Evaluating Numeracy of Language Models as a Natural Language Inference Task

Rahmad Mahendra[1], Damiano Spina[1], Lawrence Cavedon[1], Karin Verspoor[1,2]
[1]RMIT University, School of Computing Technologies
[2]The University of Melbourne, School of Computing and Information Systems
Melbourne, Australia
rahmad.mahendra@student.rmit.edu.au
{damiano.spina,lawrence.cavedon,karin.verspoor}@rmit.edu.au

Abstract

While recent advancements in large language models (LLMs) have enhanced their capabilities to solve mathematical problems, other aspects of numeracy remain underexplored. In this paper, we propose a benchmark to evaluate the ability of language models to perform basic numeracy tasks. We frame numeracy as a Natural Language Inference (NLI) task to assess the models' ability to understand both numbers and language contexts. We evaluate 49 language models (LMs), including fine-tuned LMs on NLI datasets, instruction-tuned LLMs, and specialized math-LLMs. Our findings reveal three main insights: (1) LLMs only clearly outperform smaller LMs in arithmetic tasks, indicating that mathematical reasoning cannot be generalized to other numeracy skills such as number comparison and normalization; (2) while most language models achieve fair to good accuracy for NLI entailment cases, they still struggle to predict contradiction and neutral cases; and (3) the robustness of language models' numeracy capabilities needs improvement, particularly in understanding the semantics and pragmatics of numbers in linguistic contexts.

1 Introduction

Language models (LMs) have made significant strides in the field of natural language processing (NLP) and proven to be highly adaptable to a diverse array of language-related tasks. Current benchmarks for LMs on numerical reasoning predominantly focus on answering math questions derived from test material, such as standardized assessments (Hendrycks et al., 2021; Cobbe et al., 2021; Liu et al., 2024; He et al., 2024). Recent results show remarkable progress of very large language models to decipher complex mathematics problems, such as word problems with multiple unknown variables and non-linear equations (Lu et al., 2023; Ahn et al., 2024). While solving mathematical challenges gain increasing attention in the research community, the application of numerical reasoning to real-world scenarios remains relatively unexplored. In such contexts, LLMs must not only be able to answer numerical questions but also verify the correctness of natural language text containing numerical information. It is important to expand the evaluation paradigm to include more application-oriented tasks where numerical reasoning plays a central role. For instance, fact-checking involves verifying numerical claims from textual data (Thorne and Vlachos, 2017; Venktesh et al., 2024) or scientific hypothesis testing requires numerical interpretation[1] (Lehman et al., 2019). Successful numerical reasoning in such tasks requires not only mathematical problem solving, but also a comprehension of numerical concepts (Wallace et al., 2019; Thawani et al., 2021a).

We investigate the performance of language models on foundational numeracy tasks. We test not only arithmetic reasoning but also two other essential numeracy skills, i.e., comparing numbers and understanding a number when expressed in different syntactic formats (e.g., "three" vs "3"). Our evaluation adapts natural language inference (NLI) tasks to assess the numeracy capability of language models. We analyze how well a model can use numerical information in context to make inferences or draw conclusions. Models are evaluated on their ability to disambiguate and reason about potentially conflicting information in text. NLI tasks need to deal with ambiguity, which is common when working with numerical information. Moreover, NLI tasks offer a broader evaluation of a model's overall language comprehension, including how it handles numerical information within a larger narrative. This is in line with our goal to evaluate numerical reasoning as part of a more comprehensive language understanding capability.

[1]e.g., verifying a finding "The new drug reduced symptoms by 25% more than a placebo." from a clinical trial result.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 8351–8376
April 29 - May 4, 2025

Premise	Hypothesis	Label
BASIC ARITHMETIC		
Sam had 9 dimes in his bank and his dad gave him 7 dimes.	Sam has 16 dimes now.	E
Sam had 9 dimes in his bank and his dad gave him 7 dimes.	Sam has 7 dimes now.	C
Sam had 9 dimes in his bank and his dad gave him more dimes.	Sam has 16 dimes now.	N
NUMBER COMPARISON		
Rasik walked 20 m towards north.	Rasik walked less than 80 m towards north.	E
Rasik walked not more than 20 m towards north.	Rasik walked 40 m towards north.	C
Rasik walked less than 80 m towards north.	Rasik walked less than 20 m towards north.	N
Willy has 1400 crayons. Lucy has 290 crayons.	Willy has more crayons than Lucy.	E
Willy has 1400 crayons. Lucy has 290 crayons.	Lucy has more crayons than Willy.	C
NUMBER NORMALIZATION		
John has a total of 16 shirts.	John has a total of sixteen shirts.	E
John has a total of seventy six shirts.	John has a total of 16 shirts.	C
The test of this sea urchin is up to 40 mm in diameter.	The test of this sea urchin is up to 4 cm in diameter.	E
The total weight of the hoard is 285 kilograms.	The total weight of the hoard is 285 grams.	C

Table 1: Examples of premise and hypothesis pairs from dataset covering three basic numeracy tasks: Basic Arithmetic, Number Comparison, and Number Normalization (**E:** Entailment, **C:** Contradiction, **N:** Neutral).

We introduce a benchmark consisting of over 28k sentence pairs to evaluate the numerical capabilities of language models. Our pipeline constructs diverse test sets, incorporating both existing and newly created data, by including new instances adapted from math word problems. Table 1 provides examples of this data. We conduct large-scale experiments on different pre-trained and large language models. Our findings reveal that most models achieve an F1 score below 0.75 for most Basic Arithmetic and Number Comparison tasks. Math-LLMs outperform their corresponding general-purpose models only on the Basic Arithmetic task but not on the other two tasks. On the other hand, smaller fine-tuned pre-trained language models can achieve performance comparable to large language models in both Number Comparison and Normalization tasks.

In summary, our main contributions are:

- We build a comprehensive benchmark that refines existing datasets and incorporates newly curated data for three numeracy tasks: Basic Arithmetic, Number Comparison, and Number Normalization. This benchmark serves as a valuable tool for evaluating the numerical capabilities of language models in conjunction with their language comprehension abilities.
- We conduct a large-scale empirical study on 49 open-source language models spanning 17 distinct model families using the constructed benchmark.
- Through in-depth analysis of the models' performance, we demonstrate that specialized math-based pre-training objectives do not necessarily endow models with numerical skills beyond arithmetic calculations. Furthermore, our investigation demonstrates that instruction tuning can enhance the numerical capabilities of language models.

The created dataset is released at `https://github.com/rmahendra/numeracyNLI`.

2 Related Work

Numeracy refers to the ability to understand, interpret, and use numbers (Spithourakis and Riedel, 2018; Thawani et al., 2021b). A related line of work has investigated that word embedding does not capture numeracy adequately (Naik et al., 2019; Wallace et al., 2019). Spithourakis and Riedel (2018) studied several strategies to better represent numerals in neural language models, whereas Geva et al. (2020) demonstrated that numerical reasoning skill of language models can be improved through enhanced pre-training.

Numerical reasoning ability is often investigated in the form of question answering (Xu et al., 2022; Hopkins et al., 2019), including reading comprehension (Dua et al., 2019) and math word problems (Hosseini et al., 2014; Roy and Roth, 2015; Koncel-Kedziorski et al., 2016; Huang et al., 2016; Amini et al., 2019; Miao et al., 2020; Patel et al., 2021; Cobbe et al., 2021; Hendrycks et al., 2021).

Numerical comprehension in text is also explored via different tasks, i.e., masked language modeling (Berg-Kirkpatrick and Spokoyny, 2020; Pal and Baral, 2021; Park et al., 2022; Sakamoto and Aizawa, 2023) and textual entailment (Roy et al., 2015; Ravichander et al., 2019; Akhtar et al., 2023).

EQUATE (Ravichander et al., 2019) is an NLI benchmark for quantitative reasoning, consisting of five distinct test sets, three of which involve binary classification. Although it cover various quantitative phenomena, EQUATE does not systematically evaluate specific types of numeracy. NumGLUE (Mishra et al., 2022) extends numerical reasoning evaluation to a multi-task setting, incorporating machine reading, math word problems, and NLI. While our benchmark also adopts an NLI framework, it goes beyond NumGLUE's primary focus on arithmetic tasks by introducing a broader range of numeracy challenges, such as Number Comparison and Normalization.

More recently, FERMAT (Sivakumar and Moosavi, 2023) provided a fine-grained analysis of arithmetic skills by examining number representation, mathematical operations, and training dependencies. In contrast to FERMAT, which evaluates arithmetic reasoning through QA tasks, our work leverages an NLI framework for a more structured assessment of logical reasoning over numerical facts. Additionally, while FERMAT employs template-based data augmentation to evaluate accuracy across different representations, our approach extends this by incorporating formula-based augmentations to analyze robustness across language contexts and numerical magnitudes. Furthermore, our benchmark offers a more comprehensive evaluation of Number Normalization.

3 Basic Numeracy Framework

Evaluation of numeracy and language understanding in our framework is formulated as NLI tasks (Bowman et al., 2015). Given a pair of sentences, premise and hypothesis, the task solver is asked to determine the semantic relation between them; i.e., *entailment*, where the hypothesis is true given the premise, *contradiction*, where it is false, or *neutral*, where the truth of the hypothesis cannot be determined. The data instance in this evaluation benchmark follows the *minimal pairs* principle (Warstadt et al., 2020), in which each premise sentence has minimally changed hypothesis sentences with different label.

3.1 Tasks

Basic numeracy capabilities are examined through three tasks, i.e., Basic Arithmetic, Number Comparison, and Number Normalization (Table 1). These three tasks are packaged as word problems that require numeracy which is a combination of the ability to work with numbers and understand language context.

1. Arithmetic Word Problem. Basic Arithmetic task refers to a mathematical operation that involves fundamental calculations with numbers. This task typically include basic arithmetic operations such as addition, subtraction, multiplication, and division. We do not cover complex mathematical concepts or advanced operations. Since our objective is to test basic numeracy, we constrain the arithmetic task as single-hop reasoning problem, in which all numbers needed for the calculation are provided literally in the context.

2. Number Comparison. Number comparison task involves the process of comparing two or more numerical values to determine their relative relationships. In such task, we evaluate whether one number is greater than, less than, or equal to another number (e.g., $10 > 5$, $1 \leq 2$). We cover two subtasks: number with quantifier and comparative reasoning.

In the first subtask, a premise is a sentence containing numerical information and a hypothesis is a similar sentence to the premise, except for the numerical value. The original number is substituted with different number and concatenated with a quantifier. For several instances, we use the same number in both premise and hypothesis and the semantics is changed with a quantifier. In another subtask, a premise sentence comprises two or more numerical information. A hypothesis draws a hypothetical conclusion that requires comparative reasoning between the numbers in the premise.

3. Number Normalization. A Number Normalization task is concerned with the process of converting numerical values into a standardized or normalized format. Two types of normalization are covered in this task: numeration and measurement normalization. Numeration tests the understanding of different representation systems for a number. For example, a number in English can be written as numerical digits ('9') or text ('nine'). Measurement normalization is about recognizing whether different strings refer to same or different unit of

measurement (e.g., 'm' = 'meter', 'm' ≠ 'mm') and converting a numerical value when changing unit of measurement (e.g., 1 'm' = 100 'cm').

3.2 Data Construction

To construct a benchmark for basic numeracy framework, we collect data from different sources. We select existing NLI data with a quality assurance process. We also transform the data from other NLP tasks into NLI and auto-generate new instances by leveraging template or formula change.

3.2.1 Existing NLI Data

We utilize AWP-NLI from EQUATE (Ravichander et al., 2019) for arithmetic task. We manually reviewed data and eliminated incorrect instances (incomplete hypothesis sentence or wrong label). We initially examined the numerical stress test set from Naik et al. (2018) (also part of EQUATE) for the comparison task. However, we decided to discard this data in our benchmark for two reasons: (i) Label have spurious patterns (Gururangan et al., 2018), as *neutral* instances are constructed only from *entailment* instances by swapping premise and hypothesis sentences; (ii) Many premise sentences contain multiple numbers. Since only one number is modified with quantifier, the hypothesis sentences may possess ambiguous interpretation in most cases.

3.2.2 Data Creation for Arithmetic Task

Several works (Demszky et al., 2018; Yin et al., 2021) harness other tasks to create NLI datasets. In this work, we recast the instances of arithmetic word problem (AWP) dataset that are originally for question answering task into NLI task. The AWP dataset consists of MultiArith, the Common-core multi-step arithmetic problem (Roy and Roth, 2015), arithmetic subset of ASDiv (Miao et al., 2020), and SVAMP (Patel et al., 2021).

For each instance of a word problem, we take the context of the question as premise. We concatenate the question with the correct answer and then transform it into a declarative sentence using ChatGPT [2]. We inspect the generated sentence to ensure the correctness. We use the declarative sentence as hypothesis for an *entailment* instance. We perform coreference resolution to resolve the pronoun in the hypothesis to a named-entity or noun phrase in the premise referred to. To generate *contradiction* instance, we replace the correct number in the hypothesis with a random guess number. Unlike AWP-NLI (Ravichander et al., 2019), replacement number is not constrained within specific interval, but can be any positive number. As there is overlap between the recast AWP data with existing NLI data[3], we deduplicate them in our final data.

The *neutral* case occurs when arithmetic computation cannot be performed completely in order to determine the correctness of the numerical conclusions on the hypothesis. We utilize the UMWP subset (Sun et al., 2024) to create *neutral* NLI instances. The recasting process is similar to *entailment*, except for question and answer concatenation. UMWP is derived from MultiArith, AsDiv, and SVAMP. A UMWP instance corresponds to an instance of those three AWP dataset. We concatenate the correct and incorrect answers from corresponding original AWP with the UMWP question, then transform them into hypothesis sentences.

3.2.3 Data Creation for Number Comparison with Quantifier Task

We pick the sentences containing numerical information as the premise. The sentences are sampled from AWP dataset and Wiki-Convert (Thawani et al., 2021a). The hypothesis is constructed by modifying the numerical quantity from the premise sentence. The modification is done by changing the number and/or inserting the quantifier ("more than" or "less than") preceding the number. Those strategies produce NLI instances with *entailment* or *contradiction* label. To create *neutral* instance, we flip the premise and hypothesis in *entailment* pair.

To improve the data variation, we generate more data by using the premise sentence from *neutral*. The hypothesis is slightly modified premise by changing the number and/or removing the quantifiers. In addition, we prefix some quantifiers with negation ("not"), either in the premise or hypothesis sentence, or both. Refer to Appendix A for details of the data variation.

3.2.4 Data Creation for Comparative Reasoning Subtask

A number of AsDiv and SVAMP instances are comparison word problems. Original question typically asks number difference between object being compared (usually using subtraction operation). In this task, instead of magnitude difference, we focus

[2] https://chatgpt.com/

[3] Several AWP-NLI instances were obtained from the same source with MultiArith

on correct relationship between compared objects. To create *contradiction* instance, we modify the hypothesis sentence by either switching syntactic position of compared entities or changing word comparator with its antonym. On the other hand, applying both modification strategies generate another *entailment* instance.

3.2.5 Data Creation for Number Normalization Tasks

To create the data for Number Normalization task, we sample the sentences from AWP dataset and WikiConvert. We make use of number-numeral dictionary (Gorman and Sproat, 2016) and table of unit conversion for numeration and measurement normalization subtasks, respectively. For numeration subtask, we cover numbers less than 100 and multiple of hundreds / thousands (200, ..., 3000, etc.)[4]. We exclude sentences where numbers are bound by quantifiers or used in estimation contexts, as identifying semantic relations becomes non-trivial when these numbers are replaced. Changing the numbers in such cases may not always contradict the original sentence.

The *entailment* instance consists of the sentence containing number as premise and similar sentence but substituting the number with its numeral synonym as hypothesis. To construct *contradiction* instances, we can apply different strategies. We pick a random guess number different to the number in the premise and use numeral representation for hypothesis sentence. Otherwise, we insert / remove part of the numeral text so it refers to a different number to that in the premise (e.g., use twenty ~~two~~ in the hypothesis and 22 in the premise).

For the measurement subtasks, we consider three common families of units: length (e.g., meter, feet), weight (e.g., kilogram), and area (e.g., square meter). We utilize the sentence containing number with measurement unit as premise and generate hypothesis by: (i) changing the unit with its abbreviation or writing variation (e.g., square kilometer $\rightarrow$ km2), or (ii) converting number and unit.[5] The *contradiction* case is introduced by perturbing the number and/or measurement unit. As normalization is symmetric, we can expand the data with preserving label (both *entailment* and *contradiction*) by flipping the premise-hypothesis sentences.

[4]Numbers greater than 100, e.g., 234 and 9876 are rarely verbalized as numeral in the sentence

[5]As our scope is basic numeracy, unit conversion is limited among SI unit (e.g., km to m, vice versa, but not km to mile).

3.3 Constructing Data for Further Analysis

In addition to the data prepared following the steps in section 3.2, we generate additional data to further analyze the relationship between numeracy and literacy. We expand NLI instances by varying template formulas and numerical representations.

3.3.1 Template-Formula Data Augmentation

Numeracy in natural language text is not only applying number operations but also understanding the context in which numbers are presented within the text. Identical number tuples appearing in different context may leads to dissimilar numerical interpretation and require different arithmetic operation to draw a conclusion.

To evaluate the language model's ability to understand context and interpret numbers, we prepare augmented data for the arithmetic task by combining templates and formulas. We randomly select 500 instances of the Basic Arithmetic task (100 for each arithmetic operation and another 100 neutral instances). For the non-neutral instances, we consider only those requiring a single operation with exactly two numbers in the premise (n_1, n_2), one number is in the hypothesis (n_3), and equation $n_1\ op\ n_2 = n_3$ applies, $op \in \{+, -, \times, \div\}$. We replace all numbers in both premise and hypothesis with placeholders, which we then used as *template*.

On the other hand, we auto-generate 500 equations, i.e., 4×100 correct equations for each arithmetic operation and another 100 false equations no matter chosen operator (e.g., $12\ op\ 3 = 5$). We pair each template with 10 different formula equations while each formula with 10 different templates. Hence, we collect 5000 instances in total. When template obtained from instances with addition operation is paired with addition formula, it produces *entailment* data. While, this template is assigned to formula with different operation (or false equation), it produces *contradiction* data. Templates that are originally from *neutral* instances always produce *neutral* instances when paired with any formula. *Template* and *formula* examples are presented in Table 3, and an illustration of the template-formula data augmentation methodology in Figure A1.

3.3.2 Numerical Representation Variation

To test the capability of the model perform Number Normalization in other tasks, we prepare data for Basic Arithmetic and Number Comparison task with different representation of number. We select 500, 500, and 250 instances with balanced label

	E	C	N
Basic Arithmetic	2480	2418	971
Number Comparison			
Number with Quantifier	2944	3724	2452
Reasoning	488	488	–
Number Normalization			
Numeration	494	492	–
Measurement Unit	646	646	–

	E	C	N
Template-formula based data generation			
Basic Arithmetic	2092	1908	1000
Number representation variation			
Basic Arithmetic	1200	1200	600
Number with Quantifier	558	564	378
Comparative Reasoning	378	372	–

Table 2: Basic numeracy data statistics.

(I_1)	P: A pet store had 13 siamese cats and 5 house cats. During a sale they sold 10 cats. H: The pet store have 8 cats left.	E
(I_2)	P: A pet store had 43 siamese cats and 32 house cats. During a sale they sold 22 cats. H: The pet store have 53 cats left.	E
(I_3)	P: While on vacation, Rachel took 13 pictures at the zoo and 5 at the museum. She later deleted 10 of the pictures. H: Rachel still has 8 pictures from her vacation.	E
(I_4)	P: The school cafeteria ordered 43 red apples and 32 green apples for students lunches. Only 22 students wanted some fruits. H: The cafeteria ended up with 53 extra fruits.	N
	TEMPLATE	
$\mathcal{T}_1$	P: A pet store had $\cdots$ siamese cats and $\cdots$ house cats. During a sale they sold $\cdots$ cats. H: The pet store have $\cdots$ cats left.	
	FORMULA	
$\mathcal{F}_1$	13 + 5 - 10 = 8	
$\mathcal{F}_1$	43 + 32 - 22 = 53	

Table 3: Template and formula in arithmetic task. Instances I_1 and I_2 use the same template ($\mathcal{T}_1$). Instances I_3 and I_4 share the same formula ($\mathcal{F}_1$), while I_2 and I_4 have $\mathcal{F}_2$

distribution for arithmetic, number with quantifier, and comparative reasoning, respectively. For Basic Arithmetic tasks, we generate six different subsets. Premise sentences can contain either number in digit, numeral, or mixed (one number is in digit and another one is numeral), while hypothesis sentences contain either digit or numeral form. For Number Comparison tasks, we create three subsets based on whether the numbers being compared are in the same or different formats (digit or numeral).

3.4 The Resulting Dataset

After filtering the EQUATE dataset and incorporating it with recast AWP and generated NLI data, we end up with 18,243 premise-hypothesis pairs for main dataset. We generate an additional 5,000 and 5,250 instances using template-formula augmentation and various number representation, respectively. Unless otherwise stated, the main dataset is used for experiments. The statistics of the resulting dataset for each taskare presented in Table 2.

4 Experiments and Results

4.1 Models

We conduct experiments on 49 language models spanning 17 distinct model families. Our evaluation includes pre-trained language models such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), BART (Lewis et al., 2020), XLNet (Yang et al., 2019), and DeBERTa (He et al., 2023, 2021). Those models have been fine-tuned on NLI datasets, i.e., SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER-NLI (Nie et al., 2019), and ANLI (Nie et al., 2020a).

Large language models (LLMs) such as Flan-T5 (Chung et al., 2024), Mistral (Jiang et al., 2023), Llama (Touvron et al., 2023; Dubey et al., 2024; Meta Llama Team, 2024), Gemma (Gemma Team et al., 2024a,b), Qwen (Yang et al., 2024a; Qwen Team, 2024), InternLM (Cai et al., 2024), and DeepSeek LLM (DeepSeek-AI et al., 2024) are also included in our experiments. We evaluate both

base and instruction-tuned LLMs with varying size ranging from 1.5B to 12B.

Additionally, we investigate math LLMs, specifically designed for mathematical reasoning. These models were either pre-trained with a specialized training objective or fine-tuned on math datasets. Our experiments with math LLMs include Llemma (Azerbayev et al., 2024), MetaMath (Yu et al., 2023), WizardMath (Luo et al., 2025), OpenMath (Toshniwal et al., 2024), QwenMath (Yang et al., 2024b), InternLM-Math (Ying et al., 2024), and DeepSeekMath (Shao et al., 2024).

We perform evaluation of general and math LLMs in a zero-shot setting. We test three different NLI prompts on sample data of Basic Arithmetic task using Llama-2 model, and later run the best prompt when evaluating whole data in our benchmark. The prompts are presented in Table A5.

4.2 Results

Overall Evaluation. Table 4 presents the F1 scores of the top-performing models from each model family across the five basic numeracy subtasks. For a more detailed results of precision and recall scores, refer to Tables A10 and A11. We observe that most large language models (both general-purpose and math-LLMs) outperform smaller pre-trained fine-tuned language models only on Basic Arithmetic task. Flan-T5-XXL and Qwen2.5-Math-7B-instruct are the two LLMs that consistently match performance of NLI fine-tuned models on the other four subtasks. However, even these models fall short of DeBERTa-v3's performance on all four subtasks.

Our experiments do not reveal a clear superiority of Math LLMs over other models in basic numeracy tasks. MetaMath-Mistral-7B demonstrated a marginal improvement over its corresponding general-purpose model, Mistral-7B-instruct-v0.3. In contrast, InternLM2-Math-plus-7B and DeepSeek-Math-7B-rl performed less effectively than their respective counterparts, InternLM2-7B and DeepSeek-LLM-7B-chat.

Our findings indicate that models with larger parameter sizes generally exhibit superior performance. Instruction tuning has a significant impact on enhancing the numerical abilities of LLMs, leading to a 10-20 point increase in accuracy compared to their base versions. In certain model-task combinations, this improvement can even surpass 30 points (e.g., Gemma-2-9B-it achieved an F1 score of 0.634 in Basic Arithmetic tasks, while its base version is only 0.320). Moreover, the rapid advancement of newer LLMs has also contributed to improvements in numerical capabilities. By comparing Llama, Gemma, and Qwen models from different release versions, we saw that later versions consistently outperformed their predecessors (e.g., Llama-3.1 vs. Llama-2). Complete evaluation results for all models are shown in Table A7.

Basic Arithmetic Evaluation. Top-performing LLMs successfully predict most entailment cases. Stronger math-focused LLMs, such as InternLM2-Math and DeepSeekMath, achieve a 0.9 recall for the entailment class. However, these models struggle to correctly classify non-entailment instances.

Humans generally find addition and subtraction easier to learn than multiplication and division, and face more challenges when solving arithmetic formulas with multiple operations. Our investigation, shown in Table A9, reveals the same patterns, for smaller fine-tuned models, like DeBERTa-v3. For math-LLMs, the pattern becomes less human-like, as their ability to perform complex arithmetic operations matches their proficiency with more basic ones (addition, subtraction).

Number Comparison Evaluation. Most models achieve fair to good precision for the entailment class on the number with quantifier subtask. However, they suffer from low recall for the neutral class, with scores less than 0.3 for most fine-tuned and instruction-tuned models, and even worse for Math-LLMs. Notably, OpenMath-Mistral and InternLM2-Math-plus fail to retrieve any neutral instances at all.

Number Normalization Evaluation. The results in Table 4 demonstrate that recent language models can perform Number Normalization very well. Pre-trained fine-tuned smaller LMs and instruction-tuned general-purpose LLMs, beside Mistral-7B-instruct-v0.3, achieve F1-score of 0.9 and 0.8 for numeration and measurement unit normalization subtasks, respectively. Surprisingly, half of math-LLMs score lower performance, behind LMs that are not specifically trained or tuned on mathematics reasoning purpose.

Furthermore, we are interested in exploring whether language models' capabilities in Number Normalization extend to other numeracy tasks. We compare model performance across various number representations in arithmetic and comparison tasks. Complete results are presented in Figure A3.

	Arithmetic	Comparison		Normalization	
		Quantifier	Reasoning	Numeration	Unit
Fine-tuned PLMs					
RoBERTa + NLI	0.372 / 0.362	0.580 / 0.489	0.684 / 0.697	0.939 / 0.941	0.821 / 0.821
ALBERT + NLI	0.384 / 0.381	0.672 / 0.629	0.679 / 0.669	0.963 / 0.965	**0.906 / 0.905**
ELECTRA + NLI	0.375 / 0.360	0.575 / 0.475	0.578 / 0.611	0.917 / 0.919	0.808 / 0.808
BART + NLI	0.340 / 0.333	0.596 / 0.532	0.570 / 0.588	0.931 / 0.935	0.887 / 0.887
XLNet + NLI	0.371 / 0.350	0.585 / 0.501	0.634 / 0.640	0.915 / 0.918	0.827 / 0.826
DeBERTa-v3 + NLI	**0.451 / 0.434**	**0.689 / 0.654**	**0.795 / 0.798**	**0.973 / 0.978**	0.889 / 0.892
General LLMs					
Flan-T5-XXL (**11B**)	0.440 / 0.433	**0.667 / 0.615**	**0.801 / 0.808**	0.933 / 0.936	0.848 / 0.845
Mistral-**7B**-instruct-v0.3	0.408 / 0.403	0.457 / 0.441	0.501 / 0.535	0.736 / 0.818	0.665 / 0.687
Llama-3.1-**8B**-instruct	0.577 / 0.434	0.536 / 0.397	0.701 / 0.690	0.927 / 0.927	0.781 / 0.771
Gemma-2-**9B**-it	**0.651 / 0.634**	0.632 / 0.561	0.703 / 0.748	0.885 / 0.938	0.778 / 0.811
InternLM2-**7B**	0.494 / 0.423	0.542 / 0.424	0.655 / 0.652	0.945 / 0.949	**0.876 / 0.875**
DeepSeek-LLM-**7B**-chat	0.481 / 0.402	0.472 / 0.340	0.561 / 0.562	**0.962 / 0.963**	0.783 / 0.775
Qwen2.5-**7B**-instruct	0.555 / 0.509	0.576 / 0.480	0.722 / 0.729	0.930 / 0.952	0.831 / 0.851
Math LLMs					
Llemma-**7B**	0.420 / 0.270	0.404 / 0.333	0.487 / 0.391	0.495 / 0.443	0.513 / 0.544
MetaMath-Mistral-**7B**	0.465 / 0.354	0.431 / 0.365	0.526 / 0.474	0.757 / 0.775	0.632 / 0.617
WizardMath-**7B**-v1.1	0.488 / 0.384	0.423 / 0.384	0.529 / 0.472	0.645 / 0.700	0.601 / 0.578
OpenMath-Mistral-**7B**	0.515 / 0.374	0.496 / 0.369	0.593 / 0.579	0.948 / 0.948	0.676 / 0.658
InternLM2-Math-plus-**7B**	0.498 / 0.335	0.495 / 0.366	0.502 / 0.371	0.754 / 0.737	0.689 / 0.657
DeepSeek-Math-**7B**-rl	0.567 / 0.442	0.381 / 0.284	0.535 / 0.451	0.922 / 0.923	0.797 / 0.791
Qwen2.5-Math-**7B**-instruct	**0.619 / 0.451**	**0.631 / 0.481**	**0.729 / 0.727**	**0.965 / 0.965**	**0.851 / 0.850**

Table 4: Language models evaluation on Basic Arithmetic, Number Comparison, and Number Normalization tasks. Evaluation results are presented as "accuracy / macro F1" scores. Full evaluation results for LLMs in varied size and released version are presented in Tables A7 and A8 in Appendix C.

While performance varies when switching from digits to words, there is no significant difference in scores. This suggests that language models are not only able to recognize numbers in different formats (i.e., as an isolated NLI task) but also apply this ability effectively in tasks requiring numeracy.

4.3 Analysis

Numeracy in word problems requires understanding not only numbers but also the language context. Solving these problems involves interpreting the numbers according to the textual context, identifying relevant information, and discerning the operations needed to find a solution. We analyze the interplay of numeracy (number understanding) and language comprehension. We discuss robustness and magnitude effect of LMs numeracy capability.

How robustly do LMs perform basic numeracy using different number and language context? Our benchmark evaluates model robustness by considering both accuracy and invariance with respect to templates or formulas. Template accuracy is calculated as the mean accuracy across all instances derived from that template. Template invariance measures the consistency of a model's predictions for instances generated from the same template. Formally, let an instance of an arithmetic task be $I_i =< x_i, y_i, \mathcal{T}_a, \mathcal{F}_b >$, where x_i is the premise-hypothesis text whose gold label is y_i. This instance utilizes template $\mathcal{T}_a$ and embeds formula $\mathcal{F}_b$. The accuracy and invariance for template $\mathcal{T}_j$ are computed using the following formula.

$$Acc(\mathcal{T}_j, \mathcal{M}) = \frac{1}{N} \sum_k f_{\mathcal{M}}(x_k) = y_k \ ,$$

$$Inv(\mathcal{T}_j, \mathcal{M}) = \frac{1}{N} \max_c \sum_k f_{\mathcal{M}}(x_k) = \mathcal{C}_c \ ,$$

$$I_k =< \_, \_, \mathcal{T}_j, \_ >$$

where $f_{\mathcal{M}}(x_i)$ denotes the label predicted by a model $\mathcal{M}$ for x_i, $\mathcal{C} = \{$*entailment*, *contradiction*, *neutral*$\}$, and $N = 10$ in our experiment.

Figure 1 illustrates the LMs' evaluation on arithmetic dataset clustered by the template or formula used. The LM accuracy for the formula group is comparable to that for the template group, though the latter shows more variability. This suggests that certain linguistic structures in which numbers are

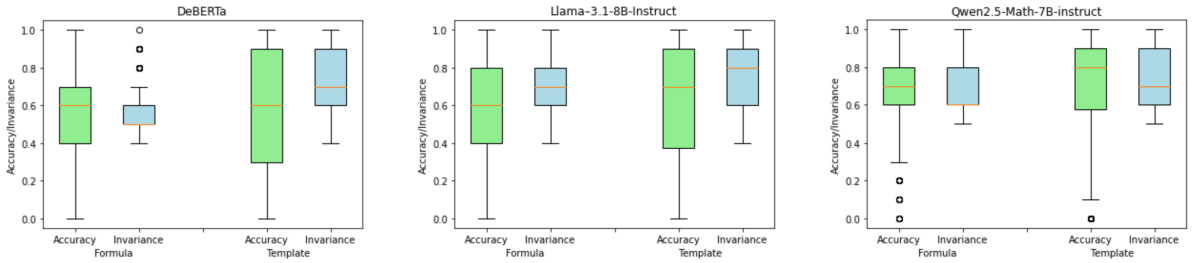

Figure 1: Basic arithmetic robustness using different template and formula for DeBERTa-v3, LLama3.1-8B-instruct, and Qwen2.5-Math-7B-instruct. Robustness evaluation for other models is reported in Figure A2.

embedded are more difficult or easier for LMs to learn. On the other hand, we find that most models exhibit low invariance when learning formulas. They generally do not output the same prediction when assigning the same number combinations to different templates, indicating that the models attempt to learn from the context. Our investigation reveals that templates with an invariance score near 1 are mostly for neutral instances. However, in most cases, predictions with an invariance score of 1 have an accuracy score of 0, indicating that the model repeatedly outputs the wrong label.

How reliably do LMs select relevant information to perform basic numeracy? Model robustness can also be evaluated by the extent to which they filter and use correct information from the context to perform numeracy. LLMs perform poorly on the neutral class, indicating that models struggle to identify whether computation cannot be verified given insufficient information context in the text. This finding aligns with Sun et al. (2024), where the models hallucinate with unanswerable math problems. For the *entailment* case, models become more confused when the premise sentences contain confounding numbers, i.e., not all numbers in the text are relevant for arithmetic calculation (see the example in Table A6).

Does number magnitude affect numeracy of LMs? Figure 2 shows model performance based on number magnitude. All models show a consistent decrease in F1-score as the numbers used in the formulas increase, with the only outlier being DeBERTa-v3 when working with large numbers. This suggests models perform calculations with greater accuracy on smaller numbers (i.e., numbers less than 100) compared to numbers in the thousands and tens of thousands. This finding aligns with previous work by Shah et al. (2023), which demonstrated a correlation between model numeracy and number magnitude in number comparison

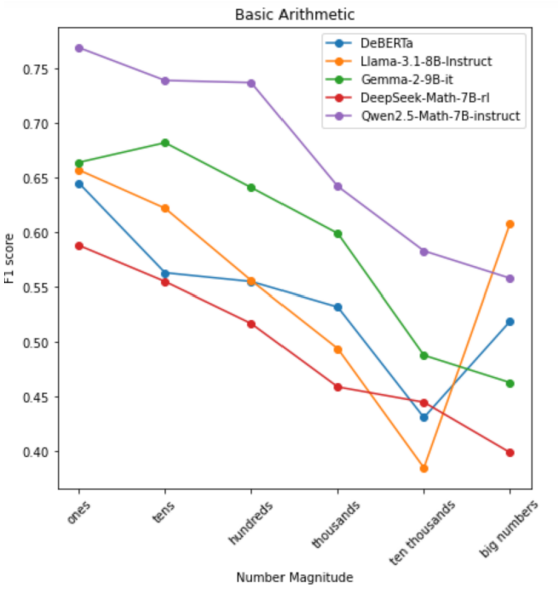

Figure 2: Number magnitude effect on numeracy.

tasks. These results indicate that LLMs may mimic aspects of human cognitive processing. Behavioral studies of arithmetic have shown that increasing the size of the numbers makes problem-solving more difficult for humans (Hamann and Ashcraft, 1985; Kong et al., 1999; Zbrodoff and Logan, 2005).

5 Conclusion

We have introduced a new benchmark dataset for evaluating basic numeracy, including Basic Arithmetic, Number Comparison, and Normalization, in NLI tasks. This provides a resource supporting evaluating progress on numeracy skills of NLP systems. Our experiments reveal most large language models struggle to correctly classify non-entailment cases. Our analysis indicates that the numeracy capabilities of models align with human behavior concerning arithmetic complexity and number magnitude. Model robustness is also an issue, particularly in relation to understanding numbers correctly in linguistic contexts.

Ethical Considerations

The data utilized in this study does not contain any personally identifiable information or offensive content. All data is in English and originates from multiple sources, each distributed under different licenses. Multi-Arith (Roy and Roth, 2015), ASDiv (Miao et al., 2020), and SVAMP (Patel et al., 2021) are sourced from crawled web pages, while Wiki-Convert (Thawani et al., 2021a) is derived from Wikipedia. ASDiv is released under the CC BY-NC 4.0 license, while EQUATE (Ravichander et al., 2019) and SVAMP are distributed under the MIT license. UMWP (Sun et al., 2024) is licensed under the CC BY-SA 4.0 protocol. The data is intended solely for academic research and non-commercial purposes.

Limitations

Our work examines a limited scope of numeracy, focusing specifically on three key types: Basic Arithmetic, Number Comparison, and Number Normalization. However, numeracy encompasses a broader range of skills that we leave for future exploration, such as counting, estimation, and understanding numerical concepts across domains. These more complex aspects of numeracy present unique challenges in real-world contexts, and their evaluation remains an open area of research. Furthermore, while numerical information can be represented in various ways within text (Mahendra et al., 2024), this study primarily uses cardinal and measurement numbers. Fractions, decimals, and percentages are underrepresented in our corpus. We leave exploring these areas to future research.

On the other hand, certain NLI instances inherently present ambiguities. Human judgments can vary when evaluating the same reasoning task (Nie et al., 2020b; Weber-Genzel et al., 2024). For example, some individuals may interpret the statement "Charles traveled for 2 hours" as contradictory to "Charles traveled for three hours" citing the discrepancy in numerical information, specifically *travel time*. However, others may perceive a monotonic reasoning (Yanaka et al., 2019), i.e., partial relationship between the two statements reasoning that "if Charles traveled for 3 hours, it implies he also traveled for 2". In this study, we adopt the former interpretation when constructing data for Number Normalization. A more comprehensive exploration of these nuances will be the subject of future research.

The current work evaluates only the zero-shot capabilities of language models, without exploring more sophisticated approaches such as few-shot learning (Brown et al., 2020) or chain-of-thought prompting (Kojima et al., 2022; Wei et al., 2022), which are left for future investigation. Furthermore, this work does not prioritize prompt engineering. Our primary objective was not to optimize task performance through prompt design but to focus on numeracy-related reasoning evaluation. We employed prompts from existing literature without fine-tuning them specifically for our experiments.

Finally, our study is limited to the English language. Although numeracy is a universally applicable skill, resources for building multilingual benchmarks remain sparse. There are only a few datasets available for math word problems or numerical reasoning corpora in languages other than English, constraining our capacity to extend this work to a broader language set.

Acknowledgments

This research was undertaken using the GPU server hosted by School of Computing Technologies at RMIT University and LIEF HPCGPGPU Facility hosted at the University of Melbourne (LIEF Grant LE170100200). This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Government.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math

word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*.

Taylor Berg-Kirkpatrick and Daniel Spokoyny. 2020. An empirical investigation of contextualized number prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4754–4764, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 technical report. *Preprint*, arXiv:2403.17297.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan

Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *Preprint*, arXiv:1809.02922.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline C. Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich,

Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 herd of models. *ArXiv*, abs/2407.21783.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024a. Gemma: Open models based on Gemini research and technology. *ArXiv*, abs/2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry,

Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, L. Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, S. Mc Carthy, Sarah Perrin, Sébastien Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Brian Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Mary Sue Hamann and Mark H Ashcraft. 1985. Simple and complex mental addition across development. *Journal of Experimental Child Psychology*, 40(1):49–72.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced BERT with disentangled attention. 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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. 2019. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 887–896, Berlin, Germany. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

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, San Diego, California. Association for Computational Linguistics.

Jian Kong, Yuping Wang, Hongyan Shang, Ying Wang, Xiuzhen Yang, and Ding Zhuang. 1999. Brain potentials during mental arithmetic-effects of problem difficulty on event-related brain potentials. *Neuroscience Letters*, 260(3):169–172.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6884–6915, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*.

Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and Karin Verspoor. 2024. Do numbers matter? types and prevalence of numbers in clinical texts. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 409–415, Bangkok, Thailand. Association for Computational Linguistics.

Meta Llama Team. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. `https://ai.meta.com/blog/meta-llama-3/`.

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, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Kuntal Kumar Pal and Chitta Baral. 2021. Investigating numeracy learning ability of a text-to-text transfer model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3095–3101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sungjin Park, Seungwoo Ryu, and Edward Choi. 2022. Do language models understand measurements? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1782–1792, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Qwen Team. 2024. Hello Qwen2. `https://qwenlm.github.io/blog/qwen2/`.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Taku Sakamoto and Akiko Aizawa. 2023. Predicting numerals in text using nearest neighbor language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4795–4809, Toronto, Canada. Association for Computational Linguistics.

Raj Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma. 2023. Numeric magnitude comparison effects in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6147–6161, Toronto, Canada. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Jasivan Sivakumar and Nafise Sadat Moosavi. 2023. FERMAT: An alternative to accuracy for numerical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15026–15043, Toronto, Canada. Association for Computational Linguistics.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

YuHong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. 2024. Benchmarking hallucination in large language models based on unanswerable math word problem. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2178–2188, Torino, Italia. ELRA and ICCL.

Avijit Thawani, Jay Pujara, and Filip Ilievski. 2021a. Numeracy enhances the literacy of language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6960–6967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021b. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2017. An extensible framework for verification of numerical claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain. Association for Computational Linguistics.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. OpenMathInstruct-1: A 1.8 million math instruction tuning dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894, Online only. Association for Computational Linguistics.

V. Venktesh, Anand Abhijit, Anand Avishek, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 650–660, New York, NY, USA. Association for Computing Machinery.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jianhan Xu, Cenyuan Zhang, Xiaoqing Zheng, Linyang Li, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2022. Towards adversarially robust text classifiers by learning to reweight clean examples. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1694–1707, Dublin, Ireland. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *Preprint*, arXiv:2402.06332.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. MetaMath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

N Jane Zbrodoff and Gordon D Logan. 2005. What everyone finds: The problem-size effect. In *The handbook of mathematical cognition*, pages 331–345. Psychology Press.

A Appendix: Data

Comparative Reasoning. Table A2 presents example of data for comparative reasoning that is reformulated from math word problem dataset.

Number Comparison with Quantifier. We summarize the type of modification and label produced for Number Comparison with quantifier task in Table A1. The data is augmented by inserting quantifier ("more" or "less") and negation, and swapping premise and hypothesis text.

Number Representation Variation. Table A3 presents examples of data for Basic Arithmetic and Number Comparison tasks using different number representations.

B Appendix: Models

Table A4 lists all models evaluated in our experiment. We utilize model implementation from Huggingface (Wolf et al., 2020). Table A5 presents prompt for evaluating LLMs in zero-shot setting.

C Appendix: Experimental Results

Table A7 and A8 shows full evaluation results of models. More detailed evaluation on Basic Arithmetic task are reported in Tables A9 and A10; Number Comparison in Table A11. The models exhibit high proficiency when comparing only numbers. The performance drops when number comparisons involve quantifier and negation (i.e., when premise or hypothesis is in form of $neg+quant+n_j$). This result confirms teh study by Truong et al. (2022) that LMs still lack comprehension of negation.

Figure A3 presents evaluation results of different variation of number representation on Basic Arithmetic and Number Comparison tasks.

Premise	Hypothesis	Possible Labels
n_i Rasik walked 20 m towards north.	n_j Rasik walked 40 m towards north.	E, C
n_i Rasik walked 20 m towards north.	$quant + n_j$ Rasik walked less than 80 m towards north.	E, C
n_i Rasik walked 40 m towards north.	$neg + quant + n_j$ Rasik walked not more than 80 m towards north.	E, C
$quant + n_i$ Rasik walked more than 40 m towards north.	n_j Rasik walked 40 m towards north.	C, N
$quant + n_i$ Rasik walked less than 80 m towards north.	$quant + n_j$ Rasik walked less than 20 m towards north.	E, N
$neg + quant + n_i$ Rasik walked not more than 20 m towards north.	n_j Rasik walked 40 m towards north.	C, N
$neg + quant + n_i$ Rasik walked not less than 40 m towards north.	$neg + quant + n_j$ Rasik walked not less than 20 m towards north.	E, N

Table A1: Dataset variation in Number Comparison with quantifier task. n_i and n_j can be same or different number. $quant$ = quantifier ("more" or "less"), neg = negation ("not").

Original QA	*There are 6 birds and 3 nests. How many are birds more than nests*?		
NLI		P: There are 6 birds and 3 nests H: There are more birds than nests	E
	entity swapping	H: There are more nests than birds.	C
	comparator replacement	H: There are few birds than nests.	C
	double change	H: There are few nests than birds.	E

Table A2: Examples of data for Comparative Reasoning task, recasted from question answering dataset.

Variant	Premise	Hypothesis
	BASIC ARITHMETIC	
P *num* H *num*	Sam had 9 dimes in his bank and his dad gave him 7 dimes.	Sam has 16 dimes now.
P *num* H *word*	Sam had 9 dimes in his bank and his dad gave him 7 dimes.	Sam has sixteen dimes now.
P *word* H *num*	Sam had nine dimes in his bank and his dad gave him seven dimes.	Sam has 16 dimes now.
P *word* H *word*	Sam had nine dimes in his bank and his dad gave him seven dimes.	Sam has sixteen dimes now.
P *mixed* H *num*	Sam had 9 dimes in his bank and his dad gave him seven dimes.	Sam has 16 dimes now.
P *mixed* H *word*	Sam had 9 dimes in his bank and his dad gave him seven dimes.	Sam has sixteen dimes now.
	NUMBER COMPARISON	
num	Adam has 10 apples. Jackie has 2 apples.	Adam has fewer apples than Jackie.
word	Adam has ten apples. Jackie has two apples.	Adam has fewer apples than Jackie.
mixed	Adam has 10 apples. Jackie has two apples.	Adam has fewer apples than Jackie.

Table A3: Examples of data for Basic Arithmetic and Number Comparison task using different number representation. *num*: numbers are represented as digit, *word*: numbers are represented as numeral word, *mixed*: some numbers are represented as digit, some others as numeral.

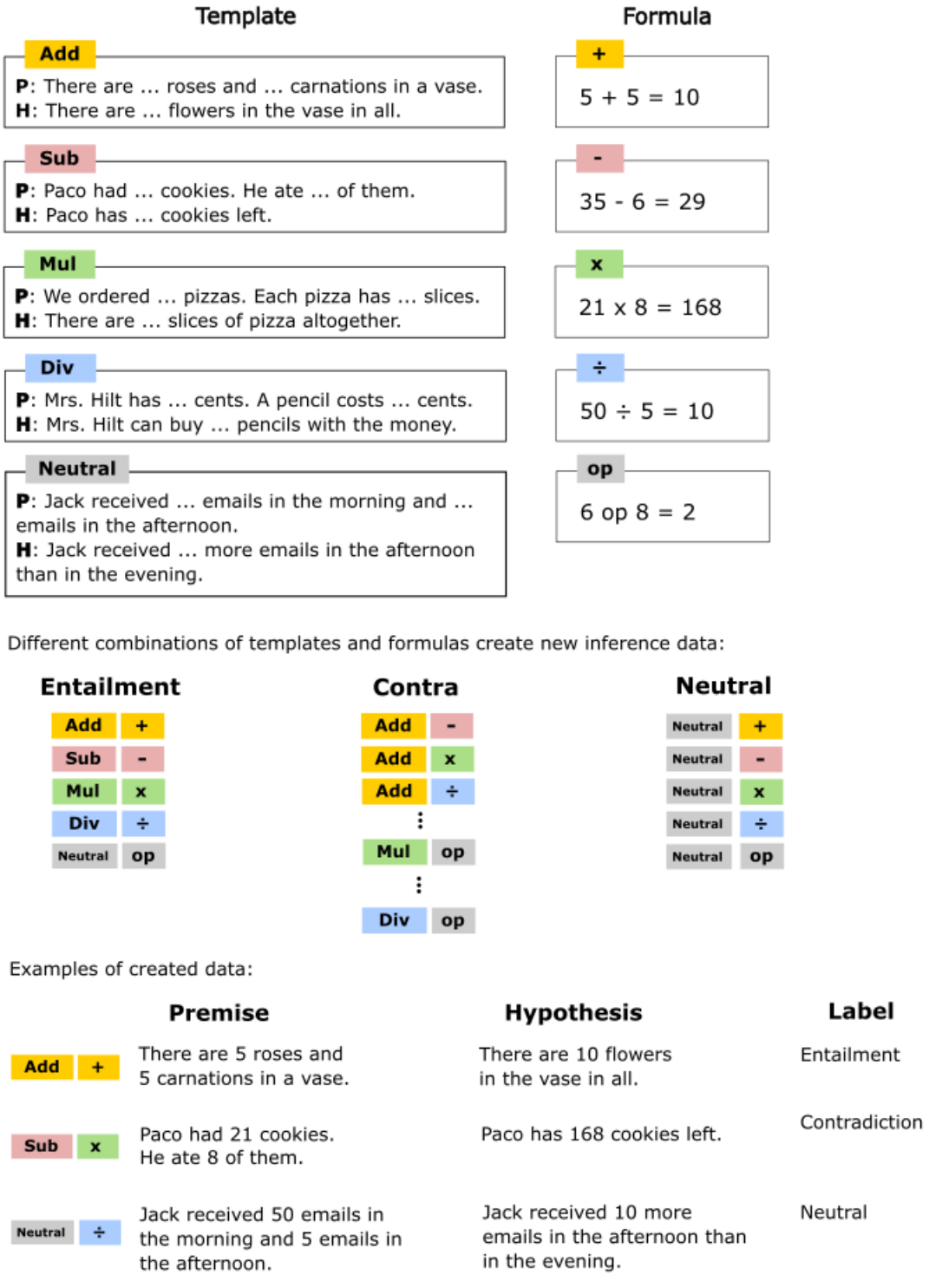

Figure A1: Template-formula data augmentation. We generate NLI data by pairing templates and formulas, following the methodology described in section 3.3.1

Models Family	HF Source
RoBERTa	ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli
ALBERT	ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli
ELECTRA	ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli
BART	ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli
XLNet	ynie/xlnet-large-cased-snli_mnli_fever_anli_R1_R2_R3-nli
DeBERTa-v3	MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli
Flan-T5	google/flan-t5-xl
	google/flan-t5-xxl
Mistral	mistralai/Mistral-7B-v0.1
	mistralai/Mistral-7B-Instruct-v0.3
Llama	meta-llama/Llama-2-7b
	meta-llama/Llama-2-7b-chat
	meta-llama/Meta-Llama-3-8B
	meta-llama/Meta-Llama-3-8B-Instruct
	meta-llama/Llama-3.1-8B
	meta-llama/Llama-3.1-8B-Instruct
Gemma	google/gemma-2b
	google/gemma-2b-it
	google/gemma-7b
	google/gemma-1.1-7b-it
	google/gemma-2-2b
	google/gemma-2-2b-it
	google/gemma-2-9b
	google/gemma-2-9b-it
InternLM	internlm/internlm2-7b
	internlm/internlm2-math-plus-7b
DeepSeek	deepseek-ai/deepseek-llm-7b-base
	deepseek-ai/deepseek-llm-7b-chat
	deepseek-ai/deepseek-math-7b-base
	deepseek-ai/deepseek-math-7b-instruct
	deepseek-ai/deepseek-math-7b-rl
Qwen	Qwen/Qwen2-1.5B
	Qwen/Qwen2-1.5B-Instruct
	Qwen/Qwen2-7B
	Qwen/Qwen2-7B-Instruct
	Qwen/Qwen2-Math-7B
	Qwen/Qwen2-Math-7B-Instruct
	Qwen/Qwen2.5-3B
	Qwen/Qwen2.5-3B-Instruct
	Qwen/Qwen2.5-7B
	Qwen/Qwen2.5-7B-Instruct
	Qwen/Qwen2.5-Math-7B
	Qwen/Qwen2.5-Math-7B-Instruct
Llemma	EleutherAI/llemma_7b
MetaMath	meta-math/MetaMath-7B-V1.0
	meta-math/MetaMath-Llemma-7B
	meta-math/MetaMath-Mistral-7B
WizardMath	WizardLMTeam/WizardMath-7B-V1.1
OpenMath	OpenMath-Mistral-7B-v0.1-hf

Table A4: Models evaluated using the basic numeracy NLI benchmark.

Selected prompt	`[PREMISE]`
	`Based on previous passage, is it true that [HYPOTHESIS]?`
	`True, False, or Maybe?`
	`Answer: ["True", "False", "Maybe"]`

Other prompt candidate (1)	`[PREMISE]`
	`Question: Does this imply that [HYPOTHESIS]? True, False, or Neither?`
	`Answer: ["True", "False", "Neither"]`

Other prompt candidate (2)	`Premise: [PREMISE]`
	`Hypothesis: [HYPOTHESIS]`
	`NLI label: ["Entailment", "Contradiction", "Neutral"]`

Table A5: Prompt used for LLMs, adapted from previous work by Brown et al. (2020), Webson and Pavlick (2022), and Cahyawijaya et al. (2023).

Problem **without** distraction	P: Jason picked 46 pears, Keith picked 47 pears, and Mike picked 12 pears from the pear tree. H: 105 pears were picked in total
Problem **with** distraction, (confounding number)	P: Sally has 9 orange balloons and 4 blue balloons. she found 2 more of the orange balloons. H: Sally has 11 orange balloons now.

Table A6: Data for Arithmetic task, highlighting the problem with distraction.

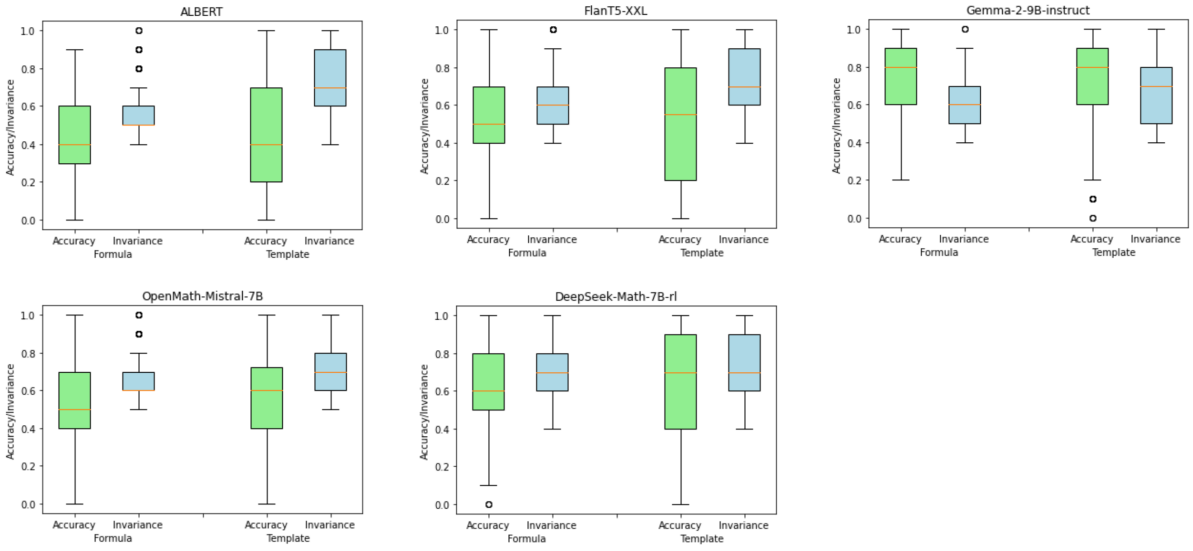

Figure A2: Basic Arithmetic robustness using different template and formula for ALBERT, FlanT5-XXL, Gemma-2-9B-it, OpenMath-Mistral-7B, and DeepSeek-Math-7B-rl.

		Arithmetic	Comparison		Normalization	
			Quantifier	Reasoning	Numeration	Unit
Flan-T5						
	XL (**3B**)	0.428 / 0.416	0.620 / 0.525	0.787 / 0.788	0.934 / 0.936	0.862 / 0.860
	XXL (**11B**)	0.440 / 0.433	0.667 / 0.615	0.801 / 0.808	0.933 / 0.936	0.848 / 0.845
Mistral	**7B**					
	base	0.345 / 0.336	0.369 / 0.348	0.412 / 0.413	0.335 / 0.443	0.461 / 0.477
	instruct	0.408 / 0.403	0.457 / 0.441	0.501 / 0.535	0.736 / 0.818	0.665 / 0.687
Llama-2	**7B**					
	base	0.419 / 0.235	0.438 / 0.335	0.490 / 0.427	0.611 / 0.575	0.575 / 0.569
	chat	0.420 / 0.235	0.422 / 0.225	0.513 / 0.493	0.520 / 0.383	0.557 / 0.476
Meta-Llama-3	**8B**					
	base	0.428 / 0.312	0.473 / 0.371	0.473 / 0.371	0.460 / 0.423	0.495 / 0.418
	instruct	0.539 / 0.503	0.521 / 0.495	0.588 / 0.606	0.895 / 0.932	0.768 / 0.786
Llama-3.1	**8B**					
	base	0.428 / 0.345	0.361 / 0.320	0.428 / 0.431	0.500 / 0.499	0.505 / 0.454
	instruct	0.577 / 0.434	0.536 / 0.397	0.701 / 0.690	0.927 / 0.927	0.781 / 0.771
Gemma	**2B**					
	base	0.377 / 0.295	0.344 / 0.293	0.496 / 0.362	0.440 / 0.397	0.503 / 0.484
	instruct	0.453 / 0.266	0.429 /0.298	0.497 / 0.423	0.563 / 0.458	0.656 / 0.611
Gemma–1.1	**7B**					
	base	0.429 / 0.262	0.377 / 0.295	0.557 / 0.364	0.580 / 0.508	0.589 / 0.541
	instruct	0.580 / 0.504	0.613 / 0.540	0.668 / 0.650	0.969 / 0.973	0.834 / 0.833
Gemma–2	**2B**					
	base	0.373 / 0.329	0.354 / 0.328	0.387 / 0.428	0.427 / 0.459	0.474 / 0.498
	instruct	0.211 / 0.186	0.453 / 0.416	0.186 / 0.281	0.647 / 0.775	0.683 / 0.742
Gemma–2	**9B**					
	base	0.338 / 0.319	0.361 / 0.350	0.330 / 0.395	0.471 / 0.540	0.448 / 0.497
	instruct	0.651 / 0.634	0.632 / 0.561	0.703 / 0.748	0.885 / 0.938	0.778 / 0.811
Qwen2	**1.5B**					
	base	0.367 / 0.348	0.415 / 0.423	0.366 / 0.418	0.444 / 0.502	0.567 / 0.580
	instruct	0.303 / 0.303	0.337 / 0.421	0.241 / 0.335	0.372 / 0.493	0.468 / 0.587
Qwen2	**7B**					
	base	0.427 / 0.414	0.465 / 0.449	0.404 / 0.510	0.559 / 0.672	0.683 / 0.712
	instruct	0.505 / 0.490	0.578 / 0.511	0.694 / 0.721	0.836 / 0.898	0.814 / 0.821
Qwen2–Math	**7B**					
	base	0.481 / 0.414	0.419 / 0.351	0.504 / 0.511	0.568 / 0.553	0.595 / 0.573
	instruct	0.563 / 0.443	0.581 / 0.448	0.656 / 0.647	0.991 / 0.992	0.837 / 0.834
Qwen2.5	**3B**					
	base	0.362 / 0.338	0.373 / 0.364	0.341 / 0.382	0.487 / 0.549	0.545 / 0.538
	instruct	0.406 / 0.400	0.469 / 0.434	0.402 / 0.478	0.648 / 0.751	0.574 / 0.638
Qwen2.5	**7B**					
	base	0.574 / 0.519	0.574 / 0.482	0.686 / 0.700	0.936 / 0.957	0.820 / 0.830
	instruct	0.555 / 0.509	0.576 / 0.480	0.722 / 0.729	0.930 / 0.952	0.831 / 0.851
Qwen2.5-Math	**7B**					
	base	0.509 / 0.452	0.460 / 0.396	0.605 / 0.614	0.651 / 0.682	0.606 / 0.631
	instruct	0.619 / 0.451	0.631 / 0.481	0.728 / 0.727	0.965 / 0.965	0.851 / 0.850

Table A7: LLMs evaluation (including models from Flan-T5, Llama, Gemma, and Qwen) on Basic Arithmetic, Number Comparison, and Number Normalization tasks. Evaluation results are presented as "Accuracy / macro F1" scores.

		Arithmetic	Comparison		Normalization	
			Quantifier	Reasoning	Numeration	Unit
InternLM2	**7B**					
	instruct	0.494 / 0.423	0.542 / 0.424	0.655 / 0.652	0.945 / 0.949	0.876 / 0.875
InternLM2-Math-Plus	**7B**					
	instruct	0.498 / 0.335	0.495 / 0.366	0.502 / 0.371	0.754 / 0.737	0.689 / 0.657
DeepSeek-LLM	**7B**					
	base	0.421 / 0.285	0.410 / 0.303	0.508 / 0.451	0.680 / 0.671	0.578 / 0.561
	chat	0.481 / 0.402	0.472 / 0.340	0.560 / 0.562	0.961 / 0.963	0.783 / 0.775
DeepSeek-Math	**7B**					
	base	0.439 / 0.332	0.423 / 0.313	0.420 / 0.420	0.532 / 0.521	0.567 / 0.546
	instruct	0.528 / 0.378	0.497 / 0.396	0.522 / 0.457	0.776 / 0.767	0.718 / 0.709
	rl	0.567 / 0.442	0.381 / 0.284	0.535 / 0.451	0.922 / 0.923	0.797 / 0.791
Llemma	**7B**					
		0.420 / 0.270	0.404 / 0.333	0.487 / 0.391	0.495 / 0.443	0.513 / 0.544
MetaMath	**7B**					
Metamath(-Llama-2)		0.454 / 0.330	0.402 / 0.304	0.498 / 0.496	0.705 / 0.701	0.541 / 0.470
MetaMath-Llemma		0.443 / 0.261	0.397 / 0.267	0.501 / 0.367	0.585 / 0.499	0.575 / 0.488
MetaMath-Mistral		0.465 / 0.354	0.431 / 0.365	0.526 / 0.474	0.757 / 0.775	0.632 / 0.617
WizardMath	**7B**					
		0.488 / 0.384	0.423 / 0.384	0.529 / 0.472	0.645 / 0.700	0.601 / 0.578
OpenMath	**7B**					
OpenMath-Mistral		0.515 / 0.374	0.496 / 0.369	0.593 / 0.579	0.948 / 0.948	0.676 / 0.658

Table A8: LLMs evaluation (including models from InternLM, DeepSeek, and MetaMath) on Basic Arithmetic, Number Comparison, and Number Normalization tasks. Evaluation results are presented as "Accuracy / macro F1" scores.

	#instances	Accuracy / F1				
		DeBERTa-v3 + NLI	Llama-3.1 -8B-Instruct	Gemma-2 -9B-it	DeepSeek-Math -7B-rl	Qwen2.5-Math -7B-instruct
addition	1,068	0.690 / 0.700	0.647 / 0.639	0.643 / 0.706	0.492 / 0.507	0.774 / 0.752
subtraction	1,299	0.628 / 0.645	0.679 / 0.669	0.628 / 0.687	0.575 / 0.578	0.808 / 0.778
multiplication	664	0.476 / 0.556	0.480 / 0.480	0.623 / 0.653	0.523 / 0.526	0.837 / 0.813
division	761	0.311 / 0.363	0.574 / 0.573	0.586 / 0.643	0.580 / 0.583	0.829 / 0.804
single operations	3,792	0.555 / 0.603	0.614 / 0.609	0.623 / 0.677	0.543 / 0.551	0.808 / 0.782
multi operations	2,632	0.433 / 0.478	0.604 / 0.593	0.580 / 0.642	0.535 / 0.545	0.706 / 0.665

Table A9: Models performance on Basic Arithmetic tasks. Breakdown by arithmetic operations and operands.

	Entailment			Contradiction			Neutral		
	P	R	F1	P	R	F1	P	R	F1
ALBERT + NLI	0.574	0.320	0.411	0.583	0.403	0.476	0.172	0.499	0.256
DeBERTa-v3 + NLI	0.646	0.397	0.492	0.611	0.517	0.560	0.178	0.422	0.251
Flan-T5-XXL	0.615	0.479	0.538	0.761	0.318	0.448	0.208	0.627	0.312
Llama–3.1–**8B**–Instruct	0.548	0.893	0.679	0.630	0.465	0.535	1.000	0.046	0.089
Gemma–2–**9B**–it	0.682	0.777	0.727	0.764	0.479	0.589	0.479	0.753	0.586
InternLM2–**7B**	0.565	0.413	0.477	0.470	0.712	0.566	0.388	0.158	0.224
DeepSeek–LLM–**7B**–chat	0.516	0.554	0.534	0.493	0.556	0.523	0.225	0.111	0.149
Qwen2.5–**7B**–Instruct	0.805	0.415	0.547	0.549	0.770	0.641	0.304	0.377	0.337
OpenMath–Mistral–**7B**	0.464	0.761	0.576	0.480	0.325	0.388	0.337	0.057	0.097
InternLM2–Math-plus–**7B**	0.474	0.901	0.621	0.594	0.284	0.384	0.000	0.000	0.000
DeepSeek–Math–**7B**–rl	0.551	0.913	0.687	0.648	0.403	0.497	0.343	0.090	0.142
Qwen2.5–Math–**7B**–instruct	0.665	0.775	0.716	0.575	0.708	0.635	1.000	0.001	0.002

Table A10: **P**recision, **R**ecall, and **F1**-score Basic Arithmetic tasks across NLI classes.

		Entailment			Contradiction			Neutral		
		P	R	F1	P	R	F1	P	R	F1
ALBERT + NLI	q	0.801	0.650	0.717	0.589	0.934	0.723	0.889	0.299	0.447
	c	0.634	0.887	0.740	0.816	0.471	0.597			
DeBERTa-v3 + NLI	q	0.903	0.659	0.762	0.606	0.938	0.737	0.702	0.347	0.465
	c	0.807	0.781	0.794	0.793	0.809	0.801			
Flan-T5-XXL	q	0.832	0.650	0.730	0.603	0.945	0.737	0.660	0.264	0.377
	c	0.778	0.875	0.824	0.872	0.727	0.793			
Llama–3.1–**8B**–Instruct	q	0.627	0.539	0.580	0.501	0.786	0.612	0.000	0.000	0.000
	c	0.646	0.887	0.748	0.820	0.514	0.632			
Gemma–2–**9B**–it	q	0.867	0.566	0.685	0.582	0.627	0.604	0.519	0.318	0.394
	c	0.779	0.773	0.776	0.833	0.633	0.719			
InternLM2–**7B**	q	0.752	0.454	0.566	0.487	0.861	0.622	0.679	0.045	0.084
	c	0.707	0.543	0.614	0.628	0.766	0.690			
DeepSeek–LLM–**7B**–chat	q	0.525	0.367	0.432	0.457	0.801	0.582	0.350	0.003	0.006
	c	0.561	0.576	0.568	0.567	0.545	0.556			
Qwen2.5–**7B**–Instruct	q	0.931	0.438	0.596	0.509	0.850	0.636	0.554	0.128	0.207
	c	0.835	0.580	0.684	0.699	0.865	0.773			
OpenMath–Mistral–**7B**	q	0.720	0.496	0.588	0.560	0.768	0.648	0.000	0.000	0.000
	c	0.568	0.777	0.656	0.647	0.410	0.502			
InternLM2–Math–plus–**7B**	q	0.655	0.722	0.687	0.619	0.542	0.578	0.000	0.000	0.000
	c	0.501	0.959	0.658	0.524	0.045	0.083			
DeepSeek–Math–**7B**–instruct	q	0.727	0.584	0.647	0.595	0.704	0.645	0.617	0.033	0.063
	c	0.513	0.867	0.645	0.573	0.176	0.270			
Qwen2.5–Math–**7B**–instruct	q	0.805	0.738	0.770	0.557	0.845	0.672	1.000	0.000	0.001
	c	0.764	0.662	0.709	0.702	0.795	0.745			

Table A11: **P**recision, **R**ecall, and **F1**-score on Number Comparison tasks (q: comparing numbers with quantifier, c: comparative reasoning) across NLI classes.

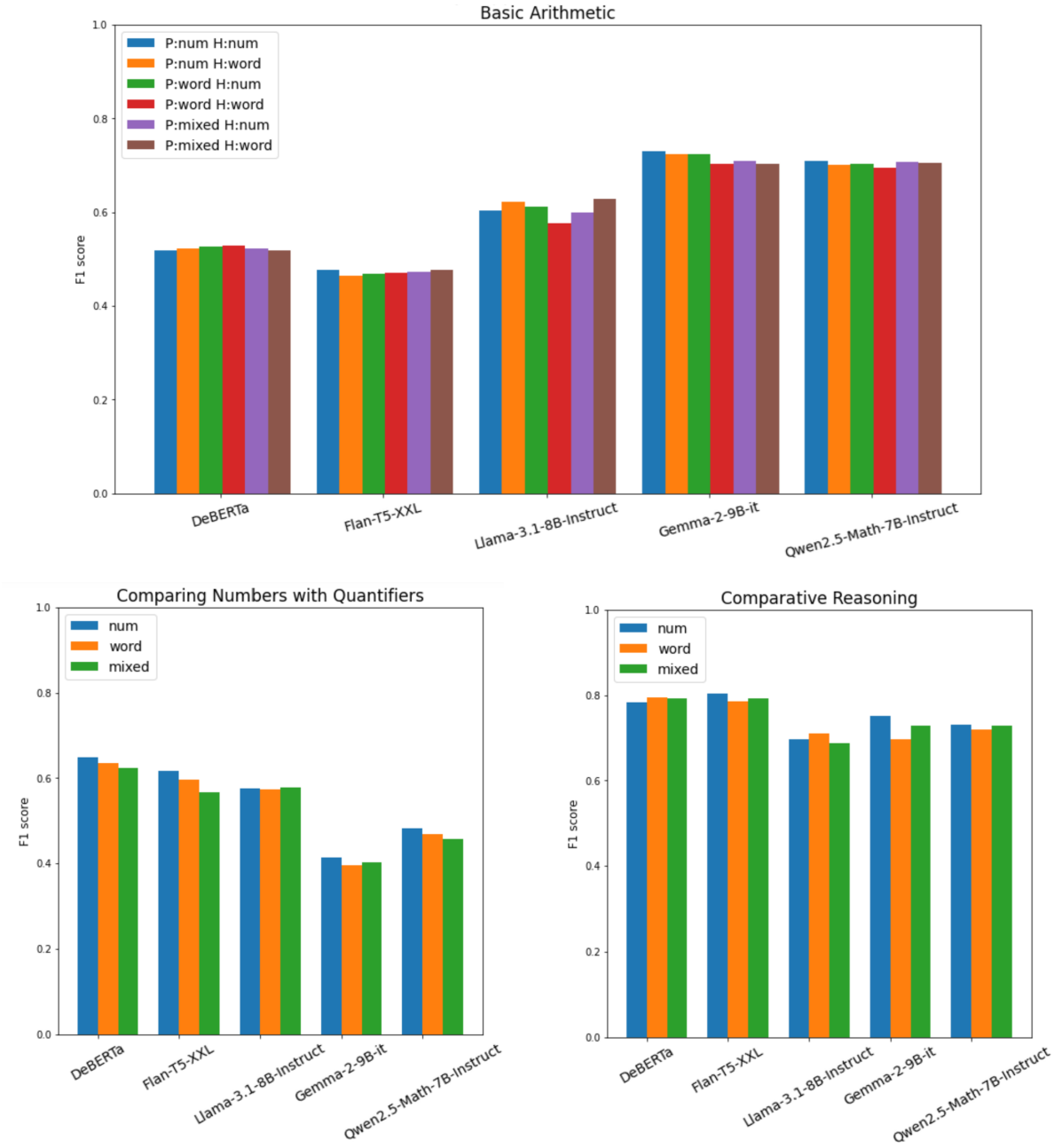

Figure A3: LMs Accuracy on Basic Arithmetic and Number Comparison Tasks Across Number Representation Variation. *num*: numbers are represented as digit, *word*: numbers are represented as numeral word, *mixed*: some numbers are represented as digit, some others as numeral.