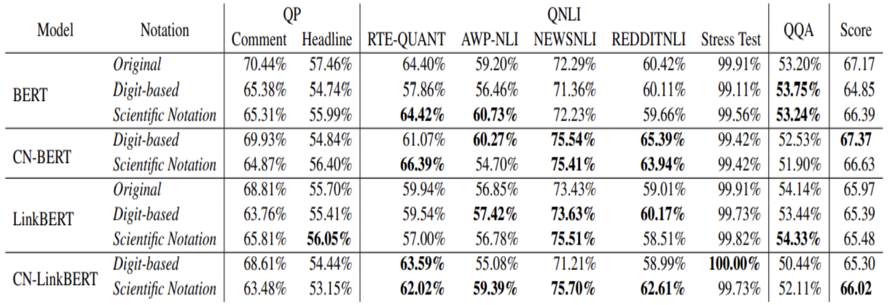
Figure 4: FinBERT's innumeracy phenomenon. (Performance Drop between CND-T1 and CND-T2.)

2



Model	Notation	Tokenized Example
BERT	*Org.*	"147", "##70", "##2"
	Digit	"1", "4", "7", "7", "0", "2"
	SN	"1", ".", "47", "##70", "##200", "##00", "##0", "##e", "+", "05"
RoBERTa	*Org.*	"147", "702"
	Digit	"1", "4", "7", "7", "0", "2"
	SN	"1", ".", "47", "70", "200000", "E", "+", "05"

Table 1: Tokenized example. Org. and SN denote original and scientific notation, respectively.



Model	Notation	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
BERT	*Original*	70.44%	57.46%	64.40%	59.20%	72.29%	60.42%	99.91%	53.20%	67.17
	Digit-based	65.38%	54.74%	57.86%	56.46%	71.36%	60.11%	99.11%	**53.75%**	64.85
	Scientific Notation	65.31%	55.99%	**64.42%**	**60.73%**	72.23%	59.66%	99.56%	**53.24%**	66.39
CN-BERT	*Digit-based*	69.93%	54.84%	61.07%	**60.27%**	**75.54%**	**65.39%**	99.42%	52.53%	**67.37**
	Scientific Notation	64.87%	56.40%	**66.39%**	54.70%	**75.41%**	**63.94%**	99.42%	51.90%	66.63
LinkBERT	*Original*	68.81%	55.70%	59.94%	56.85%	73.43%	59.01%	99.91%	54.14%	65.97
	Digit-based	63.76%	55.41%	59.54%	**57.42%**	**73.63%**	**60.17%**	99.73%	53.44%	65.39
	Scientific Notation	65.81%	**56.05%**	57.00%	56.78%	**75.51%**	58.51%	99.82%	**54.33%**	65.48
CN-LinkBERT	*Digit-based*	68.61%	54.44%	**63.59%**	55.08%	71.21%	58.99%	**100.00%**	50.44%	65.30
	Scientific Notation	63.48%	53.15%	**62.02%**	**59.39%**	**75.70%**	**62.61%**	99.73%	52.11%	**66.02**

Table 4: Experimental results of the BERT-based models. The results in bold are the ones that are better than the *Original*. The score indicates Quantitative-101 Score.

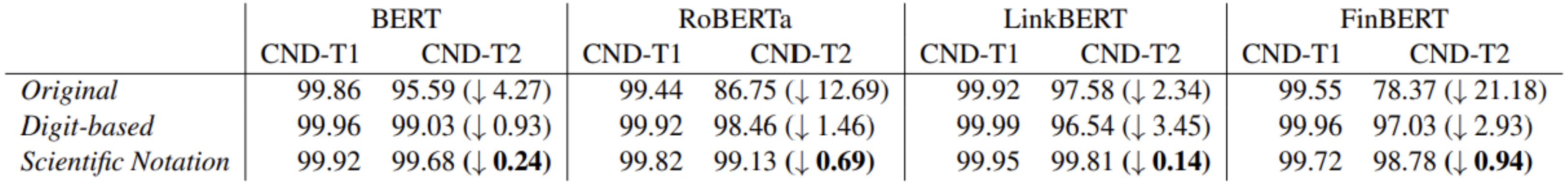
1. **Pretrained LMs exhibit some degree of "innumeracy" - difficulty in understanding numbers outside the range seen during pretraining. This phenomenon needs more investigation across different models.**

2. **Representing numbers in digit or scientific notation can help improve numeracy compared to relying on original text tokens. This indicates that better numeric tokenization and input formatting is important.**

2023 Cont’d

3. **Pre-finetuning on simple numeric reasoning tasks can enhance a model's basic skills in comprehending and comparing numbers. Designing such pretraining objectives is a promising direction.**

4. **There is room for improvement across tasks that require understanding semantics along with numerical concepts, like quantitative NLI and QA. Models still struggle with reasoning holistically over text and numbers.**



	BERT		RoBERTa		LinkBERT		FinBERT	
	CND-T1	CND-T2	CND-T1	CND-T2	CND-T1	CND-T2	CND-T1	CND-T2
Original	99.86	95.59 (↓ 4.27)	99.44	86.75 (↓ 12.69)	99.92	97.58 (↓ 2.34)	99.55	78.37 (↓ 21.18)
Digit-based	99.96	99.03 (↓ 0.93)	99.92	98.46 (↓ 1.46)	99.99	96.54 (↓ 3.45)	99.96	97.03 (↓ 2.93)
Scientific Notation	99.92	99.68 (↓ **0.24**)	99.82	99.13 (↓ **0.69**)	99.95	99.81 (↓ **0.14**)	99.72	98.78 (↓ **0.94**)

Table 3: Experimental results of ComNum task. The evaluation metric is Micro-average of F1 score (%).

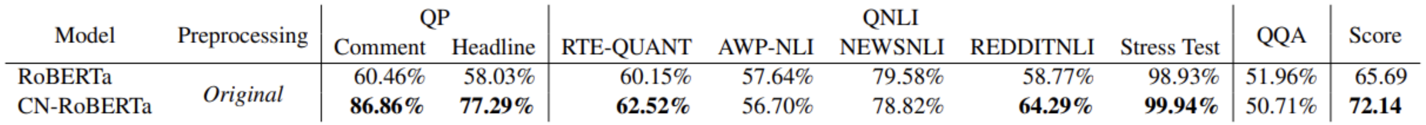
What’sGoodforyou&NotGoodforMes!

3. **There is room for improvement across tasks that require understanding semantics along with numerical concepts, like quantitative NLI and QA. Models still struggle with reasoning holistically over text and numbers.**

4. **Techniques effective for one model architecture may not directly transfer to others. More work is needed to develop methods that robustly improve numeracy across model types.**

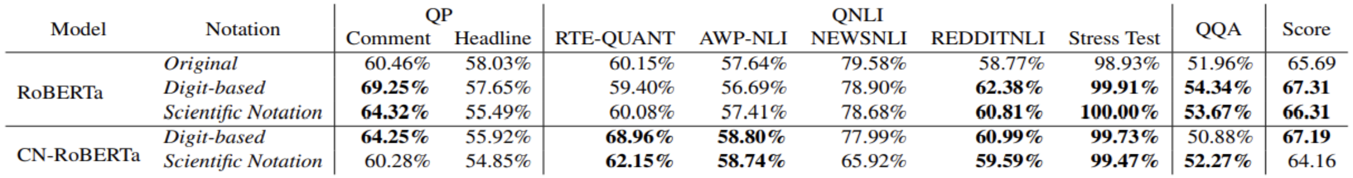
for Roberta
↑ Big Gains

for Bert
↓ less helpful



Model	Preprocessing	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
RoBERTa	*Original*	60.46%	58.03%	60.15%	57.64%	79.58%	58.77%	98.93%	51.96%	65.69
CN-RoBERTa		**86.86%**	**77.29%**	**62.52%**	56.70%	78.82%	**64.29%**	**99.94%**	50.71%	**72.14**

Table 7: Results of CN-RoBERTa without input reframing.



Model	Notation	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
RoBERTa	*Original*	60.46%	58.03%	60.15%	57.64%	79.58%	58.77%	98.93%	51.96%	65.69
	Digit-based	**69.25%**	57.65%	59.40%	56.69%	78.90%	**62.38%**	**99.91%**	**54.34%**	**67.31**
	Scientific Notation	**64.32%**	55.49%	60.08%	57.41%	78.68%	**60.81%**	**100.00%**	**53.67%**	**66.31**
CN-RoBERTa	*Digit-based*	**64.25%**	55.92%	**68.96%**	**58.80%**	77.99%	**60.99%**	**99.73%**	50.88%	**67.19**
	Scientific Notation	60.28%	54.85%	**62.15%**	**58.74%**	65.92%	**59.59%**	**99.47%**	**52.27%**	64.16

Table 5: Experimental results of the RoBERTa-based models.

Why so?

- Maybe because RoBERTa takes the original BERT model and enhances the pretraining process with more data, longer training, and an improved masking technique to achieve state-of-the-art results on many downstream NLP tasks. It demonstrates how transforming BERT's pretraining can yield better language representations.

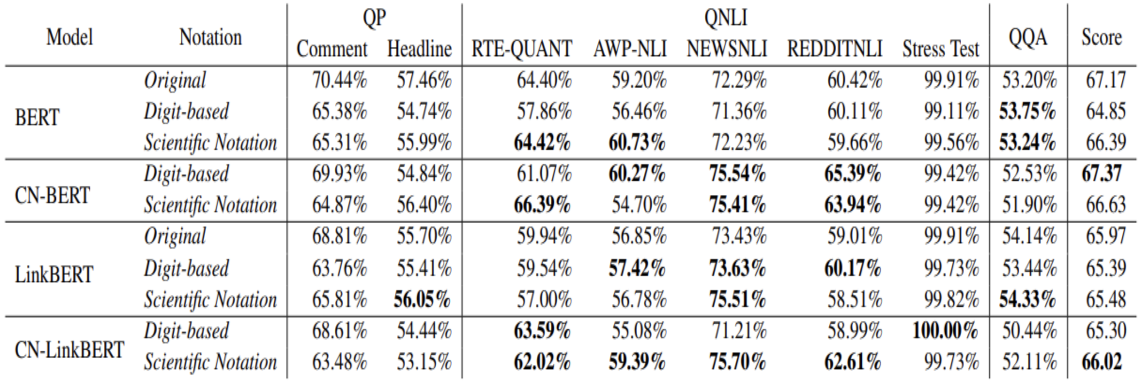
“ RoBERTa = Robustly Optimized BERT Pretraining Approach”

Fun Fact :

- Roberta uses dynamic masking, where the masked positions and values randomly change each time an example is fed to the model during pretraining.
- Dynamic masking exposes the model to more input variability and helps prevent overfitting to spurious patterns during pretraining.
- RoBERTa, which improves on BERT's pretraining, uses dynamic masking instead of BERT's static masking, in which the tokens masked out are fixed when creating each pretraining example.
- So if 15% tokens are masked, the same 15% tokens will be always be masked for that example during all epochs of pretraining.
- The mask token positions and masked words do not change across training iterations.

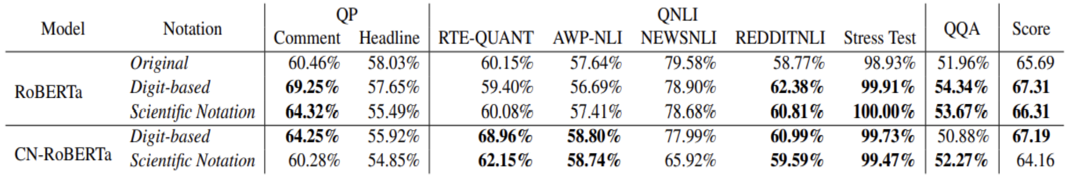
2023 Cont’d

7. **The proposed Quantitative 101 benchmark provides a diverse evaluation suite for numeric reasoning skills. But more complex, real-world tasks need to be included as models improve.**



Model	Notation	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
BERT	*Original*	70.44%	57.46%	64.40%	59.20%	72.29%	60.42%	99.91%	53.20%	67.17
	Digit-based	65.38%	54.74%	57.86%	56.46%	71.36%	60.11%	99.11%	**53.75%**	64.85
	Scientific Notation	65.31%	55.99%	**64.42%**	**60.73%**	72.23%	59.66%	99.56%	**53.24%**	66.39
CN-BERT	*Digit-based*	69.93%	54.84%	61.07%	**60.27%**	**75.54%**	**65.39%**	99.42%	52.53%	**67.37**
	Scientific Notation	64.87%	56.40%	**66.39%**	54.70%	**75.41%**	**63.94%**	99.42%	51.90%	66.63
LinkBERT	*Original*	68.81%	55.70%	59.94%	56.85%	73.43%	59.01%	99.91%	54.14%	65.97
	Digit-based	63.76%	55.41%	59.54%	**57.42%**	**73.63%**	**60.17%**	99.73%	53.44%	65.39
	Scientific Notation	65.81%	**56.05%**	57.00%	56.78%	**75.51%**	58.51%	99.82%	**54.33%**	65.48
CN-LinkBERT	*Digit-based*	68.61%	54.44%	**63.59%**	55.08%	71.21%	58.99%	**100.00%**	50.44%	65.30
	Scientific Notation	63.48%	53.15%	**62.02%**	**59.39%**	**75.70%**	**62.61%**	99.73%	52.11%	**66.02**

Table 4: Experimental results of the BERT-based models. The results in bold are the ones that are better than the *Original*. The score indicates Quantitative-101 Score.



Model	Notation	QP		QNLI					QQA	Score
		Comment	Headline	RTE-QUANT	AWP-NLI	NEWSNLI	REDDITNLI	Stress Test		
RoBERTa	*Original*	60.46%	58.03%	60.15%	57.64%	79.58%	58.77%	98.93%	51.96%	65.69
	Digit-based	**69.25%**	57.65%	59.40%	56.69%	78.90%	**62.38%**	**99.91%**	**54.34%**	**67.31**
	Scientific Notation	**64.32%**	55.49%	60.08%	57.41%	78.68%	**60.81%**	**100.00%**	**53.67%**	**66.31**
CN-RoBERTa	*Digit-based*	**64.25%**	55.92%	**68.96%**	**58.80%**	77.99%	**60.99%**	**99.73%**	50.88%	**67.19**
	Scientific Notation	60.28%	54.85%	**62.15%**	**58.74%**	65.92%	**59.59%**	**99.47%**	**52.27%**	64.16

Table 5: Experimental results of the RoBERTa-based models.

2023 Cont’d

7. **While numeracy has seen increased focus, it is still an open challenge. There are opportunities to draw ideas from numeric representations studied in other domains like knowledge bases.**

Overall, the paper provides a foundation, but substantial innovation is still needed in architectures, pretraining objectives, understanding numbers in context, and evaluation to reach human-like numeracy.

↑ Our goal ☺

mission impossible

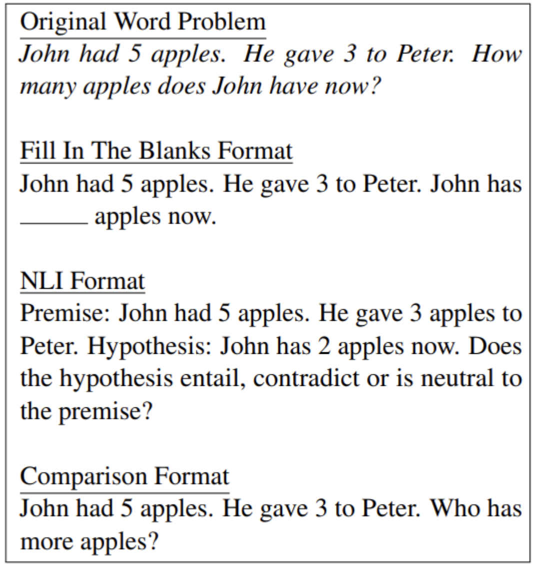
Biggest Takeaway from 2022 NumGlue paper

1. **The need for more robust arithmetic reasoning abilities in AI systems beyond narrow datasets.**
2. **The paper shows that current state-of-the-art models still struggle with basic arithmetic reasoning when tested in a general way across formats and contexts. They fail to robustly apply the skills and latch onto dataset biases instead.**

Simply put , models and systems fails to solve generic problems , failing to go beyond

Their biases of the narrow datasets they were trained on! (they can't help it, huh?)

Sad



Original Word Problem

John had 5 apples. He gave 3 to Peter. How many apples does John have now?

Fill In The Blanks Format

John had 5 apples. He gave 3 to Peter. John has ______ apples now.

NLI Format

Premise: John had 5 apples. He gave 3 apples to Peter. Hypothesis: John has 2 apples now. Does the hypothesis entail, contradict or is neutral to the premise?

Comparison Format

John had 5 apples. He gave 3 to Peter. Who has more apples?

Figure 1: A system that can robustly perform numeric reasoning over language should be able to solve problems such as the above, regardless of how the problem is posed. However, we observe existing systems are brittle; producing inconsistent solutions to such minor stylistic variations.

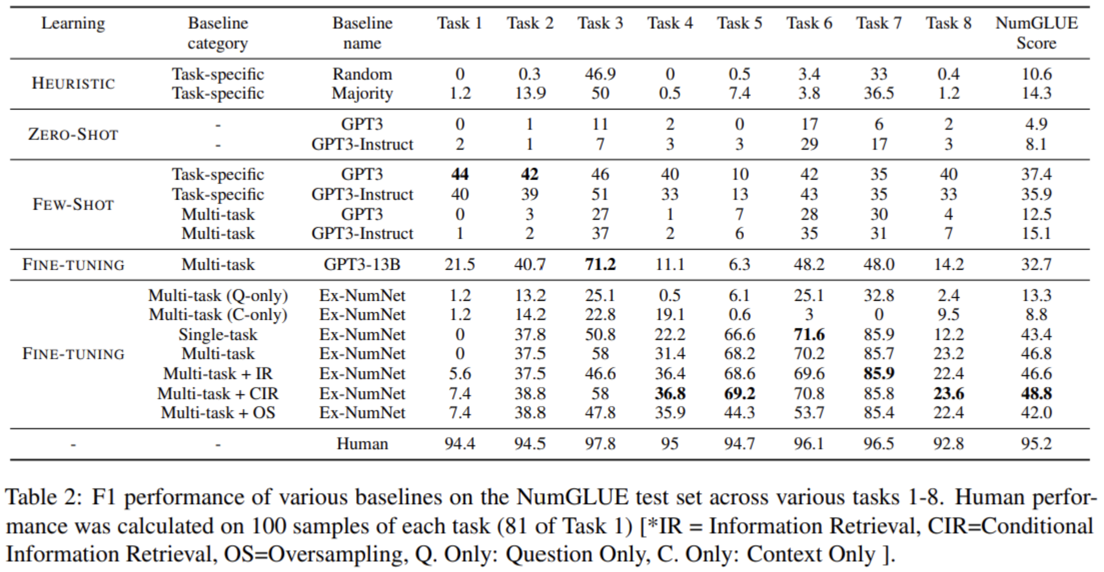
2022 NumGlue Cont’d

3. **This paper however suggests a fundamental limitation in existing models' mathematical understanding. Researchers need to focus on developing systems that can truly perform the underlying arithmetic calculations and reasoning in a generalizable way, not just optimize performance on specific datasets. (exactly my point: a model able to solve whatever problems comes in its way: beyond what it’s seen in training)**
4. **The analysis of error types points towards some areas for improvement, like better numerical parsing and reasoning skills. But more fundamentally, researchers may need to rethink model architectures and training approaches to build arithmetic reasoning abilities rather than relying on large datasets and benchmarks.**

“ It’s true !!! Usually, what we do is we rely too much on large datasets for training purposes and benchmarks for evaluating our systems. We have to rethink our model architectures and training approaches ”

Ohh you want the specifics!

- NUMGLUE is meant to test if systems can robustly perform arithmetic reasoning across different tasks and question formats. This tests if they truly understand the skills rather than just solving specific datasets.
- Experiments show NUMGLUE is challenging even for large state-of-the-art models like GPT-3, which perform much worse than humans. This indicates current AI lacks fundamental arithmetic reasoning skills.

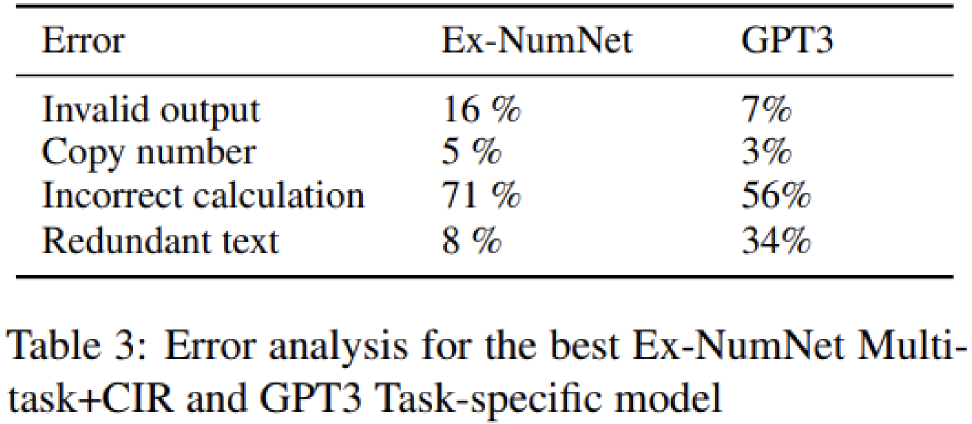


Learning	Baseline category	Baseline name	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	NumGLUE Score
HEURISTIC	Task-specific	Random	0	0.3	46.9	0	0.5	3.4	33	0.4	10.6
	Task-specific	Majority	1.2	13.9	50	0.5	7.4	3.8	36.5	1.2	14.3
ZERO-SHOT	-	GPT3	0	1	11	2	0	17	6	2	4.9
	-	GPT3-Instruct	2	1	7	3	3	29	17	3	8.1
FEW-SHOT	Task-specific	GPT3	**44**	**42**	46	40	10	42	35	40	37.4
	Task-specific	GPT3-Instruct	40	39	51	33	13	43	35	33	35.9
	Multi-task	GPT3	0	3	27	1	7	28	30	4	12.5
	Multi-task	GPT3-Instruct	1	2	37	2	6	35	31	7	15.1
FINE-TUNING	Multi-task	GPT3-13B	21.5	40.7	**71.2**	11.1	6.3	48.2	48.0	14.2	32.7
FINE-TUNING	Multi-task (Q-only)	Ex-NumNet	1.2	13.2	25.1	0.5	6.1	25.1	32.8	2.4	13.3
	Multi-task (C-only)	Ex-NumNet	1.2	14.2	22.8	19.1	0.6	3	0	9.5	8.8
	Single-task	Ex-NumNet	0	37.8	50.8	22.2	66.6	**71.6**	85.9	12.2	43.4
	Multi-task	Ex-NumNet	0	37.5	58	31.4	68.2	70.2	85.7	23.2	46.8
	Multi-task + IR	Ex-NumNet	5.6	37.5	46.6	36.4	68.6	69.6	**85.9**	22.4	46.6
	Multi-task + CIR	Ex-NumNet	7.4	38.8	58	**36.8**	**69.2**	70.8	85.8	**23.6**	**48.8**
	Multi-task + OS	Ex-NumNet	7.4	38.8	47.8	35.9	44.3	53.7	85.4	22.4	42.0
-	-	Human	94.4	94.5	97.8	95	94.7	96.1	96.5	92.8	95.2

Table 2: F1 performance of various baselines on the NumGLUE test set across various tasks 1-8. Human performance was calculated on 100 samples of each task (81 of Task 1) [*IR = Information Retrieval, CIR=Conditional Information Retrieval, OS=Oversampling, Q. Only: Question Only, C. Only: Context Only].

More specifics?

- A memory-augmented neural model called Ex-NumNet is proposed. Training it on all NUMGLUE tasks gives a 3.4% average gain over individual task training, showing the benefits of multi-task learning.
- Analysis reveals the main error types are producing invalid outputs, copying numbers from questions, incorrect calculations, and extraneous text. This suggests better numerical parsing and reasoning is needed.



Error	Ex-NumNet	GPT3
Invalid output	16 %	7%
Copy number	5 %	3%
Incorrect calculation	71 %	56%
Redundant text	8 %	34%

Table 3: Error analysis for the best Ex-NumNet Multi-task+CIR and GPT3 Task-specific model

Wait? Hold on! What’s Multi-Task Learning?

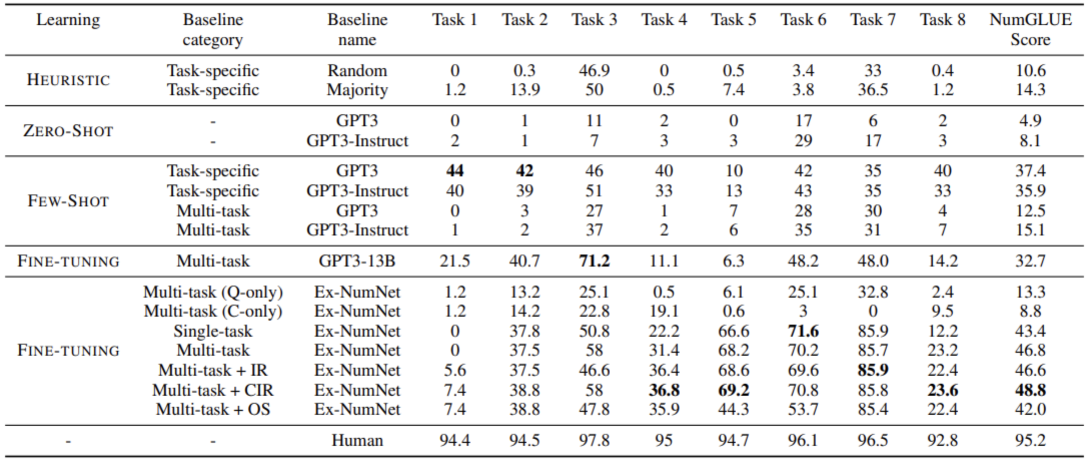
Multi-task learning is a machine learning technique where a single model is trained on multiple related tasks at the same time. The key idea is that by learning multiple tasks together, the shared knowledge across tasks can improve the model's overall performance on each individual task.

Some key points about multi-task learning:

- A single model architecture is developed that can perform multiple different tasks. For example, a single neural network can be trained for image classification, object detection, and segmentation simultaneously.
- The model shares some parameters across all the tasks, while also having some task-specific parameters. This allows it to learn shared representations that encode knowledge common across tasks.
- The model is trained on data from all the tasks jointly in an interleaved manner. The optimization process minimizes a combined loss function that incorporates the losses from each task.
- Multi-task learning introduces an inductive bias that leads the model to prefer hypotheses that explain more than one task. This acts as a regularization and helps improve generalization.
- By sharing knowledge across tasks, multi-task learning can lead to better performance compared to training the model on each task independently. It also reduces the risk of overfitting.
- Multi-task learning is especially useful when there is limited data for some tasks. The shared knowledge from related tasks acts as a useful prior and reduces data needs.

That may be off the track! So in NumGlue?

- In the NUMGLUE paper, jointly training a model on all the arithmetic reasoning tasks led to better performance compared to individual task training. This demonstrates the benefits of multi-task learning for this problem domain. (memory-augmented first , remember !)



Learning	Baseline category	Baseline name	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	NumGLUE Score
HEURISTIC	Task-specific	Random	0	0.3	46.9	0	0.5	3.4	33	0.4	10.6
	Task-specific	Majority	1.2	13.9	50	0.5	7.4	3.8	36.5	1.2	14.3
ZERO-SHOT	-	GPT3	0	1	11	2	0	17	6	2	4.9
	-	GPT3-Instruct	2	1	7	3	3	29	17	3	8.1
FEW-SHOT	Task-specific	GPT3	**44**	**42**	46	40	10	42	35	40	37.4
	Task-specific	GPT3-Instruct	40	39	51	33	13	43	35	33	35.9
	Multi-task	GPT3	0	3	27	1	7	28	30	4	12.5
	Multi-task	GPT3-Instruct	1	2	37	2	6	35	31	7	15.1
FINE-TUNING	Multi-task	GPT3-13B	21.5	40.7	**71.2**	11.1	6.3	48.2	48.0	14.2	32.7
FINE-TUNING	Multi-task (Q-only)	Ex-NumNet	1.2	13.2	25.1	0.5	6.1	25.1	32.8	2.4	13.3
	Multi-task (C-only)	Ex-NumNet	1.2	14.2	22.8	19.1	0.6	3	0	9.5	8.8
	Single-task	Ex-NumNet	0	37.8	50.8	22.2	66.6	**71.6**	85.9	12.2	43.4
	Multi-task	Ex-NumNet	0	37.5	58	31.4	68.2	70.2	85.7	23.2	46.8
	Multi-task + IR	Ex-NumNet	5.6	37.5	46.6	36.4	68.6	69.6	**85.9**	22.4	46.6
	Multi-task + CIR	Ex-NumNet	7.4	38.8	58	**36.8**	**69.2**	70.8	85.8	**23.6**	**48.8**
	Multi-task + OS	Ex-NumNet	7.4	38.8	47.8	35.9	44.3	53.7	85.4	22.4	42.0
-	-	Human	94.4	94.5	97.8	95	94.7	96.1	96.5	92.8	95.2

Table 2: F1 performance of various baselines on the NumGLUE test set across various tasks 1-8. Human performance was calculated on 100 samples of each task (81 of Task 1) [*IR = Information Retrieval, CIR=Conditional Information Retrieval, OS=Oversampling, Q. Only: Question Only, C. Only: Context Only].

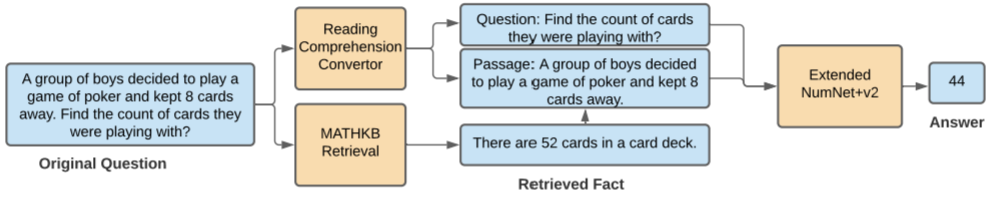

A group of boys decided to play a game of poker and kept 8 cards away. Find the count of cards they were playing with?
Original Question
Reading Comprehension Convertor
MATHKB Retrieval
Question: Find the count of cards they were playing with?
Passage: A group of boys decided to play a game of poker and kept 8 cards away.
There are 52 cards in a card deck.
Retrieved Fact
Extended NumNet+v2
44
Answer

Figure 3: Our proposed memory-augmented model that detects the type of task (1-8), uses Information Retrieval from *MATH KB* and append the information that gets fed to Ex-NumNet

Note!

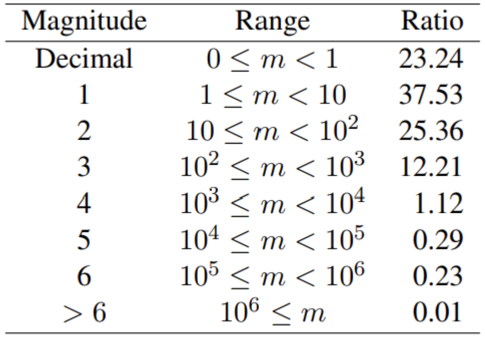
- You may skip the next two papers, if you don’t have time to read!
- But , if you are working with me, I think reading 2023 and 2022 papers would really help!

Next Paper

Numeracy 600k

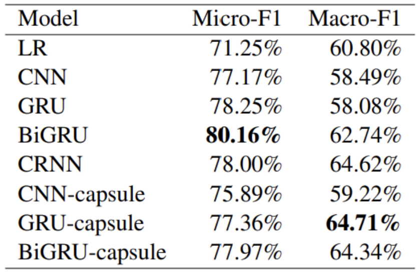
Numeracy 600k

1. Numerical reasoning is a challenging but worthwhile capability for NLP models to learn. There are many potential applications, such as detecting exaggerated claims.
2. Large labeled datasets like Numeracy-600K are valuable for training and evaluating numerical reasoning abilities. Creating datasets for different reasoning tasks is important.



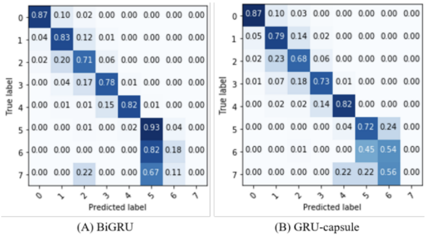
Magnitude	Range	Ratio
Decimal	$0 \leq m < 1$	23.24
1	$1 \leq m < 10$	37.53
2	$10 \leq m < 10^2$	25.36
3	$10^2 \leq m < 10^3$	12.21
4	$10^3 \leq m < 10^4$	1.12
5	$10^4 \leq m < 10^5$	0.29
6	$10^5 \leq m < 10^6$	0.23
> 6	$10^6 \leq m$	0.01

Table 3: Distribution of numerals in the dataset.



Model	Micro-F1	Macro-F1
LR	71.25%	60.80%
CNN	77.17%	58.49%
GRU	78.25%	58.08%
BiGRU	**80.16%**	62.74%
CRNN	78.00%	64.62%
CNN-capsule	75.89%	59.22%
GRU-capsule	77.36%	**64.71%**
BiGRU-capsule	77.97%	64.34%

Table 4: Experimental results.



True label	0	1	2	3	4	5	6	7
0	0.87	0.10	0.02	0.00	0.00	0.00	0.00	0.00
1	0.04	0.83	0.12	0.01	0.00	0.00	0.00	0.00
2	0.02	0.20	0.71	0.06	0.00	0.00	0.00	0.00
3	0.00	0.04	0.17	0.78	0.01	0.00	0.00	0.00
4	0.00	0.01	0.01	0.15	0.82	0.01	0.00	0.00
5	0.00	0.00	0.01	0.00	0.02	0.93	0.04	0.00
6	0.00	0.00	0.00	0.00	0.00	0.82	0.18	0.00
7	0.00	0.00	0.22	0.00	0.00	0.67	0.11	0.00

Predicted label

(A) BiGRU

True label	0	1	2	3	4	5	6	7
0	0.87	0.10	0.03	0.00	0.00	0.00	0.00	0.00
1	0.05	0.79	0.14	0.02	0.00	0.00	0.00	0.00
2	0.02	0.23	0.68	0.06	0.00	0.00	0.00	0.00
3	0.01	0.07	0.18	0.73	0.01	0.00	0.00	0.00
4	0.00	0.02	0.02	0.14	0.82	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.04	0.72	0.24	0.00
6	0.00	0.00	0.01	0.00	0.00	0.45	0.54	0.00
7	0.00	0.00	0.00	0.00	0.22	0.22	0.56	0.00

Predicted label

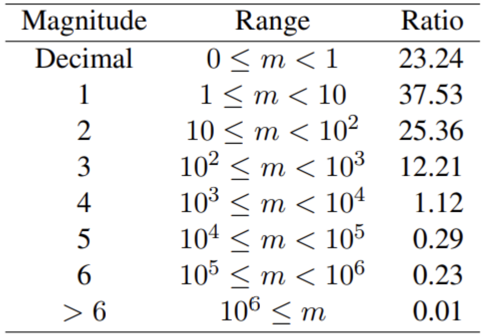
(B) GRU-capsule

Figure 1: Confusion matrices

Numeracy 600k

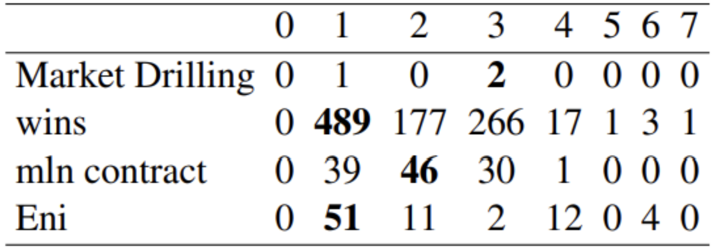
3. RNN architectures like BiGRU seem well-suited to modeling numerical reasoning, but there is room for improvement. Exploring different model architectures is still an open area.

4. Contextual information is crucial for numerical reasoning. Models need to understand cues like units of measurement when predicting magnitudes.



Magnitude	Range	Ratio
Decimal	$0 \leq m < 1$	23.24
1	$1 \leq m < 10$	37.53
2	$10 \leq m < 10^2$	25.36
3	$10^2 \leq m < 10^3$	12.21
4	$10^3 \leq m < 10^4$	1.12
5	$10^4 \leq m < 10^5$	0.29
6	$10^5 \leq m < 10^6$	0.23
> 6	$10^6 \leq m$	0.01

Table 3: Distribution of numerals in the dataset.



	0	1	2	3	4	5	6	7
Market Drilling	0	1	0	**2**	0	0	0	0
wins	0	**489**	177	266	17	1	3	1
mln contract	0	39	**46**	30	1	0	0	0
Eni	0	**51**	11	2	12	0	4	0

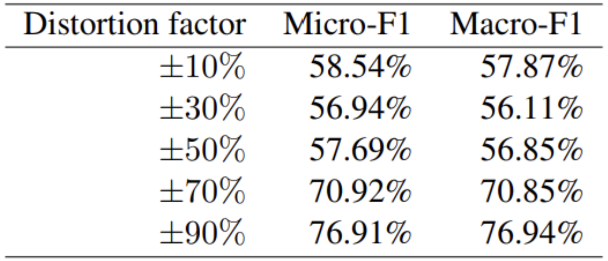
Table 6: Co-occurrence statistics of (E5)

Numeracy 600k

5. Transferring reasoning abilities across different text genres remains difficult. More work is needed on domain adaptation and generalization.

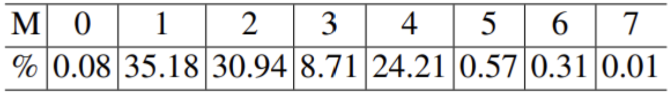
6. Capsule networks show promise on some metrics, suggesting value in using structured vector representations. But more exploration is needed.

7. There are many open challenges, like handling diverse entities and patterns, implied meaning, limited training data, etc.



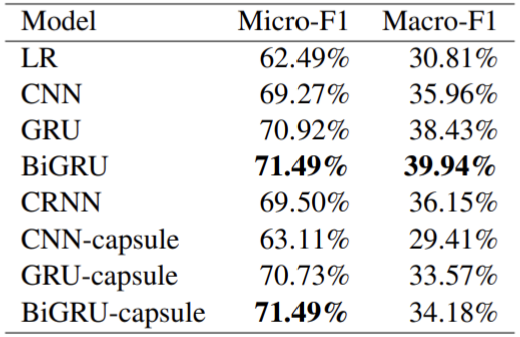
Distortion factor	Micro-F1	Macro-F1
±10%	58.54%	57.87%
±30%	56.94%	56.11%
±50%	57.69%	56.85%
±70%	70.92%	70.85%
±90%	76.91%	76.94%

Table 7: Results for exaggerated numeral detection.



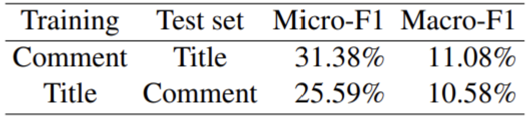
M	0	1	2	3	4	5	6	7
%	0.08	35.18	30.94	8.71	24.21	0.57	0.31	0.01

Table 8: Distribution of numerals in the title dataset. M.: magnitude; 7: M > 6.



Model	Micro-F1	Macro-F1
LR	62.49%	30.81%
CNN	69.27%	35.96%
GRU	70.92%	38.43%
BiGRU	**71.49%**	**39.94%**
CRNN	69.50%	36.15%
CNN-capsule	63.11%	29.41%
GRU-capsule	70.73%	33.57%
BiGRU-capsule	**71.49%**	34.18%

Table 9: Experimental results of titles.



Training	Test set	Micro-F1	Macro-F1
Comment	Title	31.38%	11.08%
Title	Comment	25.59%	10.58%

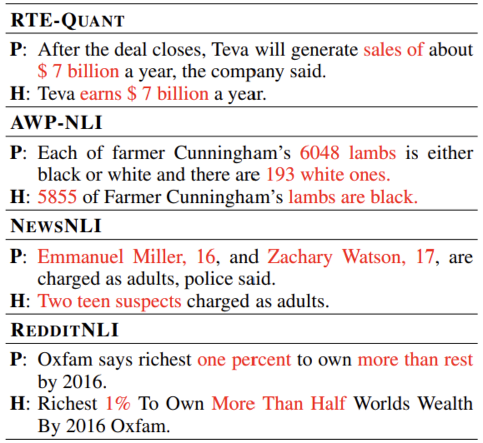
Table 10: Results of learning cross-source numeracy.

EQUATE

- A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference

EQUATE

1. **Developing models that can do deeper reasoning with the interplay between numerical and verbal information is an important challenge area.**
2. **Existing neural models rely heavily on lexical matching signals and fail to adequately perform multi-step quantitative reasoning.**
3. **Hybrid neuro-symbolic architectures that combine neural models with specialized numerical reasoning modules show promise on EQUATE, but still face challenges in complex verbal reasoning.**
4. **The field needs continued benchmarking on datasets like EQUATE that isolate quantitative reasoning, to drive progress.**



RTE-QUANT	
P:	After the deal closes, Teva will generate sales of about $ 7 billion a year, the company said.
H:	Teva earns $ 7 billion a year.
AWP-NLI	
P:	Each of farmer Cunningham's 6048 lambs is either black or white and there are 193 white ones.
H:	5855 of Farmer Cunningham's lambs are black.
NEWSNLI	
P:	Emmanuel Miller, 16, and Zachary Watson, 17, are charged as adults, police said.
H:	Two teen suspects charged as adults.
REDDITNLI	
P:	Oxfam says richest one percent to own more than rest by 2016.
H:	Richest 1% To Own More Than Half Worlds Wealth By 2016 Oxfam.

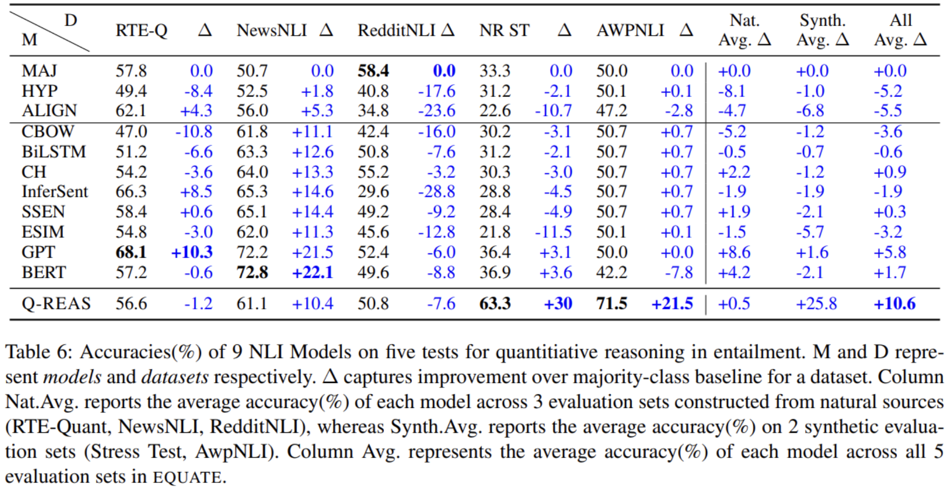
Table 1: Examples from evaluation sets in EQUATE

EQUATE

5. Hard cases require handling the intricate interdependence between numerical and linguistic phenomena like coreference, pragmatics, approximation - this complex interplay needs to be modeled.

6. Lexical inference, generating complete quantity representations, and comparing incompatible quantities also remain open issues.

7. In summary, the key takeaway appears to be that quantitative reasoning in NLI remains a significant challenge area where current models are limited, requiring continued research on hybrid reasoning approaches and benchmarking on tests like EQUATE.



D / M	RTE-Q	Δ	NewsNLI	Δ	RedditNLI	Δ	NR ST	Δ	AWPNLI	Δ	Nat. Avg. Δ	Synth. Avg. Δ	All Avg. Δ
MAJ	57.8	0.0	50.7	0.0	**58.4**	**0.0**	33.3	0.0	50.0	0.0	+0.0	+0.0	+0.0
HYP	49.4	-8.4	52.5	+1.8	40.8	-17.6	31.2	-2.1	50.1	+0.1	-8.1	-1.0	-5.2
ALIGN	62.1	+4.3	56.0	+5.3	34.8	-23.6	22.6	-10.7	47.2	-2.8	-4.7	-6.8	-5.5
CBOW	47.0	-10.8	61.8	+11.1	42.4	-16.0	30.2	-3.1	50.7	+0.7	-5.2	-1.2	-3.6
BiLSTM	51.2	-6.6	63.3	+12.6	50.8	-7.6	31.2	-2.1	50.7	+0.7	-0.5	-0.7	-0.6
CH	54.2	-3.6	64.0	+13.3	55.2	-3.2	30.3	-3.0	50.7	+0.7	+2.2	-1.2	+0.9
InferSent	66.3	+8.5	65.3	+14.6	29.6	-28.8	28.8	-4.5	50.7	+0.7	-1.9	-1.9	-1.9
SSEN	58.4	+0.6	65.1	+14.4	49.2	-9.2	28.4	-4.9	50.7	+0.7	+1.9	-2.1	+0.3
ESIM	54.8	-3.0	62.0	+11.3	45.6	-12.8	21.8	-11.5	50.1	+0.1	-1.5	-5.7	-3.2
GPT	**68.1**	**+10.3**	72.2	+21.5	52.4	-6.0	36.4	+3.1	50.0	+0.0	+8.6	+1.6	+5.8
BERT	57.2	-0.6	**72.8**	**+22.1**	49.6	-8.8	36.9	+3.6	42.2	-7.8	+4.2	-2.1	+1.7
Q-REAS	56.6	-1.2	61.1	+10.4	50.8	-7.6	**63.3**	**+30**	**71.5**	**+21.5**	+0.5	+25.8	**+10.6**

Table 6: Accuracies(%) of 9 NLI Models on five tests for quantitiative reasoning in entailment. M and D represent *models* and *datasets* respectively. Δ captures improvement over majority-class baseline for a dataset. Column Nat.Avg. reports the average accuracy(%) of each model across 3 evaluation sets constructed from natural sources (RTE-Quant, NewsNLI, RedditNLI), whereas Synth.Avg. reports the average accuracy(%) on 2 synthetic evaluation sets (Stress Test, AwpNLI). Column Avg. represents the average accuracy(%) of each model across all 5 evaluation sets in EQUATE.

Collecting our Learnings!

- Let’s collect two, three most general things
- We learn from these four papers !

Collecting our learnings from these papers

- In 2023 paper, Input-Reframing for me is a very inventive way to approach quantitative understanding, of course models like Roberta are more prone to both input-reframing and pre-finetuning approaches because of its dynamic masking capabilities
- 2022 paper presents very interesting details as : a Memory-augmented model (information retrieval method applied from the original question) is trained on all 8 tasks rather than just 1 ! Which improves average gains! : hence, multi-task learning !

Organizing Our Research

- So, how are we going to approach our own research?

Novel Pretraining?

- We could Pretrain the model further on numeric reasoning datasets like ComQA, DuReader, Dolphin18K (based on 2023 paper).
- Explore pretraining on math word problems, numerical common sense knowledge. (common sense knowledge is important since our system will be tested on unknown test dataset- 2022 paper)
- Evaluate different masking strategies for numbers during pretraining.

Model Architecture?

- Incorporate numeric processing modules such as digit embeddings, attention over numbers. (Basically different type of Input Reframing)

- wait , what does it mean by that? Incorporating numeric processing modules means :

1. Having separate embeddings for digits and number words compared to regular text tokens.
2. Attention mechanisms that allow the model to focus on key numbers in the text.
3. Dedicated numeric encoders like convolutional networks to represent numbers.

We could Experiment with MathBERT (huggingface ready) which incorporate numeric awareness.

(will add extra slide later in the end!)

- Add quantity and unit tracking mechanisms in the model architecture.

1. Components to track quantities and units as they flow through text, get transformed.
2. For example, keeping track that "10 kg" refers to the same overall quantity as "10000 g".

Model Architecture Cont’d

- Enable operating over numbers, querying knowledge bases within the model.

1. Arithmetic modules like addition, subtraction networks that allow performing computations.
2. Retrieving formulas, unit conversions from knowledge bases to support numeric operations.
3. Allowing the model to directly predict simple arithmetic expressions as output.

- The goal is to incorporate numeric-specific components tailored for quantities, units, mathematical operations into the model architecture itself.
- This equips the model with extra skills for tracking, transforming and operating over numbers by augmenting the standard NLP encoder-decoder architecture.

Training Techniques?

- Curriculum learning from simple to complex numeric reasoning.

1. Gradually increase the difficulty of numeric reasoning tasks as training progresses.
2. Start with simple number comparison, then addition/subtraction, then word problems requiring multiple steps.
3. Allows the model to slowly build up complex numeracy skills.

- Continued pretraining on in-domain datasets.

1. Quite simple, Further pretrain the model on datasets related to the target domain and tasks.
2. Helps adapt the general knowledge from broad pretraining to the niche domain.

- Jointly learn to reason over text and structured knowledge.

1. Multi-task or joint training to reason over text and structured knowledge.
2. For example, learn to link text entities to knowledge base constants.
3. Jointly predict text expressions and symbolic representations.
4. Helps fuse textual and structured reasoning.

- **The main focus is on specialized training schemes that can help improve numeric and mathematical reasoning beyond generic pretraining.**
- **Curriculum learning, targeted pretraining, and joint reasoning over multiple knowledge types are key techniques to explore.**

Evaluation?

- Benchmark performance over all NumEval subtasks.

1. Evaluate on all subsets - QP, QNLI, QQA
2. Allows comprehensive assessment on different skills

Very Simple! Each three subtask run tests!

- Test performance on math word problems.

1. Word problems require integrating text and numbers
2. Involve multiple steps of symbolic reasoning
3. Test general math problem solving abilities

Test with math word problems which means math problems written in words instead of real numbers!

Evaluation Cont’d

- Evaluate on complex numerical reasoning requiring multiple steps.

1. Require understanding relationships between quantities
2. Apply mathematical operations sequentially
3. Assess reasoning skills beyond basic numeracy

- Analyze errors to identify model limitations.

1. Identify categories of mistakes the model makes
2. Wrong operations, quantity errors, incorrect equations
3. Pinpoint limitations of the model's math abilities

Our goal here is to rigorously measure the model's quantitative reasoning capacities on diverse tasks and examples. Testing on word problems and multi-step reasoning is important to evaluate advanced skills. **Detailed error analysis provides insights into model weaknesses.**

This allows thoroughly evaluating where the model succeeds at numeric reasoning and where it still struggles compared to human abilities.

Why suggest **numericalBERT, NumNet, MathBERT ? which incorporate numeric awareness? (maybe because they have different model architecture than the original?)**

When I suggested experimenting with models like numericalBERT, NumNet, and MathBERT, here is what I meant by that:

- These are models that have been designed with specific modifications and pretraining objectives to make them more "numerically aware" compared to a standard BERT-style model.
- Some examples of how they incorporate numeric awareness:
- - numericalBERT: Has separate embeddings for numbers which helps it better understand magnitude and decimal points.
- - NumNet: Adds modules like a counting module and comparison module to explicitly model numeric reasoning.
- - MathBERT: Pretrained on math word problem datasets to acquire better mathematical reasoning abilities.

So in essence, these models build in numeric-specific components through architecture modifications and math-focused pretraining. We can learn from how these models are built and even build our own one with similar components!

So in summary, these models provide a head start for numeric reasoning that custom models likely will not match without explicit architectural and pretraining enhancements.

Why **MathBERT ?**

- By experimenting with them, hopefully we can:

- Leverage their innate numeric reasoning abilities for our task
- Fine-tune them on our datasets to adapt them
- Use them as starting points for further enhancements
- Analyze differences in their capabilities based on design choices

So, Basically, the goal would be to take advantage of the "hardcoded" numeric awareness in these models and then build upon it further using techniques like novel pretraining objectives suggested earlier.

Any Ideas and Comments?

- I would really appreciate your help and contributions to this project! <3

Research Presentation End