NUMGLUE: A Suite of Fundamental yet Challenging Mathematical Reasoning Tasks

Swaroop Mishra[1]   **Arindam Mitra**[2]   **Neeraj Varshney**[1]   **Bhavdeep Sachdeva**[1]
Peter Clark[3]   **Chitta Baral**[1]   **Ashwin Kalyan**[3]

[1]Arizona State University   [2]Microsoft Research   [3]Allen Institute for AI

Abstract

Given the ubiquitous nature of numbers in text, reasoning with numbers to perform simple calculations is an important skill of AI systems. While many datasets and models have been developed to this end, state-of-the-art AI systems are brittle; failing to perform the underlying mathematical reasoning when they appear in a slightly different scenario. Drawing inspiration from GLUE (Wang et al., 2018) that was proposed in the context of natural language understanding, we propose NUMGLUE, a multi-task benchmark that evaluates the performance of AI systems on eight different tasks, that at their core require simple arithmetic understanding. We show that this benchmark is far from being solved with neural models including state-of-the-art large-scale language models performing significantly worse than humans (lower by 46.4%). Further, NUMGLUE promotes sharing knowledge across tasks, especially those with limited training data as evidenced by the superior performance (average gain of 3.4% on each task) when a model is jointly trained on all the tasks as opposed to task-specific modeling. Finally, we hope that NUMGLUE will encourage systems that perform robust and general arithmetic reasoning within language, a first step towards being able to perform more complex mathematical reasoning[1].

1 Introduction

Reasoning with numbers is an important skill that occurs in various day-to-day scenarios and not surprisingly, numbers are ubiquitous in textual data. To train AI reasoning systems that can perform simple mathematical reasoning, many tasks have been proposed (Dua et al., 2019b; Ravichander et al., 2019; Koncel-Kedziorski et al., 2016). Despite these efforts, current state-of-the-art AI

> **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.

systems are brittle and fail when problems involving similar mathematical reasoning is posed in a slightly different manner. For instance, presenting a word problem in a different manner as shown in fig. 1, while hardly affecting human performance, is sufficient to confuse state-of-the-art AI systems[2]. This brittleness in reasoning indicates that the models latch on to spurious signals in the specific dataset resulting in "solving" the dataset while not truly understanding the underlying reasoning skill of simple arithmetic. Further, we believe that building AI systems that can truly understand and apply simple arithmetic reasoning is a mandatory first step towards successfully tackling complex

---

[1]https://allenai.org/data/numglue

[2]The recently released GPT3-Instruct, a fine-tuned model with 175B parameters produces inconsistent answers for these questions. See supplementary material: GPT3-Instruct's Response for more details.

mathematical reasoning skills (Saxton et al., 2019; Hendrycks et al., 2020, 2021).

NumGLUE. To this end, we propose NUMGLUE, a multi-task benchmark consisting of eight different tasks that at their core test for arithmetic reasoning skills. For example, as discussed in fig. 1, tasks can involve word problems presented in a slightly different manner or can involve additional reasoning strategies like commonsense reasoning or reading comprehension to be combined with the core skill of simple arithmetic. Our benchmark consists of four new tasks in addition to four existing ones; with $\sim 100K$ problems spread across eight differet tasks. The motivation behind NUMGLUE is similar to GLUE (Wang et al., 2018, 2019), a multi-task benchmark that aimed at models that demonstrated superior language understanding by learning the underlying linguistic features. NUMGLUE is designed with goal of progressing towards AI systems that are capable of performing arithmetic reasoning in a general setting; achieving superior performance on our benchmark requires the ability to correctly identify and perform the underlying arithmetic reasoning without relying on task or dataset-specific signals. Finally, we hope that NUMGLUE will encourage systems that perform robust and general numeric reasoning within language, a first step towards being able to perform more complex mathematical reasoning.

Contributions.

1. We introduce NUMGLUE– a multi-task benchmark consisting of eight different tasks, including 4 new ones, whose solution at its core requires an understanding of simple arithmetic.

2. We demonstrate that NUMGLUE is a challenging benchmark even for state-of-the-art large scale language models, obtaining poor scores not only in zero or few shot settings but also after fine-tuning. This indicates a fundamental barrier for AI systems; one that needs to be breached before complex mathematical challenges can be successfully tackled.

3. Finally, we propose a memory-augmented neural model to demonstrate the utility of such a multi-task meta dataset. Our proposed model when trained on the entirety of NUMGLUE obtains an average improvement of 3.4% on each task as opposed to task-specific training – in-

dicating that joint training leads to beneficial transfer owing to the common theme of arithmetic reasoning.

2   Related Work

Datasets for Numerical reasoning. Quantitative reasoning has been a challenging problem for a long time. Small question answering datasets were proposed to understand the quantitative aspect of natural language such as the template-based dataset which solved questions with equations as parameters (Kushman et al., 2014), addition-subtraction dataset (Hosseini et al., 2014) and arithmetic problems dataset (Koncel-Kedziorski et al., 2015). Difficulty of questions were increased in subsequent datasets (Roy and Roth, 2016), (Upadhyay et al., 2016). Later, larger datasets were created to facilitate deep learning research (Ling et al., 2017; Dua et al., 2019b). Several other maths datasets have been proposed to improve explainability (Amini et al., 2019), diversity (Miao et al., 2020), scale information in language embeddings (Zhang et al.) and hardness of math questions (Hendrycks et al., 2021).

One of the motivations behind creating this benchmark is to test for simple arithmetic reasoning independent of the context or the presentation style of the problem. Further, To the best of our knowledge, our work is the first to consider multiple tasks in the numerical reasoning space.

Multi-Task Benchmarks.   With increased success of deep learning based models on individual tasks, there has been a significant push both in the NLP community and in the broader AI community towards general purpose models that excel at multiple tasks. Naturally, various benchmarks and challenges that test for such understanding have been proposed. For instance, the BAbI dataset (Weston et al., 2015), GLUE (Wang et al., 2019) and the subsequent harder SuperGLUE (Wang et al., 2019) were proposed to both evaluate and drive progress in language understanding via shared linguistic knowledge across tasks. McCann (McCann et al., 2018) build a multi-task dataset via a novel approach – formatting each task as that of question-answering. In the more restricted setting of reading comprehension, Dua et al. (2019a) and Downey and Rumshisky build a meta-dataset that spans multiple domains

and reasoning skills.

Multi-task Models. With the growing interest towards models that go beyond specific datasets, various neural models that can perform mutliple tasks have been proposed. When the underlying reasoning is similar – eg. commonsense reasoning, problem decomposition or linguistic understanding – it has been found that training on multi-task datasets yields more robust and accurate models. For instance, the Multi-task Question Answering Network (McCann et al., 2018), T5 (Raffel et al., 2019), GPT3 (Brown et al., 2020) and GPT3-Instruct models aim to build general purpose language models that are capable of transferring linguistic understanding across tasks. A similar approach is taken by Khashabi et al. (2020) in the setting of question-answering and Lourie et al. (2021) in the scope of commonsense reasoning. Further, Muppet (Aghajanyan et al., 2021) adds an additional step of pre-finetuning between pretraining and finetuning that improves generalization to multiple tasks.

3 NUMGLUE

As mentioned previously, our NUMGLUE benchmark consists of both new and already existing arithmetic reasoning tasks. We first begin by introducing the novel datasets curated by us before providing a brief overview of existing tasks that are part of NUMGLUE. Finally, in this section, we provide an analysis of the datasets demonstrating that it contains interesting and diverse linguistic and mathematical properties.

NUMGLUE Benchmark. Our proposed NUMGLUE benchmark is a collection of eight different tasks that together include ~$100K$ questions. The tasks may either be self-contained or require additional background knowledge (e.g.commonsense reasoning) to arrive at the final solution; however, all the tasks, at their core, involve arithmetic reasoning. Table 1 shows an example question belonging to each task along with indicating the total number of data points associated with each task. It is important to note that tasks are imbalanced with only ~400 examples for Task 1 and nearly $50K$ questions under Task 5. While we could have under-sampled the questions to create a balanced suite, we retain the imbalanced dataset in order

to mimic the real world – for instance, arithmetic word problems are more abundant as opposed to word problems that may require commonsense reasoning in addition to arithmetic reasoning.

Data Partition and Evaluation. We randomly partition data in each task into training (70%), development (10%) and test (20%) sets . In the case of reading comprehension tasks (Task 5 and 6), we assign all questions corresponding to a passage to the same split – we do this in order to discourage any data leakage and thereby, allowing models to potentially rely on memorization to arrive at the correct answer.

For each task, we report the F1 measure and as an aggregate measure of performance on the NUMGLUE benchmark similar to Dua et al. (2019b), we report the (unweighted) average of the F1 scores corresponding to each task.

3.1 Novel Datasets

The novel tasks proposed as part of NUMGLUE are a combination of both freshly collected data and intelligent modifications of already existing datasets. The four novel arithmetic reasoning tasks introduced are as follows [3]:

Task 1: Commonsense + Arithmetic Reasoning. Consider the following question – *How many faces do 10 dice have?* Answering this not only requires simple arithmetic i.e.multiplying the number of faces in a die by ten but also requires knowing that a standard die has six faces. We collect this dataset by first asking the annotator to write down a numerical commonsense fact (e.g.a human has 2 hands, a day has 24 hours *etc.*) and then use frame a question that requires using this numerical fact as part of a simple arithmetic calculation.

Task 2: Domain Specific + Arithmetic Reasoning. *How many units of hydrogen are required to produce 10 units of water?* This question, similar to the previously introduced task of arithmetic reasoning questions, requires additional domain-specific knowledge – specifically, that each unit of water contains two units of hydrogen. We

---

[3] We annotate the datasets manually. We provide the exact flow used to generate questions of each task in the supplementary materials: Construction of NUMGLUE.

Task	Question Setting	Size	Example
TASK 1	Commonsense + Arithmetic	404	Question: A man can lift one box in each of his hands. How many boxes can a group of 5 people hold in total? Answer: 10
TASK 2	Domain specific + Arithmetic	1620	Question: How many units of H_2 are required to react with 2 units of C_2H_4 to form 2 units of C_2H_6? Answer: 2
TASK 3	Commonsense + Quantitative	807	Question: A person wants to get shopping done quickly. They know that they can get through the check-out at big store in 5 minutes whereas it can take 20 minutes at small store. The store they go to finish quickly is? (A) big store (B) small store? Answer: big store
TASK 4	Fill-in-the-blanks	1100	Question: Joan found 70 seashells on the beach. She gave Sam some of her seashells. She has 27 seasshells left. She gave _____ seashells to Sam? Answer: 43
TASK 5	RC + Explicit Numerical Reasoning	54212	Passage: <>. Question: How many counties were added in 1887? Answer: 2
TASK 6	RC + Implicit Numerical Reasoning	32724	Passage: <>. Question: Which player kicked the shortest field goal? Answer: David Akers
TASK 7	Quantitative NLI	9702	Statement 1: James took a 3 - hour bike ride, Statement 2: James took a more than 1 - hour bike ride, Options: Entailment or contradiction or neutral?, Answer: Entailment
TASK 8	Arithmetic word problems	1266	Question: Joe had 50 toy cars. If he gives away 12 cars, how many cars will he have remaining?, Answer: 38

Table 1: Size and example of each task in the NumGLUE benchmark. RC: Reading Comrehension

curate a dataset of such questions that require both domain-specific knowledge and arithmetic reasoning motivated by the finding that QA systems perform poorly on the ARC dataset Clark et al. (2018) consisting of grade-school level science questions. Specifically, the dataset collected by us requires understanding of a small set of chemistry (conservation of mass in chemical reactions) and physics principles ($speed = {distance}/{time}$).

Task 3: Commonsense + Quantitative Comparison. *A golf ball weighs 40g and a baseball weighs 150g. Which has a higher gravitational force?* Answering this question requires both knowing that mass is directly proportional to gravitational force and a numerical comparison via subtraction. We collect such quantitative comparison questions by using the QuaRel dataset (Tafjord et al., 2019) containing questions from diverse fields such as physics and economics as the starting point. The annotator chooses a subset of these questions that involve numerically comparable quantities (for instance, in this example, mass of the objects involved) to create the required task of quantitative comparison questions.

Task 4: Fill-in-the-blanks Format. Unlike the previously proposed tasks that require external in-

formation (e.g.commonsense knowledge) in addition to simple arithmetic reasoning, this task is self-contained but a stylistic variant of existing math word problems. We source word problems from the Arithmetic Word Problem repository (Roy and Roth, 2016, 2017, 2018) and convert them into the fill-in-the-blanks format. For an example of such a conversion, refer to fig. 1.

3.2 Existing Datasets

We now review existing datasets while discussing any modifications made when including them in NUMGLUE. In general, for all the datasets included, we perform a filtering step to clean and control for the quality of the data points being included. This step includes – a) discarding questions that do not have answer annotations b) eliminating questions with high lexical overlap with the remainder of the dataset and c) fixing any type mismatches present in the data (e.g."7.0 students" → "7 students").

Task 5: Reading Comprehension (RC) + Explicit Numerical Reasoning. We select a subset from the DROP (Dua et al., 2019b) dataset to create this task. Specifically, the selected questions involve reading comprehension and numerical reasoning but importantly, the required

answer is also a number.

Task 6: Reading Comprehension (RC) + Implicit Numerical Reasoning. Consider the following question based on a relevant passage – *Which state has the highest income tax rate?* Here, while the final answer is a name, arriving at it requires performing comparison (i.e.subtraction). We classify such questions in the DROP dataset as a separate task in NUMGLUE.

Task 7: Quantitative NLI EQUATE (Ravichander et al., 2019) introduces quantitative NLI questions that require simple arithmetic calculations to be performed in order to accurately classify the relationship between the provided premise and the hypothesis. As noted in fig. 1, many word problems can also be easily converted to this format and is therefore, a diverse and interesting task for evaluating arithmetic reasoning skills of AI systems.

Task 8: Arithmetic Word Problems Finally, we arrive at one of the earliest and extensively studied class of arithmetic reasoning problems i.e.word problems. The specific dataset included as part of our NUMGLUEbenchmark is a combination of multiple datasets proposed by Koncel-Kedziorski et al. (2016), (Koncel-Kedziorski et al., 2015) and Kushman et al. (2014). Further, to ensure that the benchmark as a whole is diverse, we eliminate questions that have a high sentence similarity with questions from the fill-in-the-blanks task.

3.3 Data Quality Analysis:

In order to ensure a high-quality test set, three independent annotators evaluate each question in the test set across all tasks. A tiny porton of the data marked as invalid or with disagreement between the annotators was excluded, resulting in a verified, high-quality NUMGLUE evaluation suite. We also perform a variety of analysis and find that the novel question tasks we created (task 1-4) have higher quality than the existing question tasks since they have higher average vocabulary (number of unique words per number of samples), higher number of unique nouns, verbs and other POS tags and have less semantic textual similarity among each other (indicating lower repetition). Detailed analysis can be found in the supplementary material: Data Quality Analysis of NUMGLUE.

4 Experiments

In this section, we establish multiple baselines on our benchmark and discuss their performance.

4.1 Baselines

We evaluate several baselines on our benchmark – (i) Heuristic, (ii) Zero-shot, (iii) Few-shot, (iv) Fine-tuning and (v) Human. We use two kinds of model architectures (i) Neuro-symbolic, a memory augmented *novel* architecture that extends Numnet+v2 (Ran et al., 2019) and (ii) End-to-end, GPT3 (Brown et al., 2020).

Architectures. In the multi-task setting where the same model is trained on all the NUMGLUE tasks, we use Reading Comprehension (RC) as the common format – converting each task to RC format via a set of hand-coded rules [4]. In addition to being capable of faithfully representing all the constituent tasks, the RC format also allows us to inject additional context in the IR setting without affecting the rest of the pipeline [5]. On the other hand, GPT3 being a generative model does not require such modifications. Importantly, note that both models are inputted the exact same information for the multi-task experiments.

Heuristic Baselines with Task Oracle. For this baseline, we assume a task oracle that knows the task a particular question belongs (in a multi-task setting) – we use this to make our heuristic baselines more competitive. The first heuristic baseline is *random*: we randomly select one of the options in case the question has multiple options (task 3 and 7), a number between 0 to 100 for questions having a numerical answer and a random entity present in the passage for questions having a text segment from the passage as the answer. In the *majority* baseline, we select the most frequent answer for each task such as "Entailment" for NLI questions and similarly, the most frequent number for questions having numerical answer and the major entity present in the passage for questions having span based answer. As the task information is known, we include these baselines under task-specific baselines when discussing results.

---

[4]More details in the supplementary material: Ex-NumNet
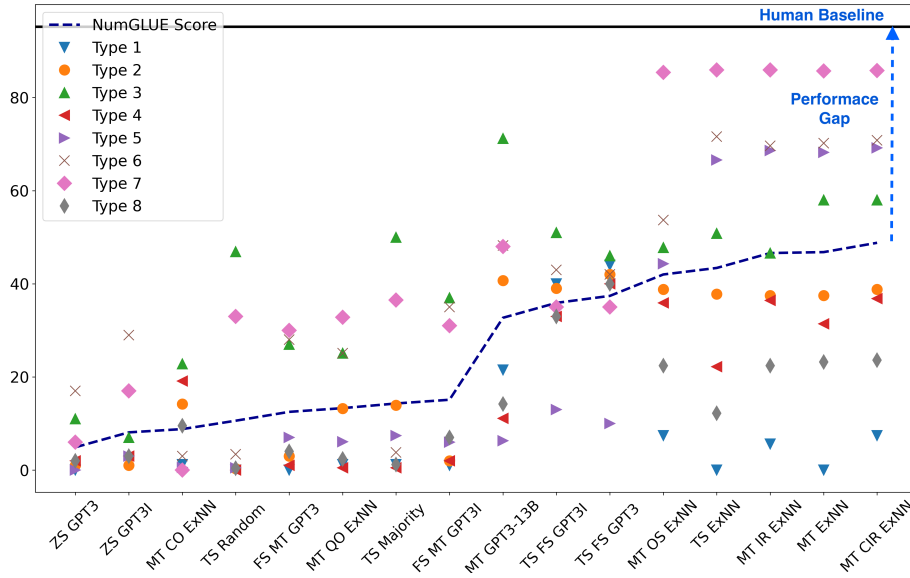[5]Henceforth we will be calling our extension to Numnet+v2 as Ex-NumNet

Figure 2: Performance of zeroshot, fewshot and finetuning baselines (Section 4) across NumGLUE. There is a signficant gap between the highest performing model and the human baseline. ZS: Zeroshot, GPT3I: GPT3-Instruct, MT: Multi-task, TS: Task-specific, QO: Question Only, CO: Context Only, EXNN: Ex-NumNet,FS: Fewshot, OS: Oversampling, IR: Information Retrieval, CIR: Conditional Information Retrieval.
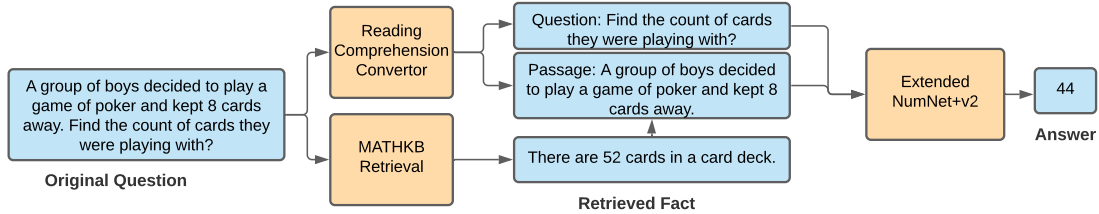

Figure 3: Our proposed memory-augmented model that detects the type of task (1-8), uses Information Retrieval from *MATH KB* and append the information that gets fed to Ex-NumNet

Zeroshot and Fewshot Baselines. We use GPT3 (Brown et al., 2020) and the more recent GPT3-Instruct[6]. We have two types of few shot baseline (i) task specific and (ii) multi task. In case of task specific fewshot baseline, instances of the same task are used as in-context examples (Brown et al., 2020) whereas in case of multitask few shot baseline, instances from all tasks are used to condition the model. Multitask fewshot is naturally a harder setting as it is task-agnostic. We use default parameters in GPT3 and GPT3-Instruct. In few-shot setting, we experiment after feeding as many examples as it can fit within the tokensize. For few shot experiments, we randomly select

examples and averaged the results over 5 runs.

Fine-tuning Baselines. We first consider variations of the fine-tuning baselines in the context of our neuro-symbolic model, Ex-NumNet.

We use it as bias-checking baseline – to ensure that solving the benchmark correctly requires considering all of the information presented to it. To this end, we evaluate the performance of our model when finetuned only on the question (Q-only) or the context (C-only). Next, we present task-specific and multi-task baselines where Ex-NumNet is fine-tuned on individual tasks and the entire NUMGLUE benchmark respectively. With the goal of addressing the data imbalance across the tasks, we include an oversampling

---

[6]newly released by OpenAI as part of the GPT3 finetuned series

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].

baseline that oversamples data from tasks with limited data so as to ensure that the model sees the same number of examples from each constituent task.

In addition, we propose a new architectural modification to Ex-NumNet. Noting that our baseline model Ex-NumNet does not take into account external knowledge, we create a new enhanced architecture in the form of a memory-augmented model that does Information Retrieval (IR) (Khot et al., 2019) with respect to a knowledge base we create, *MATH KB* to identify the needed knowledge. This is inspired by the observation that formula book and mathematical knowledge make the task easier for humans while solving math questions of various types. We then use this knowledge in the Ex-NumNet setting. Figure 3 illustrates our approach which leverages our newly created knowledge base *MATH KB*. Conditional IR model is different from the regular IR model in the sense that, IR is performed only for questions of task 1 , 2 and 4, since they require external knowledge to get answered. More details about the model and the IR process can be found in supplementary material: Proposed Memory-Augmented Model (A.5 and A.6).

Finally, we discuss fine-tuning baselines in the context of end-to-end models, specifically GPT3. We finetune the GPT3-13B model (for which the finetuning capability has been recently provided by OpenAI [7]) in the multi-task setting i.e. the desired setting of the NUMGLUE benchmark.

Human Baseline. Human baseline was calculated on 100 test set samples of each task (81 of Task 1) by averaging the scores of four annotators.

5 Results and Discussion

Table 2 shows the performance of various baseline models on the test set of our benchmark. Note that the performance of all baseline models is significantly lesser than the human baseline (Figure 2). We now discuss various insights based on these results.

Does the benchmark contain bias that a model can exploit? A challenging dataset requires the model to ideally consider all the information provided to it before arriving at an answer. To ensure that this is indeed the case, we perform ablations where only one portion of the input is provided i.e. either the question or the context. Both these "bias-checking" baselines perform poorly even in task-specific setting – indicating that both the benchmark and constituent tasks are challenging.

Which Tasks are Hard to Solve? Our results show that task 1 which requires numerical commonsense knowledge, is the hardest task to solve. Similarly, tasks 2, 4 and 8 appear to be

comparatively harder from the rest. One pattern among these tasks is that all of them expect the answer to be numeric. Numeric answer requires accurate calculation. So, models might have difficulty in learning the task directly from data. This hypothesis is also justified from the *slight* drop in human performance in these tasks..

On the other hand, task 7 has the best performance among all. Further, we see that performance on task 6 is slightly better than task 5 – although both tasks are sourced from the same dataset, we observe that models answer span based questions better as compared to numeric answers. Relatively higher performance for task 3 suggests that models find it easier to answer in an MCQ setting.

Does IR Help? Results show that knowledge help in improving performance of tasks 1, 2 and 4 – where indeed, external knowledge like commonsense or domain-specific knowledge is needed in addition to arithmetic reasoning to arrive at the correct answer. However, task 3 is an exception to this trend and in fact registers a drop in the score when provided with (unnecessary) additional information; we find that this shortcoming is fixed when using conditional information retrieval (CIR) which in fact leads to the strongest baseline presented in this work.

Does Oversampling help overcome data imbalance across tasks? Even though oversampling results in higher performance in certain tasks (in comparison with the multitask baseline), specifically the ones with smaller training data, it results in significant drop in performance in the other extreme, i.e tasks with bigger training data. Also, it never performs better than the Conditional IR module in multitask setting.

5.1 Error Analysis

We now present an analysis of the errors made by our baselines to indicate potential avenues for future research.

We analyze errors associated with 50 samples each of the 8 tasks and find that there are mainly 4 categories of error models make: (1) producing invalid output (e.g. answering text where the answer is supposed to be a number, answering a text different from the classes allowed in a classification problem), (2) copying a number

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 Multitask+CIR and GPT3 Task-specific model

from the question instead of calculating the answer, (3) incorrect calculation – this can be due to multiple reasons including (i) using an incorrect operation e.g. subtraction in place of addition, (ii) incorrect parsing of numbers or (iii) incorrect knowledge of numerical commonsense facts. (4) producing redundant text after producing correct answer. Based on error distribution in Table 3, we observe that the majority of errors come from incorrect calculation. Further, GPT3 is better than Ex NumNet+v2 in producing valid outputs, but it produces more redundant text.

Future Directions: Bigger model, more data or ... ? Table 2 shows that fine-tuned GPT3-13B outperforms other baselines on task 1, 2 and 3. Recall that these tasks require external knowledge and perhaps, this is the reason why GPT3, already pre-trained on a diverse web-scale text corpus has an edge over other baselines on these tasks. In case of the smaller Ex-NumNet, it is interesting that multitask baselines are higher than the single task baselines by 3.4% on average and that information retrieval helps in tasks that require external knowledge. Also notice that, GPT-3 is better on smaller datasets and NumNet is better on large datasets. This may indicate that GPT-3 is a better few-shot learner but not necessarily a better many-shot learner. This non-overlapping performance of GPT-3 and Ex-numnet, end-to-end and neuro-symbolic models respectively, indicates that a potential future direction for research is to combine the best of both the models.

6 Conclusion

We propose NUMGLUE, a multi-task benchmark to test for arithmetic understanding. Our benchmark consists of eight tasks including four new ones. While some of the tasks require external knowledge like commonsense or domain-specific information in addition to arithmetic reasoning, some are self-contained e.g. arithmetic word problems. Further, we demonstrate that our benchmark

is far from being solved – with state-of-the-art large scale models achieving considerably lower performance than humans. This indicates that current AI systems are incapable of performing simple arithmetic reasoning in a general setting – indicating a fundamental hurdle towards AI systems that understand complex mathematical concepts like differential equations or combinatorics. Finally, we present various baselines including a novel architecture (memory augmented Ex-NumNet) that demonstrate the advantages of various modeling choices (e.g. end-to-end vs neuro-symbolic models). Specifically, we show that training in the multi-task setting leads to meaningful sharing of knowledge across tasks as evidenced by an average gain of 3.4% on tasks compared to task-specific modeling. Finally, we hope that our benchmark not only leads to AI systems that are capable of performing simple arithmetic reasoning in a fairly general setting but also results in progress towards more complex mathematical reasoning capability.

Acknowledgements

Ethical Considerations

We have verified that all licenses of source datasets used in this paper allow for their use, modification, and redistribution in a research context. The dataset will be distributed in a manner similar to SuperGLUE (Wang et al., 2019) i.e. give full credit assignment to the original data and task creators.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Anjana Arunkumar, Swaroop Mishra, Bhavdeep Sachdeva, Chitta Baral, and Chris Bryan. 2020. Real-time visual feedback for educative benchmark creation: A human-and-metric-in-the-loop workflow.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Anna Rogers Olga Kovaleva Matthew Downey and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Sameer Singh, and Matt Gardner. 2019a. Orb: An open reading benchmark for comprehensive evaluation of machine reading comprehension. *arXiv preprint arXiv:1912.12598*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019b. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *In Conference on Empirical Methods in Natural Language Processing (EMNLP*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019. What's missing: A knowledge gap guided approach for multi-hop question answering. *arXiv preprint arXiv:1909.09253*.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. Mawps: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*.

Swaroop Mishra and Bhavdeep Singh Sachdeva. 2020. Do we need to create big datasets to learn a task? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 169–173, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. Numnet: Machine reading comprehension with numerical reasoning. *arXiv preprint arXiv:1910.06701*.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*.

Subhro Roy and Dan Roth. 2017. Unit dependency graph and its application to arithmetic word problem solving. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Subhro Roy and Dan Roth. 2018. Mapping to declarative knowledge for word problem solving. *Transactions of the Association for Computational Linguistics*, 6:159–172.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.

Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih. 2016. Learning from explicit and implicit supervision jointly for algebra word problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 297–306.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales?

A   Supplemental Material

A.1   NUMGLUE vs Other Datasets:

As figure 4 shows, we select each task from one of the clusters of numerical reasoning datasets (except the multi-model reasoning cluster since we wanted to limit our dataset to text only).

A.2   Construction of NUMGLUE :

Figure 5 and 6 illustrate detailed data creation process for task 1, task 2, task 3 and task 4 questions with the help of an example for each task. We follow the same procedure for creating other examples within the task.

A.3   GPT3-Instruct's Response

We used GPT3-Instruct on various forms of a simple arithmetic question. An expert did tuning of various parameteres such as temperature, stop condition, presence penalty, engine, maximum token size. However, GPT3-Instruct still could not solve the basic aritmetic questions reliabily (Figures 7-11).

A.4   Data Quality Analysis of NumGLUE

In this section, we discuss various linguistic and statistical properties of our benchmark; ones that we believe result in the quality, diversity and challenging nature (Gururangan et al., 2018; Mishra et al., 2020; Mishra and Sachdeva, 2020; Swayamdipta et al., 2020; Arunkumar et al., 2020) of the proposed NUMGLUE benchmark.

Vocabulary Size.   First, we calculate vocabulary size of each task by finding the number of unique words across all questions. Since our dataset is unbalanced in terms of question task, we find the average vocabulary size by dividing vocabulary size with number of data in that task.

Which Data has Higher Average Vocabulary? As illustrated in Figure 12a, most of the tasks belonging to the novel dataset category have relatively better average vocabulary size. This implies questions in those tasks have less repetitiveness. Furthermore, we expand our vocabulary analysis to understand Figure 12a better. We dive deep to analyze different parts of speech. Figure 12b summarises our analysis. Most of the novel datasets have more average number of nouns, verbs and adjectives implying there are more varieties of entities, actions and

attributes. This further means that datasets belonging to the novel category are more diverse in nature.

Sentence Similarity Analysis We extend our analysis to reinforce our inference from the word vocabulary analysis. We find Semantic Textual Similarity (STS) of a sentence with every other sentence.

Which Data Consists of Most Dissimilar Sentences? As depicted by Figure 12c-12f, most questions in QuaRel have high similarity value with other questions indicating the repetitiveness of data. Same is true for majority of EQUATE data. DROP also has high similarity among questions. However, similarity among questions in our dataset is significantly less. Some similarity boxes can be seen in the chart. They are mostly due to task 2 data, and partly due to task 3 data. Lesser similarity implies that our dataset is far less repetitive than others. Also, the repetition in our dataset is sparse and is not equally distributed among the whole dataset unlike others. This way, our dataset is more diverse.

Note that question in Task 2 have lower vocabulary and further, a higher similarity as well. As a small set of chemistry and physics principles are used to generate questions, the result is a fairly templated or uniform-looking dataset – leading to the observed reversal of trends in this particular task.

A.5   Ex-NumNet

Figure 13 illustrates our baseline model: Ex-NumNet. This contains a Reading Comprehension Converter module which converts each task of question to reading comprehension format. Figure 14 illustrates various examples of how each task of questions get converted to the reading comprehension format. We add a task converter module to detect task of a question. We design task converter heuristically based on the features associated with questions (e.g. NLI contains "Sentence 1" and "Sentence 2" whereas completion contains a blank). We convert each of the tasks to RC format. For NLI questions, we use the premise sentence as passage, hypothesis as the question and append the string "Entailment, contradiction or neutral?"
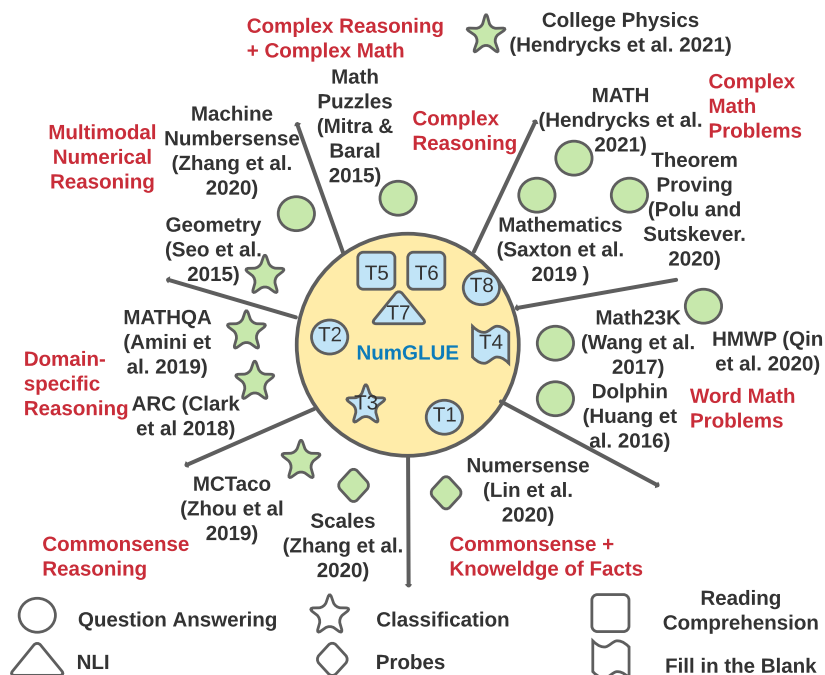
Figure 4: Our dataset NUMGLUE (center in the yellow circle) has been positioned with respect to existing datasets. T1-T8 represents 8 tasks. Note that, NUMGLUE contains the feature of being format invariant unlike other datasets. Position of datasets within clusters is done based on their semantic category, for example T1 Numerical Commonsense QA is closer to the cluster of Commonsense Reasoning + Knowledge of Facts; its position reflects the same
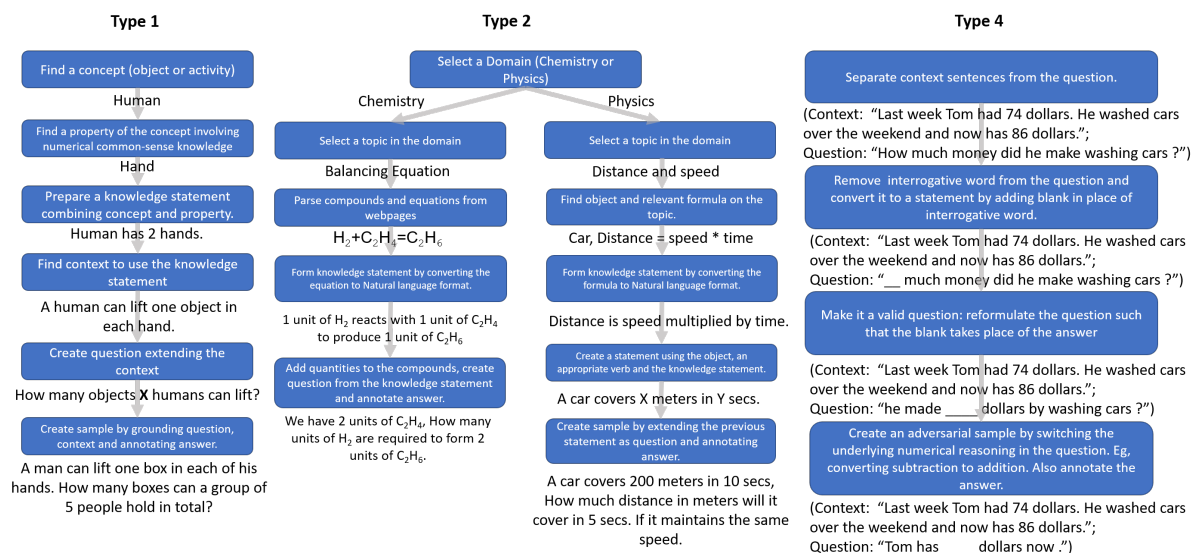

Figure 5: Step by step data creation process for task 1, 2 and 4 questions

to the question so that it has a span based answer. For other questions, we tokenize the question string into its constituent sentences and use a heuristic approach to split the question string into passage and question. Furthermore, for option based questions, we append all the options at the end of the question.

A.6 Proposed Memory-Augmented Model

Figure 13 illustrates our baseline model Ex-NumNet. We add an IR mechanism as described in Algorithm 1 and illustrated in Figure 3 of the

Type 3

> Look for entity pairs in QuaRel which are numerically comparable.

('mass', 'gravity')

> Select a question where numbers can be introduced to quantify qualitative relationship.

A golf ball has a smaller mass then a baseball. Which item has a weaker gravitational field? (A) **golf ball** (B) baseball

> Break the question in to two parts by adding numerical knowledge explicitly in place of numerically comparable quantities.

A golf ball has a mass of 78 grams and a baseball has a mass of 0.159 Kg. Which item has a weaker gravitational field? (A) **golf ball** (B) baseball

> Often, Create adversarial samples so that model can not ignore number and answer directly based on text.

A golf ball has a mass of 156 grams and a baseball has a mass of 84 gms. Which item has a weaker gravitational field? (A) golf ball (B) **baseball**

> Sometimes, add numbers in option so that model does not overfit to text data.

A golf ball has a mass of 250 grams and a baseball has a mass of 84 gms. Which item has a weaker gravitational field? (A) half a golf ball (B) **a baseball**

Figure 6: Step by step data creation process for task 3 questions

---

Algorithm 1: Our Information Retrieval Approach

```
1  Dataset D, MATH KB K Hyper-Parameters: Z, v,
     th, b v Knowledge sentences  forall s ∈ D do
2  │   Concat Question and Answer ;
3  │   Generate Query by retaining only verbs,
   │     adjectives and adverbs;
4  │   forall j ∈ K do
5  │   │   Create Index using Elastic Search ;
6  │   │   Retrieve top Z sentences from MATH KB.
7  │   end
8  │   while size(Z)>v do
9  │   │   forall k ∈ Z do
10 │   │   │   forall u ∈ k − 1 do
11 │   │   │   │   if STS(Z(u),Z(k))>th then
12 │   │   │   │   │   Delete k;
13 │   │   │   │   end
14 │   │   │   end
15 │   │   end
16 │   │   th=th-b;
17 │   end
18 end
```

main paper. As mentioned in the 'Baselines' subsection (Experiments section) of the main paper, we convert each task to RC format in our baseline and append the knowledge retrieved using IR from *MATH KB* at the end of the passage. In our experiments, we use the following hyperparameters in the IR process: $Z = 50$, $v = 10$, $th = 0.75$ and $b = 0.1$.

Formalization Let D represents dataset, s represents sample, K represent the *MATH KB*, v represents the number of knowledge statements retrieved for each sample, th is the cut off STS (Semantic Textual Similarity) value above which knowledge statements are treated redundant and removed, b is the reduction we do iteratively on th until v statements remain.

We create a knowledge base, *MATH KB* by accumulating all tasks of external knowledge which are needed to solve questions of various tasks (e.g. human has 2 hands, cow has 4 legs, there are 24 hours in a day *etc.*.). We also add math formulae required to solve questions in our benchmark (e.g. the formula of speed in terms of distance and time). We add alll these in the form of plain text separated by new line. We use Elasticsearch to retrieve relevant knowledge sentences. We further filter them using a heuristic threshold of relevance. We append this knowledge in the beginning of the passage so that continuity is not broken between passage and question. Figure 3 of the main paper illustrates our approach.

A.7 Hyper Parameters Used

All the experiments were ran with the following hyper parameters, batch size was kept at 16 where as the eval batch size was 5. The maximum number of epoch ran for the experiments were 5 with the warm-up kept at 0.06. The learning rate used was 1.5e-5 and the weight decay was 0.01.

All above hyper parameters were selected using a grid search; we kept rest of the hyper parameters unaltered. All the experiments were performed on "TeslaV100-SXM2-16GB", with which the model takes 24hrs to train on nearly 100k samples.

A.8 Additional Examples

We provide additional examples of task 1, 2, 3 and 4 questions here to better illustrate the novel datasets we have created as part of our NUMGLUE.
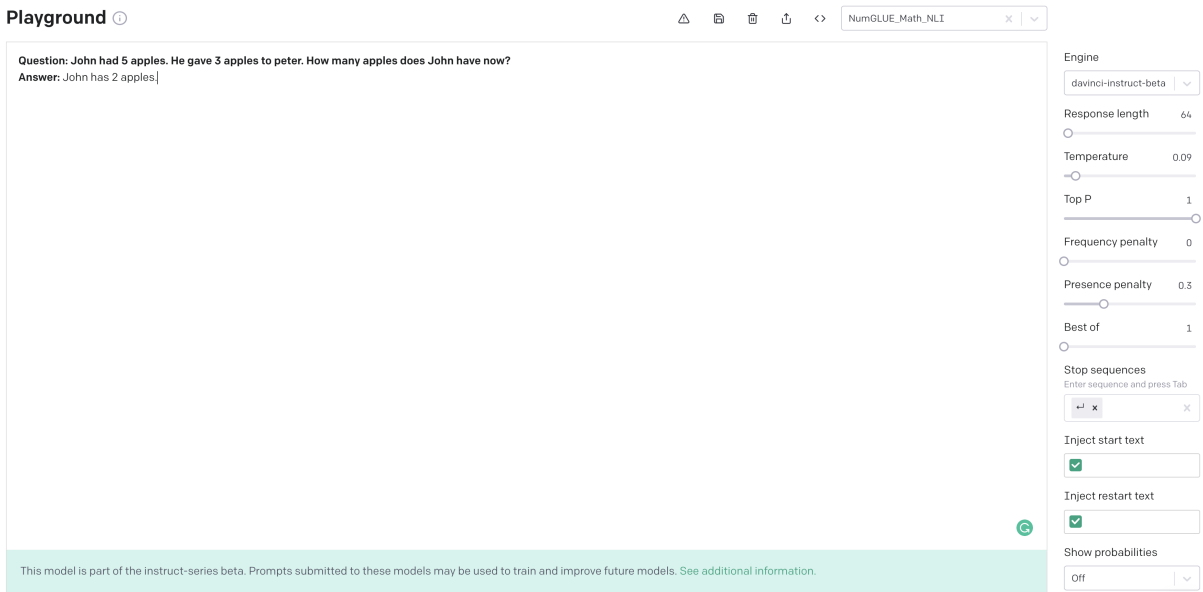
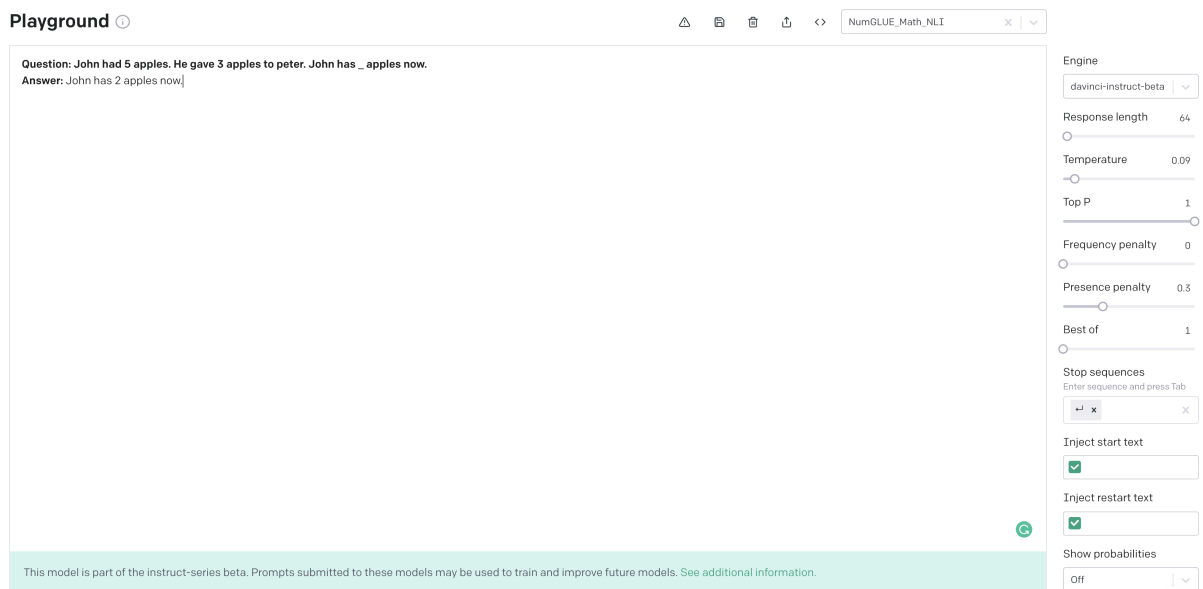Figure 7: GPT3-Instruct's response to a simple numerical reasoning question.

Figure 8: GPT3-Instruct's response to a simple numerical reasoning question expressed in fill in the blanks format.
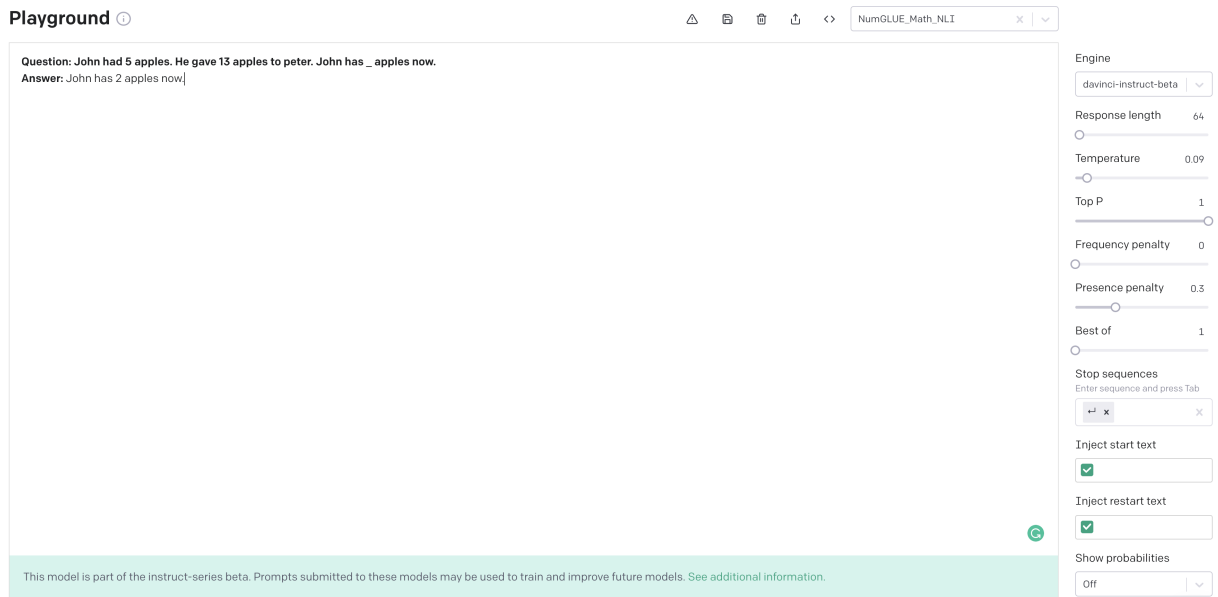
Figure 9: GPT3-Instruct's response to a simple numerical reasoning question expressed in fill in the blanks format where numbers are changed.
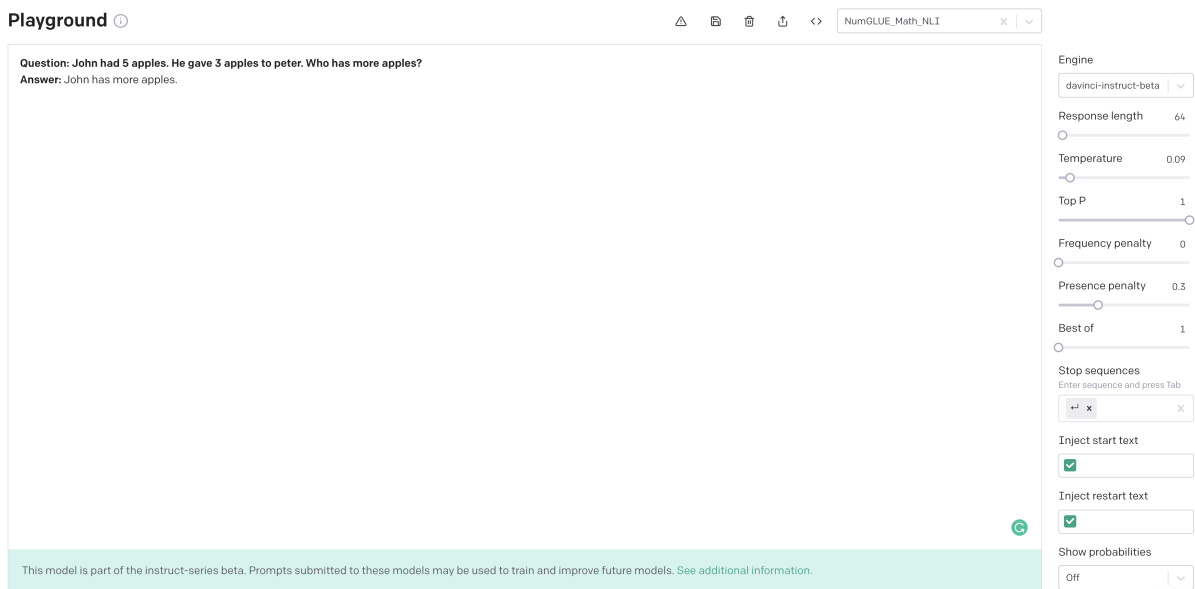

Figure 10: GPT3-Instruct's response to a simple numerical reasoning question expressed in comparison format.

Figure 11: GPT3-Instruct's response to a simple numerical reasoning question expressed in NLI format.

Question	Knowledge Required	Answer
Ella and Lily are playing a game that requires 10 die. Find out the total number of faces in 10 die.	A die has 6 faces	60
Jacob and Lillian are running a km long race. Jacob finished the race when Lillian was 190 meters from the finish line. How many meters did Lillian cover till that time?	1000 meters make a km	810
A man can lift one box in each of his hands. How many boxes can a group of 5 people hold in total?	A human being has 2 hands	10

Table 4: Example questions where numerical knowledge required to answer is not explicitly provided in the question.

Question	Knowledge Required	Answer
Find the mass percentage of H in C6H6	Mass of C is 12 units and mass of H is 1 unit	7.69
How many units of H2 are required to react with 2 units of C2H4 to form 2 units of C2H6	H2 + C2H4 = C2H6	2
A car covers 912 meters in 19 seconds. If bike's speed is one fourth of the car. Find the distance covered by the bike in 4 seconds.	distance travelled = speed * time	48

Table 5: Example questions where domain knowledge is required to answer a question.

(a) Average vocabulary represents the average number of unique words across various tasks. On an average, novel datasets (task 1-4) have higher vocabulary.

(b) Average number of unique Part of Speech (POS) tags is higher for task 1 and task 4 in the novel datasets in contrast to other tasks.

(c) STS plot for the QuaReL dataset shows significant repetition across samples

(d) STS plot for the EQUATE dataset shows considerable repetition across samples.

(e) STS plot for the DROP dataset shows repetitions for most part of the data.

(f) STS plot for the novel datasets show relatively lower repetition than other datasets

Figure 12: Data quality analysis of NUMGLUE across various tasks of data. On an average, novel datasets have higher quality than the others since they have higher average vocabulary, higher average POS tag numbers and lower Semantic Textual Similarity (STS) among each other. X-axis and Y-axis represents samples ordered in the same way, an ideal high quality dataset would have a bright line in the diagonal and rest of the places it should be dark signifying lower repetition across instances.
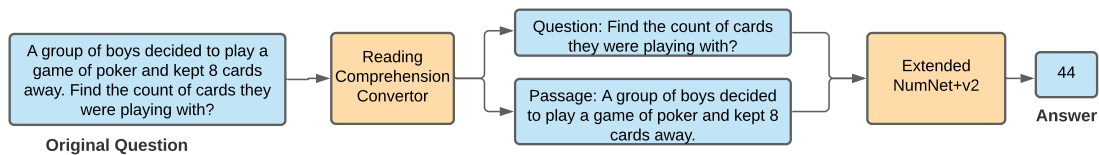
Figure 13: Architecture of Ex-NumNet

Figure 14: Conversion of various tasks to reading comprehension format

QuaRel Question	Transformed Question
A person wants to get shopping done quickly. They know that they can get through the checkout at big store faster than they can at small store. The store they go to to finish quickly is (A) **big store** (B) small store	A person wants to get shopping done quickly. They know that they can get through the checkout at big store in 5 minutes whereas it can take 20 mintues at small store. The store they go to to finish quickly is (A) **big store** (B) small store
Tina is racing her two dogs. Her greyhound is slim, her rottweiler is heavy. The dog that gets faster more quickly is the (A) rottweiler (B) **greyhound**	Tina is racing her two dogs. Her greyhound weighs 88 lbs and her rottweiler weighs 79 lbs. The dog that gets faster more quickly is the (A) **rottweiler** (B) greyhound
A golf ball has a smaller mass then a baseball. Which item has a weaker gravitational field? (A) **golf ball** (B) baseball	A golf ball has a mass of 78 grams and a baseball has a mass of 0.159 Kg. Which item has a weaker gravitational field? (A) **golf ball** (B) baseball

Table 6: Examples showing conversion of QuaRel questions to quantitative comparison questions

Arithmetic Word Problem	Transformed Question
Joan found 70 seashells on the beach. She gave Sam some of her seashells. She has 27 seashell left. How many seashells did she give to Sam ? **43**	Joan found 70 seashells on the beach . She gave Sam some of her seashells . She has 27 seashells left. She gave _____ seashells to Sam. **43**
Last week Tom had 74 dollars. He washed cars over the weekend and now has 86 dollars. How much money did he make washing cars ? **12**	Last week Tom had 74 dollars. He washed cars over the weekend and made another 86 dollars. Tom has _____ dollars now . **160**

Table 7: Examples showing MAWPS questions and corresponding questions in Completion format