Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data

Mubashara Akhtar[1], **Abhilash Shankarampeta**[2], **Vivek Gupta**[3], **Arpit Pital**[4], **Oana Cocarascu**[1] and **Elena Simperl**[1]

[1]King's College London [2]Meesho [3]University of Pennsylvania [4]University of Utah

mubashara.akhtar@kcl.ac.uk

Abstract

Numbers are crucial for various real-world domains such as finance, economics, and science. Thus, understanding and reasoning with numbers are essential skills for language models to solve different tasks. Various benchmarks have assessed the numerical reasoning abilities of language models and revealed their limitations. However, these benchmarks are limited to specific numerical aspects. In this paper, we propose a hierarchical taxonomy for numerical reasoning skills. It comprises more than ten reasoning types across four levels: representation, number sense, manipulation, and complex reasoning. We conduct a comprehensive evaluation of state-of-the-art models across all reasoning types to identify reasoning challenges specific to model types. Therefore, we develop a diverse set of numerical probes employing a semi-automated approach. We focus on the tabular Natural Language Inference (TNLI) task as a case study and measure models' performance shifts. While no single model excels in all reasoning types, FlanT5 (few-/zero-shot) and GPT3.5 (few-shot) demonstrate strong overall numerical reasoning skills compared to other models based on our probing framework.

1 Introduction

Numerical data is ubiquitous in the real-world. Many applications in domains such as finance, economics and science require understanding and reasoning with numbers. In recent years, benchmarks were introduced to study language models' numerical reasoning skills (Zhang et al., 2020; Wallace et al., 2019). However, these datasets mostly concentrate on few, specific numerical reasoning types (e.g. scales (Zhang et al., 2020)). Limitations of language models' numerical abilities, as discussed in prior research, include tokenization and representation of numbers in text (Thawani et al., 2021b), hallucination (Ji et al., 2023; Chen et al., 2023; Ye et al., 2023), and generalizability/robustness issues (Razeghi et al., 2022; Geva et al., 2020).

Hulk	
Directed by	Ang Lee
Release date	June 20, 2003
Running time	138 minutes
Budget	$137 million
Box office	$245.4 million

H1: Hulk was released on 20th June, 2003. *(E)*
Date: Hulk was released on 20-06-2003. *(E)*
Date Flip: Hulk was released on 12-08-2009. *(E)*

H2: The movie has a length of 138 minutes. *(E)*
Appr: The movie has a length of about 150 minutes. *(C)*

H3: The movie can be watched in about two hours. *(E)*
Num: The movie can be watched in about 2 hours. *(E)*
Num Flip: The movie can be watched in about 3 hours. *(C)*

Arith: Hulk brought in $108.4 million profit. *(E)*
Arith Flip: Hulk brought in $120.9 million profit. *(C)*

Table 1: Base hypotheses (H1, H2, H3) and (**flip**ped) probes for heterogeneous numbers (i.e. **date**), **appr**imation, **num**eracy, and **arith**metic. Labelled as **E**ntail or **C**ontradict.

Successful numerical reasoning requires a combination of skillsets: understanding representation of numbers (Thawani et al., 2021a,b) and their meaning in a given context (Loukas et al., 2022), applying operations (Geva et al., 2020; Patel et al., 2021), and integrating factual and commonsense numerical knowledge to solve numerical problems (Lin et al., 2020; Park et al., 2022). For example, classifying the hypotheses *"The movie can be watched in about* 2 *(or 'two') hours."* from Table 1 requires understanding that both *"*2*"* and *"two"* depict the same numerical value, converting *"*2 *hours"* to another unit (i.e. 120 minutes), and applying approximation to map *"*120 *minutes"* to *"*138 *minutes"* in the table.

In this paper, we evaluate state-of-the-art models on various numerical reasoning types. To assess which reasoning types are challenging for specific models, we create a diverse and large set of numerical probes and measure shifts in models' performance.

We organize all probed reasoning types in a hierarchical taxonomy. Inspired by how humans understand and reason with numbers, as well as previous numerical benchmarks, we include eleven reasoning types across four level: *representation*, *number sense*, *manipulation*, and *complex reasoning* (Figure 1). We apply a semi-automated approaches for probe creation. We select tabular NLI (TNLI) as a case study task, given three criteria: (i) numerical data (numbers, percentages, dates, etc.) is prevalent in tables; (ii) tables are common in real-world data sources such as in scientific publications, database systems and financial documents; (iii) tables as structured data facilitate automated perturbations to create large-scale probing sets. See Table 1 for some examples of probes created from hypotheses (H1, H2, H3) and the given table.

Our experiments conclude that large language models (LLMs) like FlanT5 and GPT3.5 perform better than other models on various numerical reasoning tasks. Both table-based and numerical models struggled to understand data with flipped labels and negative values. Moreover, we observe that some models' performance improves significantly for counterfactual probes (e.g. NT5 and TAPAS) and label-flipping probes (e.g. FlanT5 zero-shot), which indicates that models might exploit dataset artifacts and are biased towards one label. These findings emphasize the importance of further systematically investigating numerical reasoning capabilities across various NLP models.

Our **contributions** are as follows:

- We introduce a taxonomy for numerical reasoning skills, including representation/number sense/manipulation skills and complex reasoning with numbers.
- We propose a semi-automated approach to create large-scale, numerical probe sets using table NLI datasets.
- We evaluate three different categories of language models (LMs) on our numerical probe sets: (i) numerical LMs; (ii) LMs for tabular data; and (v) zero-/few-shot LMs.

2 A Taxonomy for Numerical Reasoning

This section introduces a hierarchical taxonomy for numerical reasoning, inspired by previous works on numeracy in NLP (Thawani et al., 2021b; Xu et al., 2022) and psychology (Barrouillet and Fayol, 1998a; Whyte and Bull, 2008; Bofferding, 2019). We group numerical reasoning skills given their complexity level in four categories: $R1 - R4$.

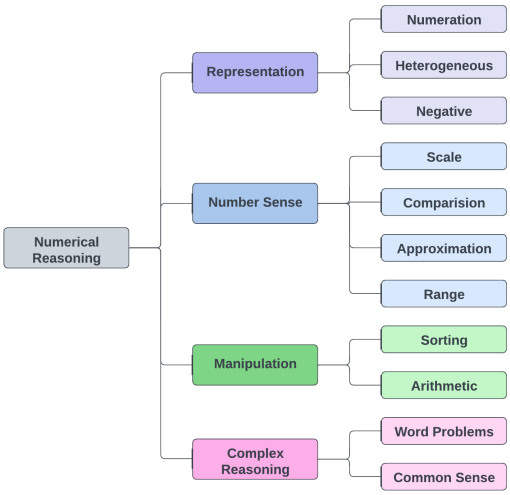

Figure 1: Overview of numerical reasoning types.

2.1 Number Representation ($R1$)

This category includes skills for understanding the *form* of numerical data. Similar to the notion of form in language (Bender and Koller, 2020), this is the realization of numbers in text; the way they are represented and expressed.

Numeration. Numeration studies language model's understanding of representation systems common for numbers in English: the Arabic ("2") and English ("two") numeration systems. Specifically, we probe if LMs can link between distinct symbols used for the same number. For example in Figure 1, H3 contains "two" as a word, which can be also represented through "2".

Heterogeneous Number Types. Formatted numbers (e.g. dates, times, and fractions) are frequently used to convey additional information associated with a numerical value. Numbers are formatted in a specific way given their context and purpose, such as expressing times and dates using full-stop ("."), using the "%" symbol to indicate fractions, and different currency symbols for money (i.e. "$" or "€"). See H1 and "Arith" in Figure 1 for example sentences.

Negative Numbers. Early on in their development, children develop some mental model for negative numbers (see experiments with first-graders in Bofferding (2019)). Using negative numbers requires understanding the notation of negatives (i.e. "−" followed by a number). This also includes distinguishing between minus in subtractions ($1 - 3$),

dates (12-12-2022), counts (i.e. from one to three) and in negative number (-2).

2.2 Number Sense ($R2$)

Number sense includes reasoning skills for conceptualizing number quantities and understanding their meaning in a given context.

Scale. In everyday communication, numbers often occur together with measurement scales, e.g. weights, distances, or heights. Understanding numbers in context of scales is a basis for question answering applications (e.g. *"We are driving* 80 *km h, is this within the speed limit?"*), commonsense (e.g. *"Cats weight between four and five kilograms."*) (Lin et al., 2020), temporal reasoning (e.g. *"She left the office thirty minutes ago."*) (Zhou et al., 2020; Zhang et al., 2020), and other applications.

Comparison. Comparing numbers allows understanding numerical relationships. It involves identifying which numbers are greater than, less than, or equal to others. For example, given the table in Figure 1, understanding *"The running time of Hulk is longer than* 120 *minutes."* requires comparison.

Range. The question *"Was the budget of the movie between* $130 *and* $245.4*?"* about the table in Figure 1 requires understanding number ranges. Already at an age between two and three years, children develop numerical abilities to understand sequences of numbers and start reciting numbers in an appropriate order (Fuson, 2012; Laski and Siegler, 2007). Models' that understand the notation of ranges, can correctly answer the question by knowing that 137 is in the range $130 - 245.4$.

Approximation. Humans commonly approximate number values in everyday life (Odic and Starr, 2018; Bonny and Lourenco, 2013). H3 in Figure 1 requires approximation among other skills to map "about two hours" to "138 minutes" in the table. As a reasoning skill, it allows to make quick estimations and metric unit conversations, and understand the approximate values of numbers without calculating them explicitly.

2.3 Manipulation ($R3$)

Manipulation reasoning types are used to apply basic operations on numbers such as addition. Successful manipulation of numbers requires understanding their *representations* and meaning in the given context (i.e. *number sense*).

Sorting. The sentence *"Out of all Ang Lee's directed movies, 'Hulk' was the one with the second highest box office income."* requires sorting all movies according to their box office income in order to select the one with the second highest income. Sorting objects according to some criteria is a basic milestone for developing cognitive skills. By age 2, children already begin to understand the concept of sorting.

Simple arithmetic. Arithmetic reasoning is the ability of manipulating numbers with basic operations (addition, subtraction, multiplication, division). While adults commonly retrieve results of simple calculations from memory, children use different operations (Barrouillet and Fayol, 1998b).

2.4 Complex Reasoning ($R4$)

This category builds on all previous reasoning categories ($R1 - R3$) to solve numerical word problems (NWP). NWP are expressed through natural language and require multistep reasoning. Extracting information from the problem description and applying numerical/mathematical reasoning using the retrieved information and world/commonsense knowledge is required (Upadhyay and Chang, 2017; Amini et al., 2019; Huang et al., 2016).

3 Numerical Probing Framework

This section provides an overview of the probing framework. We use tabular Natural Language Inference (TNLI) for automated probe creation.

3.1 Preliminaries

Tables for numerical probing. Tables align well with our objectives given three key criteria: (i) numerical data is common in tables; (ii) tables are frequent in real-world data sources; (iii) tables, due to their structured formats, facilitate automated perturbations for probe creation. Tables' semi-structured format, the alignments available between table cells and column/row headers, and the frequency of numbers, make them well suitable for creating numerical probes automatically.

Table NLI. Given a natural language sentence as hypothesis and a tabular premise, the aim of TNLI is to classify if the hypothesis *entails* or *contradicts* the table (Gupta et al., 2020). We use the table NLI datasets *TabFact* (Chen et al., 2020) and InfoTabs (Gupta et al., 2020), as well as recast the table QA datasets TAT-QA (Zhu et al., 2021) and

TabMWP (Lu et al., 2023) to NLI (i.e. *TATQA-NLI*, *TabMWP-NLI*). TAT-QA includes metadata, i.e. annotations of cells and operations per correct answer. This information is not available for any TNLI dataset and is crucial to create probes for specific reasoning types, e.g. *arithmetic reasoning*. Table 2 provides an overview of the TNLI datasets.

Preprocessing. For each of numerical reasoning type, we first identify base TNLI hypotheses and/or tables in the datasets that can be used for automated probe creation. Hereby, we defined a list of reference tokens specific for each reasoning type and to filter relevant dataset samples. For example, we used units of measurements such as "hour", "meter", or "kilogram" filter hypotheses for *scale* probes (see §4 for more details). To recast the TAT-QA dataset, we follow the simple yet effective, rule-based approach proposed by Demszky et al. (2018) for QA to NLI conversion.

3.2 Probes through Structural Perturbation

Overall, we our framework includes three types of probes, created through hypotheses perturbation and counterfactual tables.

1. Hypothesis label-preserving probes We create label-preserving probes changing the base hypothesis such that its meaning is not changed and the initial hypothesis label is preserved. The probes are used to evaluate model's ability to reason and predict the correct label given semantically-equivalent changes.

2. Hypothesis label-flipping probes To generate label-flipping probes, we modify the base hypothesis such that its meaning alters and the label of the probe flips, e.g. from entailment to contradiction. We aim to overcome potential dataset artefacts that might be exploited for label prediction instead of performing numerical reasoning.

These changes are specific to the reasoning types. For example, to flip labels for *scale* probes, we substitute measurement units for a particular scale (e.g. "kilograms") by another unit (e.g. "meters") or introduce errors in conversion of units (e.g. 3 kilometers replaced by $3,000$ meters).

3. Table Probes through Counterfactual Table Editing We also probe with counterfactual tables to evaluate if models rely on spurious patterns in the premise table for label prediction. We filter the counterfactual datasets by Jena et al. (2022) consisting of *{hypothesis; original table; counterfactual table}* for numerical hypotheses.

Dataset	Hypotheses	Tables	Num cells	Probes
TabFact	118,275	16,573	59.00%	214,440
InfoTabs	23,738	2,540	53.6%	19,779
TATQA-NLI	4,947	2,156	59.7%	15,139
ToTTo	1,000	892	45.7%	1,000
TabMWP	283	283	38.3%	238

Table 2: TNLI probing datasets; *num cells* refers to the average ratio of numerical cells in tables.

4 Probing with TNLI Datasets

This section discussed probes in detail and how we created them for each reasoning type from §3.[1]

Numeration. To study models' understanding of string ("two") and numerical (e.g. "2") number representations, we create two types of numeration probes. Onw converting number representations from strings to numeric, while the second category applies the conversion vice versa. We filter hypotheses with numbers written as strings ("two") and substitute them by their numeric counterpart (e.g. "2"). The label-preserving probes are semantically equivalent to the base hypotheses and the label (e.g. *entailment*) is not changed. Label-flipping probes replace the converted number x by a random number in the range of $[x - x*0.5; x + x*0.5]$. For example, the numeration flipping probe of H1 (Table 3) replaces 112 by one hundred and forty-four and flips the label from *entailment* to *contradiction*.

Heterogeneous number types. We created heterogeneous probes for the following categories frequent in the TNLI datasets: date formats, ordinals, percentage, currencies, and scientific notation. To filter base hypotheses, we applied a simple, rule-based approach specific to each category (i.e. dates formats, percentage, ordinals, etc.). To create label-preserving probes we applied representation-level changes which did not change the semantic meaning. For H3, we substituted 3rd June, 1986 by another English date format 03-06-1986. To flip the label, we replaced the date in the adjusted format by a random date, i.e. 15-01-1999. We replaced percentage signs by the token *"percentages"* and vice versa. Similarly, ordinals written as words (*first*) were exchanged by numerical representations ($1st$) and the other way around. For hypotheses with large numbers (e.g. " $116,111,561 " in H3), we introduced scientific notations ($\$116.111561e-6$).

[1]Find details on probe statistics in Appendix A.

Rafael Nadal	
Plays	Left-handed
Born	3 June 1986 (age 37)
Height	1.85 m
Turned pro	2001
Prize money	US$116,111,561 (3rd all-time leader in earnings)

Base Hypothesis H_1	Born in 1986 , Nadal is age 37 currently.
Numeration Probe H_1	Born in nineteen eighty six, Nadal is age thirty seven currently.
Num Flip Probe H_1	Born in nineteen ninety two, Nadal is age forty one currently.
Range Probe H_1	Born in 1986, Nadal is age between 31-43 currently.
Base Hypothesis H_2	The player's birth date is on 3rd June, 1986 .
Heterog Probe H_2	The player's birth date is on 03-06-1986.
Heterog Flip Probe H_2	The player's birth date is on 15-01-1999.
Base Hypothesis H_3	With $116,111,561 prize money, he is the 3rd highest earning all-time player.
Heterog Probe H_3	With $\$116.111561e-6$ prize money, he is the third highest earning all-time player.
Approx Probe H_3	With about $\$116,000,000$ prize money, he is the 3rd highest earning all-time player.
Base Hypothesis H_4	Rafael Nadal has a height of 1.85 meters.
Scale Probe H_4	Rafael Nadal has a height of 185 centimeters.
Scale Flip Probe H_4	Rafael Nadal has a height of 5.2 ft.
Base Hypothesis H_5	After the year 2000, the player Nadal turned pro.
Comparison Probe H_5	After the year 1995, the player Nadal turned pro.
Comparison Flip Probe H_5	Before the year 1990, the player Nadal turned pro.

Table 3: Exemplary hypotheses and non-/flipping probes for evaluated reasoning types

Negative numbers. To create negative probes, we replaced negative numbers $-n$ (e.g. -3) by string equivalents (e.g. *minus* 3; *negative* 3) and evaluated changes in model performances on these semantically same sentence pairs. For label-flipping probes, we converted negative numbers into the positive counterpart n. For example, converting *"The company's monthly closing resulted in -5 million USD."* to *"The company's monthly closing resulted in 5 million USD."* flips the label.

Scale. We created two types of scale probes: (i) *conversion*; (ii) *mapping*. Conversion convert numbers within a measurement scale. For H4 in Table 3, we converted the number and measurement unit (1.85 meters) to the next smaller unit within the same scale (185 centimeters) for the label-preserving probe. For label-flip, we introduced an error in the converted number, i.e. converting 1.85 meters. to 5.2 ft instead of 6.07 ft. Mapping probes replace the number and measurement unit by an equivalent (e.g. $1.85m$ by 1.85 meters) for label-preserving probes and a random measurement unit e.g. $1.85m$ to 1.85 kilograms) to flip the base hypotheses.

Comparison. We first created a list of signal word-pairs by prompting GPT3.5. The list includes pairs such as {"bigger":"smaller"}, {"taller":"shorter"}, and {"faster":"slower"}. Using these pairs and their synonyms, we filtered base hypotheses and created three types of comparison probes. First, changing the signal word with its opposite counterpart to flip labels (see H5 in Table 3 flipping "after" to "before"). Second, altering the number such that the comparison and label do not change: replacing "after 2000" by "after 1995" ($H5$). Finally, we combine both prior approaches to create label-flipping probes, e.g. "Before the year 1990, the player Nadal turned pro."s

Approximation. We first extract a number n from our base hypothesis and given the value of n, we decide the magnitude of rounding to apply. While smaller numbers are rounded to tens, larger number are rounded to hundreds, thousands or larger decimal points. For example, we created the probe *"With about $\$116,000,000$ prize money, he is the 3rd highest earning all-time player"* by rounding n equal $116,111,561 to "about $\$116,000,000$" (H3 in Table 3).

Range. To create range probes, we substitute number n in the base hypothesis by an appropriate range, e.g. 37 by "between 31-43" (H1). We define the radius of the range and its boundaries automatically given the value of n. For example, given $n < 10$, we randomly sample a radius be-

tween $1-5$. For $n=7$ and a sampled radius of 2, the range will be $[5-9]$. We select decimal boundaries if n is a decimal number.

Sorting. We utilized table columns as number sequences to create sorting probes. We generated a list of position indicators in number sequences (e.g. "top", "second" "3rd","biggest", "lowest"). These words were used to filter base hypotheses. To create label-flipping probes, we changed the position of the sequence to another one. For instance, we modified "in the first quarter of 2018" to "in the third quarter of 2018" by selecting the value from the third row instead of the first.

Simple arithmetic. Using on TATQA-NLI its metadata indicating the involved numbers and operations for numerical reasoning, we created arithmetic probes. We extracted probes involving addition, subtraction, multiplication, and division. Additionally, we generated label-flipping probes by replacing the operation output (e.g. result of subtraction) in the hypothesis with a different number. In Table 1, the *"Arith"* probe involves calculating the difference between the *budget* and *box office* values to determine the correctness of 108.4. The flipped arithmetic probe produces a close but incorrect subtraction output, 120.9.

Numerical word problems. We converted TabMWP questions and answers into declarative hypotheses. TabMWP is a dataset of free-text math word problems that involve reasoning with tabular data. For label-flipping probes, we substituted numbers in the hypotheses with random numbers from the same column.

Counterfactual Table NLI Probes We filtered the counterfactual ToTTo (Parikh et al., 2020) dataset by Jena et al. (2022) for numerical hypothesis. To create counterfactual tables, they swap two or more table cells to modify the tables such that the label of the respective hypothesis changes from entailment to contradiction and vice versa.

5 Experiments and Analysis

Next, we provide an overview of all models that were evaluated on the probes from §4. We also discuss the obtained results and insights.

5.1 Probed Models

We use state-of-the-art models which are divers in terms of architecture, size, and training setup, grouped into three categories:

$(C1)$ **Numerical LMs.** This category includes LMs adapted for numerical reasoning. *LUNA* (Han et al., 2022) is a recent transformer-based model with an adapted tokenization approach for numbers. The model encodes numbers as single tokens (e.g. $3,201$) instead of splitting them down into subwords or binned tokens. *NT5* (Yang et al., 2021) is a variation of the T5 model. It has been modified for numerical reasoning through additional pretraining objectives and fine-tuning using numerical datasets. *PASTA* (Gu et al., 2022) is based on DeBERTa and is pretrained with objectives that use table-based numeric operations.

$(C2)$ **LMs for tabular reasoning.** *TAPAS* (Herzig et al., 2020) extends the BERT encoder with table-specific embeddings. We used a TAPAS model variant pretrained through intermediate pretraining on synthetic and counterfactual data (Eisenschlos et al., 2020). Previous works have also shown the success of the **BERT family of models* on tabular NLI tasks (Herzig et al., 2020; Yin et al., 2020; Neeraja et al., 2021; Shankarampeta et al., 2022). Tables are either linearized or preprocessed into sentences or structured formats. These transformed tables are then used as input to the models. We used a DeBERTa model (He et al., 2021) trained on multiple NLI datasets.

$(C3)$ **Large LMs.** For few-/zero-shot evaluation, we selected FlanT5 and GPT3.5. We probed FlanT5 in both a few-shot and zero-shot setting. However, we limited the probing on GPT3.5 to few-shot due to its accessibility through a paid API.[2]

5.2 Training and Evaluation

To finetune models, we used the base hypotheses of the training datasets (e.g. InfoTabs) and evaluated models only on probes created with their testsets. The few-shot models were prompted with 2-shot extrapolation. We evaluated all models in a 3-step process: (1) evaluation of base hypotheses H; (2) evaluation of probes P, created using H; (3) calculating changes in model performance by comparing accuracy of P to H. As our TNLI task is a binary classification task, we used accuracy for evaluation.

[2]Unlike GPT3.5, FlanT5 is available free of charge.

Model	Table Specific		Numerical Specific			Large LMs		
Reasoning Type	**TAPAS**	**DeBERTa**	**NT5**	**LUNA**	**PASTA**	**FlanT5 (few)**	**FlanT5 (zero)**	**GPT3.5 (few)**
Numeration	-0.32	-1.82	-4.18	-5.22	-7.7	1.28	-8.84	-1.94
Heterogeneous	- 4.03	-2.36	-3	-10.09	-7.76	0.34	-5.49	-3.86
Negative Numbers	- 46.11	- 13.77	-94.48	-75.55	-10.68	19.21	42.3	5.58
Numeration Flipping	-38.87	4.09	-48.53	-71.35	-25.85	-78.37	33.38	-13.5
Heterogeneous Flipping	-9.57	8.53	-1.97	-43.48	-23.59	-53.44	86.6	-6.52
Negative Numbers Flipping	-64.81	- 41.56	-17.87	76.85	-70.58	- 83.92	173.14	79.27
Scale	0.03	- 6.25	0	-11.43	-1.56	- 9.45	-7.05	2.21
Comparison	- 21.8	- 18.18	- 29.19	- 30	- 35.11	29.38	140.82	65.52
Approximation	- 5.61	- 6.65	- 9.55	- 7.67	- 27.44	-9.66	-12.94	- 6.45
Range	- 18.89	- 33.77	-20.43	-86.77	-84.66	22.44	178.13	44.82
Scale Flipping	- 23.73	-64.58	0	- 68.44	-51.66	-69.56	93.77	-9.6
Comparison Number Flipping	57.67	-19.36	-29.62	-0.28	-19.10	- 8.47	-40.75	21.6
Simple Arithmetic Flipping	- 58.62	-24.96	-27.1	7.07	-49.06	-71.53	265.07	15.3
Sorting Flipping	-34.8	28.66	-22.6	54.31	-4.9	-86.67	25.0	-36.89
Complex Reasoning	63.37	6.93	-3.22	41.41	-50.84	-89.77	-40	-67.66
Counterfactual	44.5	55.54	159.30	0.98	-6.09	61.5	12.23	-5.03

Table 4: Probing results given as accuracy difference (in %) between base hypotheses and probes.

5.3 Results and Discussion

Table 4 gives on overview of all probing results. If available, we separately list scores for flipped probes, e.g. *numeration* and *numeration flipping*.

$(Q1)$ **Do any models excel in all numerical reasoning types?** While there is not one best-performing model across all reasoning types and different models struggle with different types, FlanT5 and GPT3.5 show overall good numerical reasoning skills. While GPT3.5 performance drops by -67.66% for complex reasoning probes, the model's average accuracy change is around 5.0% for other types. This can be related to (1) models pretraining data and (2) training on chain-of-thought reasoning tasks (Wei et al., 2022). GPT3.5 was trained on more than 300 TB data Common Crawl, allowing the model to memorize much more numerical data than other probed models. In comparison, DeBERTa was trained on only $78GB$ of data (He et al., 2021). Interesting is also the performance difference between NT5 and FlanT5. Both models use the T5 model as the base model. FlanT5 was finetuned using instructions and chain-of-thought reasoning, and outperforms NT5.

$(Q2)$ **How do models perform on different types of numerical reasoning? Representation.** In Table 4, comparing representation probes (rows $2-7$), TAPAS and few-shot FlanT5 perform best on non-flipping numeration probes. FlanT5 (few) also performs well on heterogeneous probes, followed by DeBERTa (-2.4%) and NT5 (-3%). TAPAS, NT5, and LUNA show significant performance drops (between -38.87 and -71.35) on negative number probes. This could be because the models exploit correlations between the "$-$" sign and labels for predicting base hypotheses. Interestingly, few- and zero-shot models like FlanT5 and GPT3.5 show improvements on negative number probes. This may be because the models understand "minus 22" as a negative number but not "-22." We discuss label-flipping probes for numeration below.

Number sense. Comparing model performance on number sense probes (rows $8-13$), we observe different patterns for fine-tuned models and few-/zero-shot models. Fine-tuned models struggle especially with comparison probes, with a -29% average performance drop. Scale probes show a -3.4% decrease, while approximation probes report a -13.3% decrease. In contrast, FlanT5 and GPT3.5 perform better on comparison and range probes, sometimes surpassing predictions on the base hypotheses. All models demonstrate lower performance on approximation probes compared to the base hypotheses, with PASTA showing the largest decrease of -27.44%.

Manipulation and Complex Reasoning. Fine-tuned models exhibit an average accuracy drop of -23.5%, except for LUNA which shows performance increases. In contrast, few-/zero-shot models slightly improve performance by 9.5%. Unlike most other reasoning types, fine-tuned models outperform few-/zero-shot models on complex reasoning probes. TAPAS achieves the highest accuracy, followed by LUNA and DeBERTa. FlanT5 and GPT3.5 demonstrate the largest performance drops on complex reasoning probes.

$(Q3)$ **Do models perform similarly for flipped and non-flipped probes?** We observe higher performance drops for label-flipping probes compared to non-flipping probes across models. Models that struggle with flipping probes but perform

well on their non-flipping counterparts indicate a reliance on spurious patterns for label prediction. For example, TAPAS experiences an accuracy drop of -2.28% on numeration probes, but a drop of -45.98% on numeration flipping probes. Similarly, DeBERTa performs comparatively well on scale probes (-6.25%) compared to scale flipping probes (-64.58%). Minor performance drops are found across models for numeration, heterogeneous, and scale probes, suggesting good reasoning skills of models in these categories. Additionally, DeBERTa exhibits robust performance on number flipping probes for sorting and FlanT5 on negative numbers, as well as arithmetic probes.

($Q4$) **Are numerical and table-specific models better for numerical tabular reasoning?**

Numerical models. LUNA, a transformer model that uses a specific tokenization method for numbers, performs similarly to other models on many reasoning types. The only reasoning type where LUNA outperforms others is comparison flipping probes, with a small improvement of 0.28%. PASTA is a DeBERTa-based model trained on numerical data and pretraining objectives. However, compared to DeBERTa, it only performs better on negative number and scale probes.

Table-based models. Comparing non-flipping and flipping probing results for TAPAS, we observe huge accuracy decreases for label flipping cases. For example, for numeration non-flipping probes, TAPAS shows a small decrease (-2.28%). However, when the labels are flipped, the model misclassifies almost half of these probes, resulting in a large accuracy drop of 45.98%. Similarly, for scale probes, the accuracy of non-flipping probes is 2.33%, but it decreases to -31.09% for flipped probes. Compared to other models, TAPAS performs well on heterogeneous probes, non-flipping scale probes, and complex reasoning probes.

6 Related Work

Numeracy Taxonomies in NLP Prior works have introduced different taxonomies to organise numeracy in NLP research. Thawani et al. (2021b) discuss number representations in NLP systems and introduce a taxonomy based on on granularity (exact vs. approximate) and units (abstract vs. grounded) of numbers in text. Xu et al. (2022) focus on the robustness of NLP models in handling numerical data and organize their numeracy probing tasks in two broad categories: (i) number detection and extraction and (ii) semantic parsing of numbers. For (i), they evaluate number mapping between numerals and numbers as words and consider float numbers while studying number detection. In category (ii), they concentrate on arithmetic reasoning. The DROP benchmark (Dua et al., 2019) requires various arithmetic (e.g. substraction, count, sort) and NLU skills (e.g coreference resolution) to answer questions over paragraphs.

Language Model / Numerical Skills Various studies have evaluated LMs' numerical skills in recent years. Earlier works probed word embeddings for numeration (e.g. 4=four) (Naik et al., 2019), comparison (e.g. $3 < 4$) (Wallace et al., 2019), scale (Zhang et al., 2020), and superlatives (Wallace et al., 2019). More recent works evaluate LMs on out-of-distribution numbers (Kim et al., 2021), numeration/magnitude/sorting/superlatives (Pal and Baral, 2021), and arithmetic (Muffo et al., 2022).

Numerically-tuned Language Models Various numerical LMs have been developed in recent times. Geva et al. (2020) and Liang et al. (2022) inject numerical skills into BERT through numerical pretraining objectives. PASTA (Gu et al., 2022) and NT5 (Yang et al., 2021), which are based on DeBERTa and T5 respectively, fall into the same category of models. Another line of work adjusts LMs' architectures for numerical reasoning through numerical tokenization (Han et al., 2022) or additional, numerical embeddings (Jin et al., 2021).

Augmenting LMs with external Calculators todo

Systematic Probes for Tables Tables have been utilized previously used to create probes for table grounding (Gupta et al., 2022b) or recasting non-NLI datasets (e.g. question-answering) to NLI (Jena et al., 2022). Unlike unstructured text data, tables have a natural structure that allows creating controlled experiments more easily (Gupta et al., 2022a). We drew inspiration from prior tabular probing approaches and extended them for automating probing of numerical tabular data. Jena et al. (2022) systematically applying adjustments to table QA datasets to generate NLI data. While we follow a similar approach for probe creation, they focus on transforming QA transformation for data creation, emphasizing the end-result (i.e. the NLI data), rather than the reasoning behind the

answers.

Comparison to Prior Work All the above mentioned prior works on numerical reasoning have provided motivation for our research. However, their evaluations have focused on a narrow range of reasoning types and models. Most study only concentrated on one specific model such as T5 (Pal and Baral, 2021), GPT3 (Muffo et al., 2022), or BERT (Park et al., 2022). In contrast, our framework provides a comprehensive evaluation of numerical reasoning skills. We cover a wide spectrum of complexity levels, ranging from representation to complex reasoning. Moreover, we assess a variety of models with diverse architectures, sizes, and training settings for numerical reasoning.

7 Conclusion

This paper presents a framework for probing language models' numerical reasoning skills. We organise skills in a taxonomy and generate large-scale sets of probes covering more then ten reasoning types. Using table NLI as a case study, we evaluate the numerical reasoning abilities of seven models. These models belong to the categories numerical LMs, tabular LMs, and few-/zero-shot LLMs. We discuss reasoning types that prove challenging for the probed models and explore promising directions for future research.

Limitations

This work proposes a taxonomy and framework to probe numerical reasoning skills in LMs. It involves the creation of large-scale probing sets using an automated approach. However, the evaluation of this approach is currently limited to the task of table NLI. For future research, it is interesting to extend this to include additional tasks and datasets. This extension serves two purposes: first, it allows evaluating a more diverse range of datasets. Second, it enables including challenges specific to other tasks.

In this paper, the evaluation of most reasoning types primarily involves structural changes at the hypotheses level. While we include counterfactual table probes, they are limited to one dataset and perturbations method only. Further research is needed to study models' performance on numerical data in the premise data. Therefore, we need table-based probes for all reasoning types of the proposed taxonomy.

Ethics Statement

In this paper, we study the numerical reasoning skills of different LMs. However, to deploy these systems in real-world applications, further studies and evaluations specific to the intended use cases are required. In order to support future research, we plan to release all the scripts and resources used for probe creation and model evaluation. This will facilitate and encourage further research in this field.

References

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.

Pierre Barrouillet and Michel Fayol. 1998a. From algorithmic computing to direct retrieval: Evidence from number and alphabetic arithmetic in children and adults. *Memory &amp Cognition*, 26(2):355–368.

Pierre Barrouillet and Michel Fayol. 1998b. From algorithmic computing to direct retrieval: Evidence from number and alphabetic arithmetic in children and adults. *Memory & Cognition*, 26:355–368.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Laura Bofferding. 2019. *Understanding Negative Numbers*, pages 251–277. Springer International Publishing, Cham.

Justin W. Bonny and Stella F. Lourenco. 2013. The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology*, 114(3):375–388.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. PURR: efficiently editing language model hallucinations by denoising language model corruptions. *CoRR*, abs/2305.14908.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations,*

ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR*, abs/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.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Karen C Fuson. 2012. *Children's counting and concepts of number*. Springer Science & Business Media.

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.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4971–4983, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vivek Gupta, Riyaz A. Bhat, Atreya Ghosal, Manish Shrivastava, Maneesh Singh, and Vivek Srikumar. 2022a. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *Transactions of the Association for Computational Linguistics*, 10:659–679.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. 2022b. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3283, Dublin, Ireland. Association for Computational Linguistics.

Hongwei Han, Jialiang Xu, Mengyu Zhou, Yijia Shao, Shi Han, and Dongmei Zhang. 2022. LUNA: language understanding with number augmentations on transformers via number plugins and pre-training. *CoRR*, abs/2212.02691.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. 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.

Aashna Jena, Vivek Gupta, Manish Shrivastava, and Julian Eisenschlos. 2022. Leveraging data recasting to enhance tabular reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4483–4496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Zhihua Jin, Xin Jiang, Xingbo Wang, Qun Liu, Yong Wang, Xiaozhe Ren, and Huamin Qu. 2021. Numgpt: Improving numeracy ability of generative pre-trained models. *CoRR*, abs/2109.03137.

Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang, and Sung-Hyon Myaeng. 2021. Have you seen that number? investigating extrapolation in question answering models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7031–7037, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elida V. Laski and Robert S. Siegler. 2007. Is 27 a big number? correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, 78(6):1723–1743.

Zhenwen Liang, Jipeng Zhang, Lei Wang, Wei Qin, Yunshi Lan, Jie Shao, and Xiangliang Zhang. 2022. MWP-BERT: Numeracy-augmented pre-training for math word problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 997–1009, Seattle, United States. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Matteo Muffo, Aldo Cocco, and Enrico Bertino. 2022. Evaluating transformer language models on arithmetic operations using number decomposition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 291–297. European Language Resources Association.

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.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Darko Odic and Ariel Starr. 2018. An introduction to the approximate number system. *Child Development Perspectives*, 12(4):223–229.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. 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.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhilash Shankarampeta, Vivek Gupta, and Shuo Zhang. 2022. Enhancing tabular reasoning with pattern exploiting training. 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 706–726, Online only. Association for Computational Linguistics.

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.

Shyam Upadhyay and Ming-Wei Chang. 2017. Annotating derivations: A new evaluation strategy and dataset for algebra word problems. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 494–504, Valencia, Spain. Association for Computational Linguistics.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. 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*.

Jemma Catherine Whyte and Rebecca Bull. 2008. Number games, magnitude representation, and basic number skills in preschoolers. *Developmental Psychology*, 44(2):588–596.

Jialiang Xu, Mengyu Zhou, Xinyi He, Shi Han, and Dongmei Zhang. 2022. Towards robust numerical question answering: Diagnosing numerical capabilities of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7950–7966, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng-Jian Yang, Ying-Ting Chen, Yuechan Chen, and Daniel Cer. 2021. Nt5?! training T5 to perform numerical reasoning. *CoRR*, abs/2104.07307.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *CoRR*, abs/2301.13808.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Probe Statistics

Reasoning Type	Count
Word Problems	238
Sorting	379
Counterfactual	1,000
Currency	1,014
Negative	3,316
Range	4,208
Scientific notation	6,274
Arithmetic	8,082
Ordinal	10,569
Percentage	16,851
Date	18,642
Approximation	20,440
Comparison	30,763
Numeration	166,319
Total	288,095
Flipped probes	77,687

Table 5: Breakdown of probes per reasoning type.

Table 2 gives an overview of probes per dataset. Most probes (i.e. $214,440$) are created from TabFact hypotheses as this is also the biggest dataset available, followed by InfoTabs ($19,779$). Table 5 provides a breakdown of probes per reasoning type. In total, we have $286,857$ probes, of which $76,404$ are label-flipping probes.

B Insights

Main Insights. We investigated the language models and found that LLMs like FlanT5 and GPT3.5 perform better than other models on various numerical reasoning tasks. When the labels are switched around and when dealing with negative values, we found that both table-based and numerical models had difficulty comprehending the data. In contrast, DeBERTa performs relatively well compared to models like LUNA and PASTA, which are tuned for improved numerical reasoning skills.

In the ideal scenario with counterfactual tables, the models' performance should be similar to the performance on the original tables. However, we observed that TAPAS and DeBERTa's performance improved significantly, which leads to the conclusion that models are biased toward one label.

Overall no language model excels in all the numerical reasoning tasks. Surprisingly, models per-

form relatively well in complex tasks like Numerical Word Problems but struggle at simple reasoning tasks like numeration and comparison.