# Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, Dan Roth
Microsoft Research, Singapore Management University, University of Pennsylvania



Do **Large Language Models (LLMs)** have the capability to **engage in mathematical reasoning** when presented with a **combination of tables and text?**

Do **Large Language Models (LLMs)** have the capability to **engage in mathematical reasoning** when presented with a **combination of tables and text?**

Page 91 from the annual reports of GRMN (Garmin Ltd.)

The fair value for these options was estimated at the date of grant using a Black-Scholes option pricing model with the following weighted-average assumptions for 2006, 2005 and 2004:

| | 2006 | 2005 | 2004 |
|---|---|---|---|
| Weighted average fair value of options granted | $20.01 | $9.48 | $7.28 |
| Expected volatility | 0.3534 | 0.3224 | 0.3577 |
| Distribution yield | 1.00% | 0.98% | 1.30% |
| Expected life of options in years | 6.3 | 6.3 | 6.3 |
| Risk-free interest rate | 5% | 4% | 4% |

... The total fair value of shares vested during 2006, 2005, and 2004 was $9,413, $8,249, and $6,418 respectively. The aggregate intrinsic values of options outstanding and exercisable at December 30, 2006 were $204.1 million and $100.2 million, respectively. *( ... abbreviate 10 sentences ... )*

**Question:** Considering the weighted average fair value of options , what was the change of shares vested from 2005 to 2006?

**Answer:** - 400

**Calculations:**

$$\left(\frac{9413}{20.01}\right) - \left(\frac{8249}{9.48}\right) = -400$$

**Program:**

divide ( 9413, 20.01 ) divide ( 8249, 9.48 )

substract ( #0, #1 )

Figure 1: An example from FINQA: The system needs to learn how to calculate the number of shares, then select relevant numbers from both the table and the text to generate the reasoning program to get the answer.

Do **Large Language Models (LLMs)** have the capability to **engage in mathematical reasoning** when presented with a **combination of tables and text?**

**A. Yes**

Page 91 from the annual reports of GRMN (Garmin Ltd.)

The fair value for these options was estimated at the date of grant using a Black-Scholes option pricing model with the following weighted-average assumptions for 2006, 2005 and 2004:

| | 2006 | 2005 | 2004 |
|---|---|---|---|
| Weighted average fair value of options granted | $20.01 | $9.48 | $7.28 |
| Expected volatility | 0.3534 | 0.3224 | 0.3577 |
| Distribution yield | 1.00% | 0.98% | 1.30% |
| Expected life of options in years | 6.3 | 6.3 | 6.3 |
| Risk-free interest rate | 5% | 4% | 4% |

... The total fair value of shares vested during 2006, 2005, and 2004 was $9,413, $8,249, and $6,418 respectively. The aggregate intrinsic values of options outstanding and exercisable at December 30, 2006 were $204.1 million and $100.2 million, respectively. *( ... abbreviate 10 sentences ... )*

**Question:** Considering the weighted average fair value of options , what was the change of shares vested from 2005 to 2006?

**Answer:** - 400

**Calculations:**

$$\left(\frac{9413}{20.01}\right) - \left(\frac{8249}{9.48}\right) = -400$$

**Program:**

divide ( 9413, 20.01 ) divide ( 8249, 9.48 )

substract ( #0, #1 )

Figure 1: An example from FINQA: The system needs to learn how to calculate the number of shares, then select relevant numbers from both the table and the text to generate the reasoning program to get the answer.

Do **Large Language Models (LLMs)** have the capability to **engage in mathematical reasoning** when presented with a **combination of tables and text?**

**A. Yes, but only upto some extent**

Page 91 from the annual reports of GRMN (Garmin Ltd.)

The fair value for these options was estimated at the date of grant using a Black-Scholes option pricing model with the following weighted-average assumptions for 2006, 2005 and 2004:

| | 2006 | 2005 | 2004 |
|---|---|---|---|
| Weighted average fair value of options granted | $20.01 | $9.48 | $7.28 |
| Expected volatility | 0.3534 | 0.3224 | 0.3577 |
| Distribution yield | 1.00% | 0.98% | 1.30% |
| Expected life of options in years | 6.3 | 6.3 | 6.3 |
| Risk-free interest rate | 5% | 4% | 4% |

... The total fair value of shares vested during 2006, 2005, and 2004 was $9,413, $8,249, and $6,418 respectively. The aggregate intrinsic values of options outstanding and exercisable at December 30, 2006 were $204.1 million and $100.2 million, respectively. *( ... abbreviate 10 sentences ... )*

**Question:** Considering the weighted average fair value of options , what was the change of shares vested from 2005 to 2006?

**Answer:** - 400

**Calculations:**

$$\left(\frac{9413}{20.01}\right) - \left(\frac{8249}{9.48}\right) = -400$$

**Program:**

divide ( 9413, 20.01 ) divide ( 8249, 9.48 )

substract ( #0, #1 )

Figure 1: An example from FINQA: The system needs to learn how to calculate the number of shares, then select relevant numbers from both the table and the text to generate the reasoning program to get the answer.

Revenue from external customers, classified by significant product and service offerings, was as follows:

| (in **millions**) | | | |
|---|---|---|---|
| Year Ended June 30, | 2019 | 2018 | 2017 |
| Server products and cloud services | 32,622 | 26,129 | 21,649 |
| Office products and cloud services | 31,769 | 28,316 | 25,573 |
| Windows | 20,395 | 19,518 | 18,593 |
| Gaming | **11,386** | **10,353** | 9,051 |
| Search advertising | 7,628 | 7,012 | 6,219 |
| **LinkedIn** | 6,754 | **5,259** | **2,271** |
| **Enterprise Services** | **6,124** | 5,846 | 5,542 |
| **Devices** | **6,095** | **5,134** | 5,062 |
| **Other** | 3,070 | 2,793 | **2,611** |
| Total | **$125,843** | **$110,360** | **$96,571** |

Our commercial cloud revenue, which includes Office 365 Commercial, Azure, the commercial portion of LinkedIn, Dynamics 365, and other commercial cloud properties, was **$38.1 billion**, **$26.6** billion and **$16.2** billion in fiscal years 2019, 2018, and 2017, respectively. These amounts are primarily included in Office products and cloud services, Server products and cloud services, and LinkedIn in the table above.

| # | Reasoning | Question | Answer | Scale | Derivation |
|---|---|---|---|---|---|
| 1 | Word Matching (38.06%) | How much revenue came from Linkedin in 2018? | **5,259** | million | - |
| 2 | Set of spans (11.94%) | Which were the bottom 2 revenue items for 2017? | **LinkedIn, Other** | - | - |
| 3 | Comparison (5.65%) | Which year has the lowest revenue? | **2017** | - | - |
| 4 | Counting (2.28%) | How many revenue items are between 6,000 million and 6,500 million in 2019? | 2 | - | **Devices ## Enterprise Services** |
| 5 | Addition (2.37%) | What is the total revenue of commercial cloud from 2017 to 2018? | 42.8 | billion | **26.6 + 16.2** |
| 6 | Subtraction (16.17%) | How much of the total revenue in 2018 did not come from devices? | 105,226 | million | **110,360 - 5,134** |
| 7 | Division (3.84%) | How much does the commercial cloud revenue account for the total revenue in 2019? | 30.28 | % | **38.1 billion / 125,843 million** |
| 8 | Composition (19.69%) | What was the percentage change in gaming between 2018 and 2019? | 9.98 | % | **(11,386 - 10,353) / 10,353** |

Figure 1: An example of TAT-QA. The left dashed line box shows a *hybrid context*. The rows with blue background are *row header* while the column with grey is *column header*. The right solid line box shows corresponding question, answer with its scale, and derivation to arrive at the answer.

**Financial report:**

… the total income tax benefit recognized for share-based compensation in the accompanying statements of income is also presented.

| | 2010 | 2009 | 2008 |
|---|---|---|---|
| share-based compensation cost | $18.10 | $14.60 | $13.80 |
| income tax benefit | -$6.30 | -$5.20 | -$4.90 |

**Conversational QA:**

**Q1:** In the year of 2010, what was the share-based compensation cost?
**A1:** 18.1

**Q2:** and what was the income tax benefit?
**A2:** -6.3

**Q3:** what was, then, the sum of both?
**A3:** **add**(18.1, -6.3) = 11.8

**Q4:** and what was that sum in 2009?
**A4:** **add**(14.6, -5.2) = 9.4

**Q5:** what, then, was the change in the sum of those amounts from 2009 to 2010?
**A5:** **add**(18.1, -6.3), **add**(14.6, -5.2), **subtract**(#0, #1) = 2.4

**ConvFinQA**
Conversation dialogue version

Questions similar FinQA

**MultiHiertt**
Hierarchical Tabular

**Document:**

( ... abbreviate... )

The following table presents product and service sales and operating expenses by segment (dollar in millions):

hierarchical column headers

hierarchical row headers

| | Year Ended December 31 | | | |
|---|---|---|---|---|
| | 2018 | | 2017 | |
| Segment | Sales | Expenses | Sales | Expenses |
| Innovation Systems | | | | |
| Product | 2,894 | 2,582 | — | — |
| Service | 382 | 351 | — | — |
| Aerospace Systems | | | | |
| Product | 11,087 | 9,889 | 10,064 | 8,988 |
| Service | 2,009 | 1,796 | 2,067 | 1,854 |
| Mission Systems | | | | |
| Product | 7,329 | 6,335 | 7,012 | 6,088 |
| Service | 4,380 | 3,854 | 4,458 | 3,940 |
| Technology Service | | | | |
| Product | 485 | 450 | 391 | 360 |
| Service | 3,812 | 3,404 | 4,296 | 3,878 |

Product sales for 2018 increased $4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of $2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

( ... abbreviate... )

The table below reconciles funds provided to each segment (dollar in millions):

| | 2018 | 2017 | |
|---|---|---|---|
| Segment | Funded | Funded | % Change |
| Innovation Systems | 5,928 | — | — |
| Aerospace Systems | 11,448 | 9,560 | 19.7 % |
| Mission Systems | 9,676 | 9,277 | 4.3 % |
| Technology Services | 2,883 | 2,792 | 3.3 % |

Approximately $26.6 billion of the $53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. ( ... abbreviate... )

**Question:** *In 2018, what was the total sales increase in the segment with most funds in 2017?*

**Answer:** *965*

**Numerical expression:** *(11087 – 10064) + (2009 – 2067)*

# High Level Motivation

## Why semi-structured tabular information?

1. Structured Data - a challenge
2. Abundant Numerical Information
3. Diverse Reasoning
4. Multiple Diverse TQA Datasets

### *Why Probing LLM for Numerical Reasoning?*

1. Prior works show that LLM struggle on numerical aspects
2. Only broadly studied, several fine-grained aspect not studies

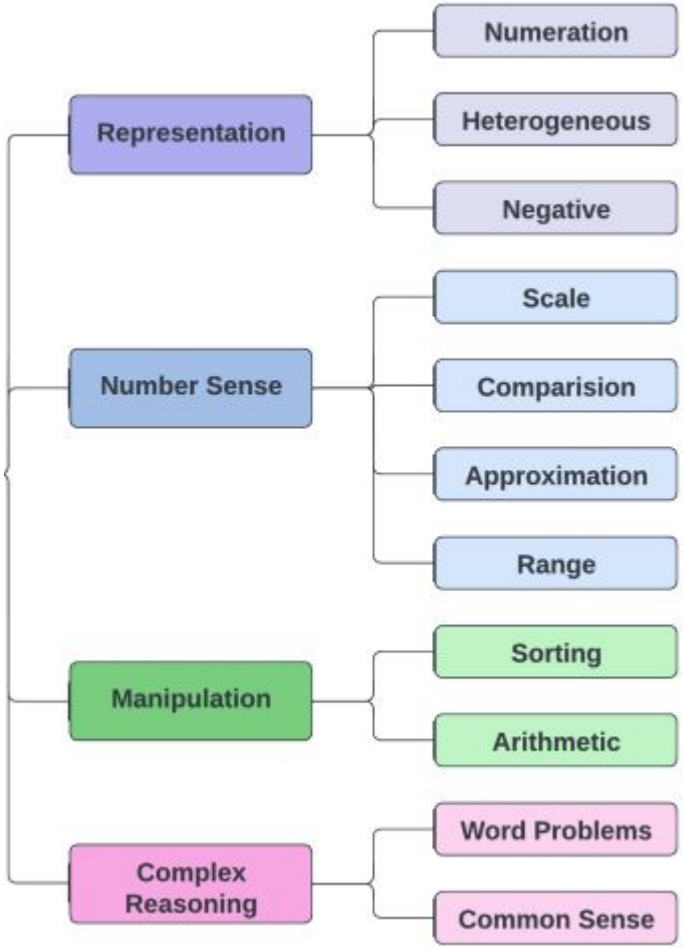



Figure from : Mubashara et.al. Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data, EMNLP 2023

# Our Contributions

1. **Evaluation** of LLM in tabular question answering **for mathematical reasoning tasks** using open-source **financial tabular datasets.**

   We **annotated metadata of reasoning steps (test/dev sets) for four TQA datasets.**

# Our Contributions

1. **Evaluation** of LLM in tabular question answering **for mathematical reasoning tasks** using open-source **financial tabular datasets.**

   We **annotated metadata of reasoning steps (test/dev sets) for four TQA datasets.**

2. Our analysis is **thorough and multifaceted**, encompassing **both qualitative and quantitative aspects**.

   Insights into the **strengths and limitations of LLMs** on tabular data (QA tasks), especially in scenarios **involving mathematical reasoning.**

# Our Contributions

1. **Evaluation** of LLM in tabular question answering **for mathematical reasoning tasks** using open-source **financial tabular datasets.**

   We **annotated metadata of reasoning steps (test/dev sets) for four TQA datasets.**

2. Our analysis is **thorough and multifaceted**, encompassing **both qualitative and quantitative aspects**.

   Insights into the **strengths and limitations of LLMs** on tabular data (QA tasks), especially in scenarios **involving mathematical reasoning.**

3. Building upon qualitative analysis, we introduce a **novel unified multi-step prompting method** which promote evidence extraction and correct reasoning namely ***EEDP***.

   Our proposed prompt work well across complex data structures such as hybrid table and text, as well as multi-hierarchical tables.

# Our Probing Criteria

1. **#Reasoning Steps:** More arithmetic operations involved mean increased complexity.

**Financial report:**

… the total income tax benefit recognized for share-based compensation in the accompanying statements of income is also presented.

| | 2010 | 2009 | 2008 |
|---|---|---|---|
| share-based compensation cost | $18.10 | $14.60 | $13.80 |
| income tax benefit | -$6.30 | -$5.20 | -$4.90 |

**Conversational QA:**

**Q1:** In the year of 2010, what was the share-based compensation cost?
**A1:** 18.1

**Q2:** and what was the income tax benefit?
**A2:** -6.3

**Q3:** what was, then, the sum of both?
**A3:** **add**(18.1, -6.3) = 11.8

**Q4:** and what was that sum in 2009?
**A4:** **add**(14.6, -5.2) = 9.4

**Q5:** what, then, was the change in the sum of those amounts from 2009 to 2010?
**A5:** **add**(18.1, -6.3), **add**(14.6, -5.2), **subtract**(#0, #1) = 2.4

Figure 1: An example from CONVFINQA: each question may depend on previous questions to answer.

# Our Probing Criteria

1. **#Reasoning Steps:** More arithmetic operations involved mean increased complexity.

2. **Question Types:** Questions classification based on operations combinations and linguistic complexity.

**Financial report:**

... the total income tax benefit recognized for share-based compensation in the accompanying statements of income is also presented.

| | 2010 | 2009 | 2008 |
|---|---|---|---|
| share-based compensation cost | $18.10 | $14.60 | $13.80 |
| income tax benefit | -$6.30 | -$5.20 | -$4.90 |

**Conversational QA:**

**Q1:** In the year of 2010, what was the share-based compensation cost?
**A1:** 18.1

**Q2:** and what was the income tax benefit?
**A2:** -6.3

**Q3:** what was, then, the sum of both?
**A3:** **add**(18.1, -6.3) = 11.8

**Q4:** and what was that sum in 2009?
**A4:** **add**(14.6, -5.2) = 9.4

**Q5:** what, then, was the change in the sum of those amounts from 2009 to 2010?
**A5:** **add**(18.1, -6.3), **add**(14.6, -5.2), **subtract**(#0, #1) = 2.4

Figure 1: An example from CONVFINQA: each question may depend on previous questions to answer.

| | |
|---|---|
| SUM | Questions that implicitly or explicitly require ONLY the knowledge of Addition. |
| DIFF | Questions that implicitly or explicitly require ONLY the knowledge of Subtraction. |
| PRODUCT | Questions that implicitly or explicitly require ONLY the knowledge of Multiplication. |
| DIVISION | Questions that implicitly or explicitly require ONLY the knowledge of Division. |
| RATIO | Questions that implicitly or explicitly require knowing fractional forms, for example: Percentages, Ratios. |
| CHANGE RATIO | Questions involving the difference between two fractional forms, e.g., Percentages changes, Difference in Ratios. |
| RANGE | Questions requiring knowledge of the minimum and maximum of data observations. |
| COMPARE | Questions necessitating a comparison between mathematical quantities, e.g., greater than, less than. |
| AVERAGE | Questions needing knowledge of the average, used to calculate the central tendency of a group of data points. |
| IN-DOMAIN-INFO | Questions that require implicit knowledge to understand domain-specific mathematical formulations, for example: Return on Investment (RoI), Cost of Goods sold (COGS), Amortization Rate, etc. |
| TIME | Questions explicitly involving mathematical operators for time-spans not in the table or context. |
| COUNTING | Questions requiring the counting of elements in a set or group of data points. |

3. **#Rows:** Larger tables pose challenges for **multi-hop reasoning tasks**

increase information → harder to **extract relevant (right) evidence**

**Document:**

( ... abbreviate... )

The following table presents product and service sales and operating expenses by segment (dollar in millions):

hierarchical column headers

hierarchical row headers

| | Year Ended December 31 | | | |
|---|---|---|---|---|
| | 2018 | | 2017 | |
| **Segment** | **Sales** | **Expenses** | **Sales** | **Expenses** |
| **Innovation Systems** | | | | |
| Product | 2,894 | 2,582 | — | — |
| Service | 382 | 351 | — | — |
| **Aerospace Systems** | | | | |
| Product | 11,087 | 9,889 | 10,064 | 8,988 |
| Service | 2,009 | 1,796 | 2,067 | 1,854 |
| **Mission Systems** | | | | |
| Product | 7,329 | 6,335 | 7,012 | 6,088 |
| Service | 4,380 | 3,854 | 4,458 | 3,940 |
| **Technology Service** | | | | |
| Product | 485 | 450 | 391 | 360 |
| Service | 3,812 | 3,404 | 4,296 | 3,878 |

Product sales for 2018 increased $4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of $2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

( ... abbreviate... )

The table below reconciles funds provided to each segment (dollar in millions):

| | 2018 | 2017 | |
|---|---|---|---|
| **Segment** | **Funded** | **Funded** | **% Change** |
| Innovation Systems | 5,928 | — | — |
| Aerospace Systems | 11,448 | 9,560 | 19.7 % |
| Mission Systems | 9,676 | 9,277 | 4.3 % |
| Technology Services | 2,883 | 2,792 | 3.3 % |

Approximately $26.6 billion of the $53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. ( ... abbreviate... )

**Question:** *In 2018, what was the total sales increase in the segment with most funds in 2017?*

**Answer:** *965*

**Numerical expression:** *(11087 – 10064) + (2009 – 2067)*

3. **#Rows:** Larger tables pose challenges for **multi-hop reasoning tasks**

increase information → harder to **extract relevant (right) evidence**

4. **Hierarchical Depth:** As the hierarchy depth increases,

**understand the information structure** and **connections** in **important data cells is challenging**.

**Document:**

( ... abbreviate... )

The following table presents product and service sales and operating expenses by segment (dollar in millions):

hierarchical column headers

hierarchical row headers

| | Year Ended December 31 | | | |
|---|---|---|---|---|
| | 2018 | | 2017 | |
| **Segment** | **Sales** | **Expenses** | **Sales** | **Expenses** |
| **Innovation Systems** | | | | |
| Product | 2,894 | 2,582 | — | — |
| Service | 382 | 351 | — | — |
| **Aerospace Systems** | | | | |
| Product | 11,087 | 9,889 | 10,064 | 8,988 |
| Service | 2,009 | 1,796 | 2,067 | 1,854 |
| **Mission Systems** | | | | |
| Product | 7,329 | 6,335 | 7,012 | 6,088 |
| Service | 4,380 | 3,854 | 4,458 | 3,940 |
| **Technology Service** | | | | |
| Product | 485 | 450 | 391 | 360 |
| Service | 3,812 | 3,404 | 4,296 | 3,878 |

Product sales for 2018 increased $4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of $2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

( ... abbreviate... )

The table below reconciles funds provided to each segment (dollar in millions):

| | 2018 | 2017 | |
|---|---|---|---|
| **Segment** | **Funded** | **Funded** | **% Change** |
| Innovation Systems | 5,928 | — | — |
| Aerospace Systems | 11,448 | 9,560 | 19.7 % |
| Mission Systems | 9,676 | 9,277 | 4.3 % |
| Technology Services | 2,883 | 2,792 | 3.3 % |

Approximately $26.6 billion of the $53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. ( ... abbreviate... )

**Question:** *In 2018, what was the total sales increase in the segment with most funds in 2017?*

**Answer:** *965*

**Numerical expression:** *(11087 – 10064) + (2009 – 2067)*

3. **#Rows:** Larger tables pose challenges for **multi-hop reasoning tasks**

increase information → harder to **extract relevant (right) evidence**

4. **Hierarchical Depth:** As the hierarchy depth increases,

**understand the information structure** and **connections** in **important data cells is challenging**.

5. **%Empty Cells:** More empty cells in a dataset indicate **increased data ambiguity**, making it challenging to derive meaningful insights

**Document:**

( ... abbreviate... )

The following table presents product and service sales and operating expenses by segment (dollar in millions):

hierarchical column headers

hierarchical row headers

| | Year Ended December 31 | | | |
|---|---|---|---|---|
| | 2018 | | 2017 | |
| **Segment** | **Sales** | **Expenses** | **Sales** | **Expenses** |
| **Innovation Systems** | | | | |
| Product | 2,894 | 2,582 | — | — |
| Service | 382 | 351 | — | — |
| **Aerospace Systems** | | | | |
| Product | 11,087 | 9,889 | 10,064 | 8,988 |
| Service | 2,009 | 1,796 | 2,067 | 1,854 |
| **Mission Systems** | | | | |
| Product | 7,329 | 6,335 | 7,012 | 6,088 |
| Service | 4,380 | 3,854 | 4,458 | 3,940 |
| **Technology Service** | | | | |
| Product | 485 | 450 | 391 | 360 |
| Service | 3,812 | 3,404 | 4,296 | 3,878 |

Product sales for 2018 increased $4.3 billion, or 25 percent, as compared with 2017. The increase was primarily due to the addition of $2.9 billion of product sales from Innovation Systems and higher restricted and F-35 volume at Aerospace Systems.

( ... abbreviate... )

The table below reconciles funds provided to each segment (dollar in millions):

| | 2018 | 2017 | |
|---|---|---|---|
| **Segment** | **Funded** | **Funded** | **% Change** |
| Innovation Systems | 5,928 | — | — |
| Aerospace Systems | 11,448 | 9,560 | 19.7 % |
| Mission Systems | 9,676 | 9,277 | 4.3 % |
| Technology Services | 2,883 | 2,792 | 3.3 % |

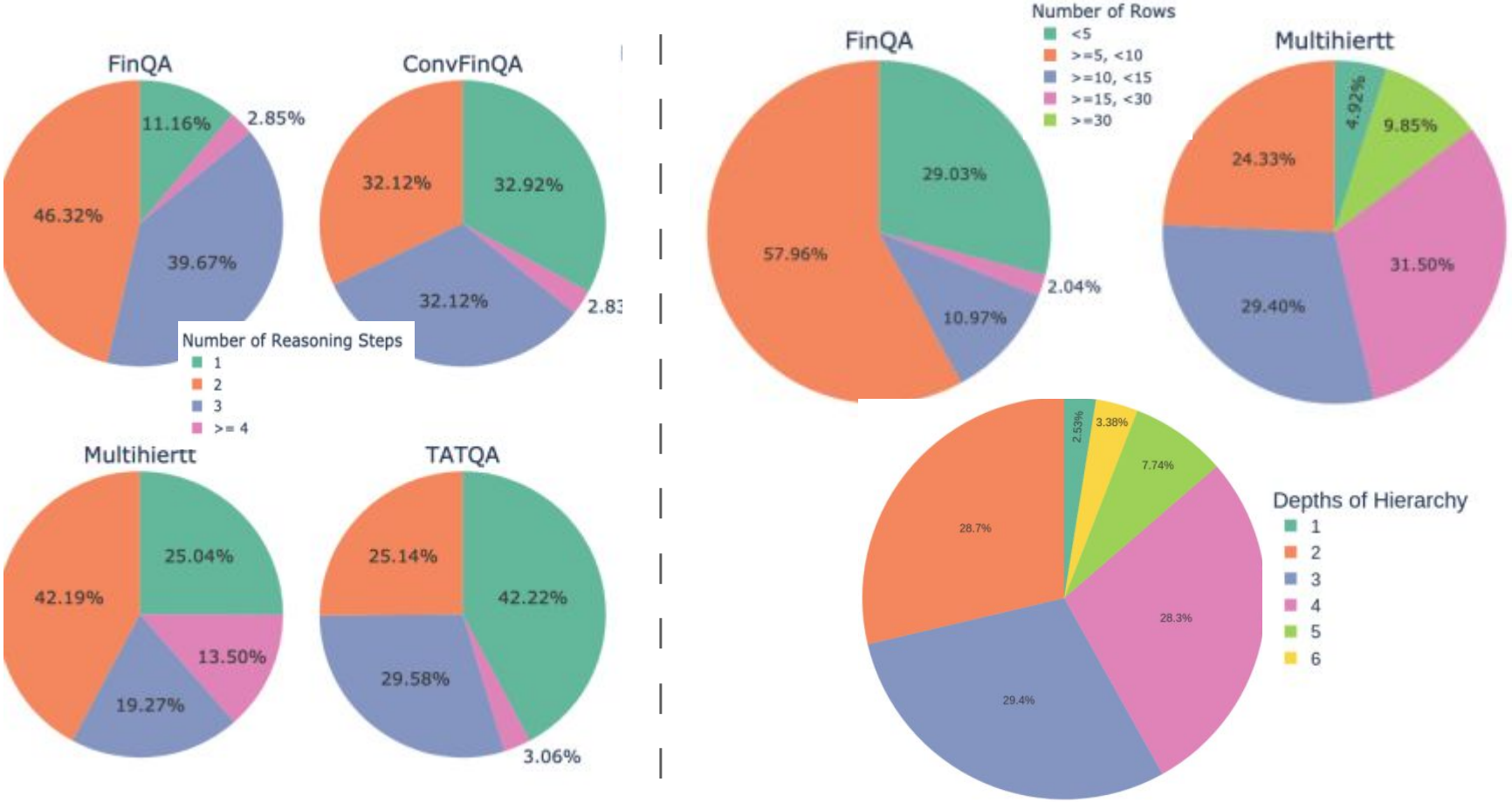
Approximately $26.6 billion of the $53.5 billion total at December 31, 2018 is expected to be converted into sales in 2019. ( ... abbreviate... )

**Question:** *In 2018, what was the total sales increase in the segment with most funds in 2017?*

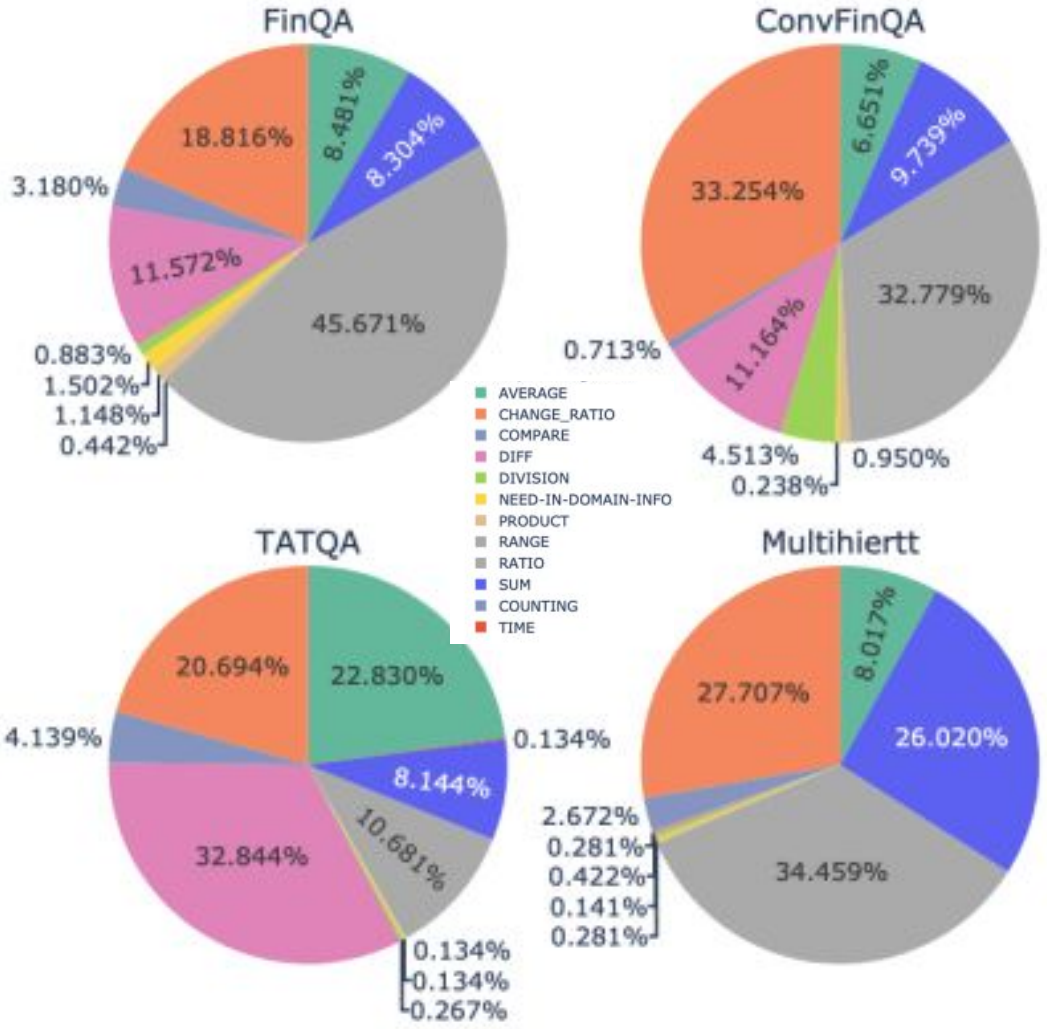
**Answer:** *965*

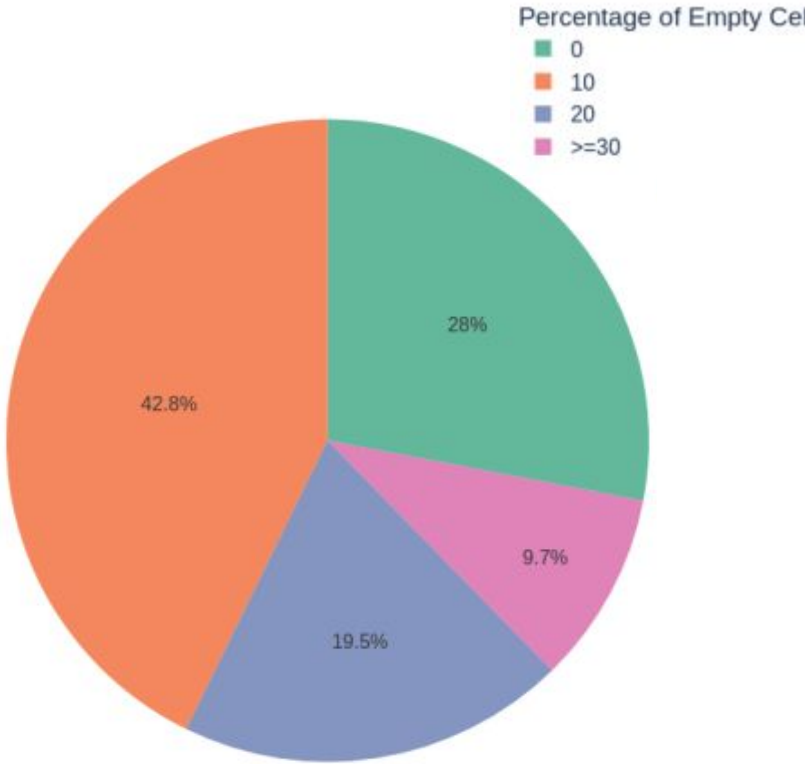
**Numerical expression:** *(11087 – 10064) + (2009 – 2067)*

# MetaData Annotations: Reasoning Steps, #Rows



# QType & Empty Cell



# LLM Performance - API models

| Dataset (Split) | Model | 0-Direct | 0-CoT | k-Direct | k-CoT |
|---|---|---|---|---|---|
| FinQA (full test set) | GPT-4 | 70.18 | 75.49 | 72.74 | 74.51 |
| | GPT-3.5-TURBO | 43.36 | 50.71 | 47.79 | 55.75 |
| | PaLM 2-540B | 29.91 | 33.01 | 29.38 | 33.2 |
| ConvFinQA (full dev set) | GPT-4 | 48.46 | 66.51 | 65.08 | 75.53 |
| | GPT-3.5-TURBO | 34.92 | 42.99 | 38.95 | 56.53 |
| | PaLM 2-540B | 16.74 | 26.16 | 18.99 | 21.94 |
| Multihiertt (68% of test set) | GPT-4 | 38.53 | 61.6 | 45 | 64.41 |
| | GPT-3.5-TURBO | 23.21 | 39.52 | 23.07 | 43.18 |
| | PaLM 2-540B | 16.74 | 26.16 | 18.99 | 21.94 |
| TATQA (45% of dev set) | GPT-4 | 78.67 | 80.4 | 83.2 | 86.4 |
| | GPT-3.5-TURBO | 40.13 | 63.2 | 62.93 | 76.53 |
| | PaLM 2-540B | 46.27 | 47.33 | 49.87 | 57.07 |

# Types of Errors

1. ***Extraction:*** Highlights the difficulties in **correctly identifying relevant pieces** of evidence.

**examples in the paper*

# Types of Errors

1. ***Extraction:*** Highlights the difficulties in **correctly identifying relevant pieces** of evidence.

2. ***Reasoning:*** Model struggles to formulate **a suitable and contextually relevant approach/technique/steps**.

**examples in the paper*

# Types of Errors

1. ***Extraction:*** Highlights the difficulties in **correctly identifying relevant pieces** of evidence.

2. ***Reasoning:*** Model struggles to formulate **a suitable and contextually relevant approach/technique/steps**.

3. ***Calculation:*** Emphasizes the inherent constraints of LLMs when it comes to **performing precise calculations.**

**examples in the papers*

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

***E****licit* ➜ ***E****xtract* ➜ ***D****ecompose* ➜ ***P****redict*

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

***E****licit*

1. Start ***Eliciting***: Prompt the model to bring up key domain knowledge relevant to the question.

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

***E**licit* ➜ ***E**xtract*

1. Start ***Eliciting***: Prompt the model to bring up key domain knowledge relevant to the question.

2. Next, ***Extract***: Gather evidence while considering the relevant domain knowledge required for answering the question.

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

***E**licit* ➜ ***E**xtract* ➜ ***D**ecompose*

1. Start ***Eliciting***: Prompt the model to bring up key domain knowledge relevant to the question.

2. Next, ***Extract***: Gather evidence while considering the relevant domain knowledge required for answering the question.

3. Then, ***Decompose***: Break down a complex math problem into much smaller and more simpler steps.

# Mitigation: Our EEDP Prompt

New prompting method: **Simplifies complex problem** by **breaking** them into **smaller, multiples, manageable steps,** making tough tasks easier to solve.

***E**licit* ➜ ***E**xtract* ➜ ***D**ecompose* ➜ ***P**redict*

1. Start ***Eliciting***: Prompt the model to bring up key domain knowledge relevant to the question.

2. Next, ***Extract***: Gather evidence while considering the relevant domain knowledge required for answering the question.

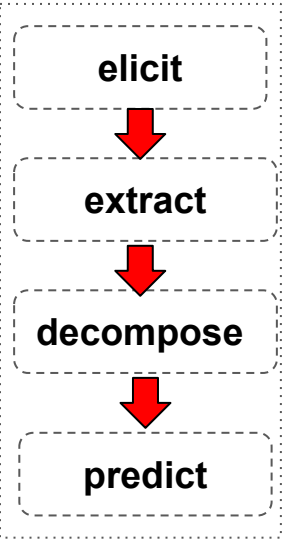
3. Then, ***Decompose***: Break down a complex math problem into much smaller and more simpler steps.

4. Finally, ***Predict***: Solve each small step one by one by performing the necessary calculation to reach the overall solution.

# Actual Prompt: <Instructions>

Please ***carefully analyze*** the provided information, perform ***necessary numerical calculations*****, and** ***accurate answers*** to the given question using the provided data.

Take into ***account the pre-text, table, and post-text*** when formulating your response. When performing numerical calculations, ensure you *gather the* ***required information and follow a step-by-step approach*****.**




***Elicit*** ***the most*** ***relevant domain fact or knowledge*** that might be useful for you to extract the right operands and **forming the right approach to the problem.**

***Translate*** ***table data into sentences*** ***and*** ***identify gold evidence*** for answering the question.

Define a **sequence of** ***atomic operations*** *(Add, Subtract, Divide, Multiply, Greater, Min, Max, Exp etc.)* which take into **account only one or two operands** at a time.

Divide a ***complex task into a sequence of atomic operations*** defined above. Finally, conclude as : ***The final answer is <answer>.***

Here are a few demonstrations of the task: ***<Demonstration>***

# Actual Prompt: <Demonstration>

Effective Income Tax Rate: A reconciliation of the United States federal statutory income tax rate to our effective income tax rate is as follows:

| Description | Dec 31, 2018 (%) | Dec 31, 2019 (%) |
|---|---|---|
| US Federal Statutory Rate | 21.00 | 21.00 |
| Effective Income Tax Rate | 1.99 | -0.01 |

In 2019 and 2018 we had pre-tax losses of $19,573 and $25,403, respectively, which are available for carry forward to offset future taxable income.

# Actual Prompt: <Demonstration>

Effective Income Tax Rate: A reconciliation of the United States federal statutory income tax rate to our effective income tax rate is as follows:

| Description | Dec 31, 2018 (%) | Dec 31, 2019 (%) |
|---|---|---|
| US Federal Statutory Rate | 21.00 | 21.00 |
| Effective Income Tax Rate | 1.99 | -0.01 |

In 2019 and 2018 we had pre-tax losses of $19,573 and $25,403, respectively, which are available for carry forward to offset future taxable income.

***Question:*** What was the 2019 percentage change in pre-tax losses?

**Output:**

elicit

**Domain Knowledge:** Pre-tax losses, or operating losses, refer to financial losses that a company incurs before considering the effects of income taxes.

To find the 2019 percentage change in pre-tax losses, we need to find the difference between the new and the old value of the pre-tax losses, then divide the obtained difference by the old value and multiply this value by 100.

extract

**Gold Evidences:** The **pre-tax losses** in **2019 is $19,573.** The pre-tax losses in **2018 is $25,403.**

**Steps:**

*Step 1:* Subtract 25403 from 19573. → *Response 1:* **19573 - 25403 = -5830**

*Step 2:* Divide #1 by 25403. → *Response 2:* **\frac{#1}{25403} = -0.2295**

*Step 3:* Multiply #2 by 100. → *Response 3:* **#2 * 100 = - 22.95**

decompose

predict

The **final answer** is -**22.95**

# LLM Performance - API models

**EEDP outperform other prompting techniques**

| Dataset (Split) | Model | k-CoT | *k-EEDP (ours)* |
|---|---|---|---|
| FinQA (full test set) | GPT-4 | 74.51 | **78.41** |
| | GPT-3.5-TURBO | 55.75 | **59.47** |
| | PaLM 2-540B | 33.2 | **43.19** |
| ConvFinQA (full dev set) | GPT-4 | 75.53 | **76.96** |
| | GPT-3.5-TURBO | 56.53 | **58.19** |
| | PaLM 2-540B | 21.94 | **35.72** |
| Multihiertt (68% of test set) | GPT-4 | 64.41 | **67.79** |
| | GPT-3.5-TURBO | 43.18 | **46.13** |
| | PaLM 2-540B | 21.94 | **35.72** |
| TATQA (45% of dev set) | GPT-4 | 86.4 | **88.13** |
| | GPT-3.5-TURBO | 76.53 | **79.06** |
| | PaLM 2-540B | 57.07 | **67.87** |

Table 2: Comparison of accuracy across LLMs and prompting strategies for datasets assessing numerical reasoning over semi-structured data. Note that 0-Direct implies zero-shot direct prompting, while $k$-CoT implies few-shot CoT prompting.

# Open-source

- Open models, including LLAMA-70B (+later), underperform due to size constraints.

| Dataset (Split) | Model | 0-Direct | 0-CoT | k-Direct | k-CoT | *k-EEDP (ours)* |
|---|---|---|---|---|---|---|
| FinQA (full test set) | GPT-4 | 70.18 | 75.49 | 72.74 | 74.51 | **78.41** |
| | GPT-3.5-TURBO | 43.36 | 50.71 | 47.79 | 55.75 | **59.47** |
| | PaLM 2-540B | 29.91 | 33.01 | 29.38 | 33.2 | **43.19** |
| | LLaMA-2-13B | 16.2 | 18.1 | 14.9 | 12.9 | **38.13** |
| | MAmmoTH-13B | 15.24 | 19.96 | 10.68 | **23.29** | 20.58 |
| | Mistral-7B-Instruct | 19 | **23.37** | 13.16 | 18.83 | 12.81 |

Table 2: Comparison of accuracy across LLMs and prompting strategies for datasets assessing numerical reasoning over semi-structured data. Note that 0-Direct implies zero-shot direct prompting, while $k$-CoT implies few-shot CoT prompting.

# Open-source Models

- Open models, including LLAMA-70B (+later), underperform due to size constraints.
- Numerically specialized models like MAmmoTH-13B & Mistral-7B-Instruct also fall much behind.

| Dataset (Split) | Model | 0-Direct | 0-CoT | k-Direct | k-CoT | *k-EEDP (ours)* |
|---|---|---|---|---|---|---|
| FinQA (full test set) | GPT-4 | 70.18 | 75.49 | 72.74 | 74.51 | **78.41** |
| | GPT-3.5-TURBO | 43.36 | 50.71 | 47.79 | 55.75 | **59.47** |
| | PaLM 2-540B | 29.91 | 33.01 | 29.38 | 33.2 | **43.19** |
| | LLaMA-2-13B | 16.2 | 18.1 | 14.9 | 12.9 | **38.13** |
| | MAmmoTH-13B | 15.24 | 19.96 | 10.68 | **23.29** | 20.58 |
| | Mistral-7B-Instruct | 19 | **23.37** | 13.16 | 18.83 | 12.81 |

Table 2: Comparison of accuracy across LLMs and prompting strategies for datasets assessing numerical reasoning over semi-structured data. Note that 0-Direct implies zero-shot direct prompting, while $k$-CoT implies few-shot CoT prompting.

# Open-source Models

- Open models, including LLAMA-70B (+later), underperform due to size constraints.
- Numerically specialized models like MAmmoTH-13B & Mistral-7B-Instruct also fall much behind.
- Consistent underperformance observed across four datasets: FinQA, ConvFinQA, MultiHiertt, TAT-QA.
- EEDP ~= CoT in performance on most sets

| Dataset (Split) | Model | 0-Direct | 0-CoT | k-Direct | k-CoT | *k-EEDP (ours)* |
|---|---|---|---|---|---|---|
| FinQA (full test set) | GPT-4 | 70.18 | 75.49 | 72.74 | 74.51 | **78.41** |
| | GPT-3.5-TURBO | 43.36 | 50.71 | 47.79 | 55.75 | **59.47** |
| | PaLM 2-540B | 29.91 | 33.01 | 29.38 | 33.2 | **43.19** |
| | LLaMA-2-13B | 16.2 | 18.1 | 14.9 | 12.9 | **38.13** |
| | MAmmoTH-13B | 15.24 | 19.96 | 10.68 | **23.29** | 20.58 |
| | Mistral-7B-Instruct | 19 | **23.37** | 13.16 | 18.83 | 12.81 |

Table 2: Comparison of accuracy across LLMs and prompting strategies for datasets assessing numerical reasoning over semi-structured data. Note that 0-Direct implies zero-shot direct prompting, while $k$-CoT implies few-shot CoT prompting.

# Qualitative Examples - Elicit

Question : What is the five-year total return on the Goldman Sachs group inc.?

| Table 1 | 12/26/08 | 12/31/09 | 12/31/10 | 12/31/11 | 12/31/12 | 12/31/13 |
|---|---|---|---|---|---|---|
| The Goldman Sachs Group Inc. | $100.00 | $224.98 | $226.19 | $123.05 | $176.42 | $248.36 |
| S&P 500 Index | $100.00 | $130.93 | $150.65 | $153.83 | $178.42 | $236.20 |
| S&P 500 Financials Index | $100.00 | $124.38 | $139.47 | $115.67 | $148.92 | $201.92 |

**Few-shot CoT Response**

*The five year total return* on the Goldman Sachs Group Inc. is 248.36.

The answer is 248.36. ✗

**EEDP Response (*Ours*)**

***Domain Knowledge:***
*The five year total return on the goldman sachs group inc. is the difference between the value of the stock at the end of the five year period and the value of the stock at the beginning of the five year period.*

***Gold Evidences:***
the goldman sachs group inc . | $ 100.00 | $ 224.98 | $ 226.19 | $ 123.05 | $ 176.42 | $ 248.36

Step 1: Find the value of the stock at the end of the five year period
Response 1: $248.36
Step 2: Find the value of the stock at the beginning of the five year period
Response 2: $100.00
Step 3: Find the difference between the values obtained in #1 and #2
Response 3: #1 - #2 = 148.36

The final answer is 148.36. ✓

EEDP

**Elicit the require knowledge and reasoning**

For *Debt Maturities*,

By strategically mining domain knowledge about Return

unlike the few-shot CoT prompting.

# Qualitative Examples - Extract and Reasoning

Question : What's the sum of Debt maturities of Thereafter, and Capital lease obligations of Less than 1 year ?

| Contractual Obligation Categories | Total | 2013 | 2014 to 2015 | 2016 to 2017 | Thereafter |
|---|---|---|---|---|---|
| Reclamation and environmental obligations | $5,243 | $246 | $471 | $329 | $4,197 |
| Debt maturities | $3,527 | $2 | $500 | $500 | $2,525 |
| Take-or-pay contracts | $2,200 | $976 | $731 | $286 | $207 |
| Scheduled interest payment obligations | $1,289 | $121 | $241 | $226 | $701 |
| Operating lease obligations | $205 | $32 | $38 | $31 | $104 |
| **Total** | **$12,464** | **$1,377** | **$1,981** | **$1,372** | **$7,734** |

| Contractual Obligation Categories | Less than 1 year | 1-3 years | 3-5 years | More than 5 years | Total |
|---|---|---|---|---|---|
| Operating lease obligations | $8,765 | $9,422 | $5,851 | $3,190 | $27,228 |
| Capital lease obligations | $4,380 | $1,309 | - | - | $5,689 |
| Notes payable, including accrued interest | $102,493 | $46,210 | $225,213 | - | $373,916 |
| Purchase obligations | $9,707 | - | - | - | $9,707 |
| **Total** | **$125,345** | **$56,941** | **$231,064** | $3,190 | **$416,540** |

### Few-shot CoT Response

The sum of *Debt maturities of Thereafter*, and *Capital lease obligations of Less than 1 year* is 500 + 4380 = 4880.

The answer is 4880. ✗

### EEDP Response (*Ours*)

***Domain Knowledge:***
To find the sum of Debt maturities of Thereafter, and Capital lease obligations of Less than 1 year, we need to add the two values.

***Gold Evidences:***
Table shows *Debt maturities of Thereafter* is 2,525 .
Table shows *Capital lease obligations of Less than 1 year* is 4,380 .

Step 1: Add 2,525 and 4,380.
Response 1: 2,525 + 4,380 = 6,905

The final answer is 6,905 ✓

EEDP **rightly extract evidence and reason**

unlike few-shot CoT, which suffers from lack of domain knowledge.

# Qualitative Examples - Decomposition

Question : What was the percentage change in revenue from external customers from UK in 2019 from 2018?

6 Segment Information continued The Group's revenue is diversified across its entire end customer base and no single end user accounted for greater than 10 per cent of the Group's revenue in either 2018 or 2019. In 2019 two distributors accounted for 15 per cent each, and one distributor for 11 per cent of Group billings which were attributable to all segments of the Group (2018: three distributors accounted for 15 per cent, 14 per cent and 12 per cent each).

| Revenue from external customers by country | Year-ended 31 March 2019 | Year-ended 31 March 2018 |
|---|---|---|
| UK | 83.2 | 73.5 |
| USA | 222.2 | 199.0 |
| Germany | 143.5 | 128.4 |
| Other countries | 261.7 | 238.1 |
| Total revenue from external customers | 710.6 | 639.0 |

### Few-shot CoT Response

The revenue from external customers from UK in 2019 is 83.2 million and the revenue from external customers from UK in 2018 is 73.5 million. The percentage change is 13.5%.

The answer is 13.5% ✗

### EEDP Response (*Ours*)

***Domain Knowledge***: To find the percentage change in revenue from external customers from UK in 2019 from 2018, we need to subtract the old value (in 2018) of the revenue from the new one (in 2019), then divide the obtained difference by the old value and multiply this value by 100.

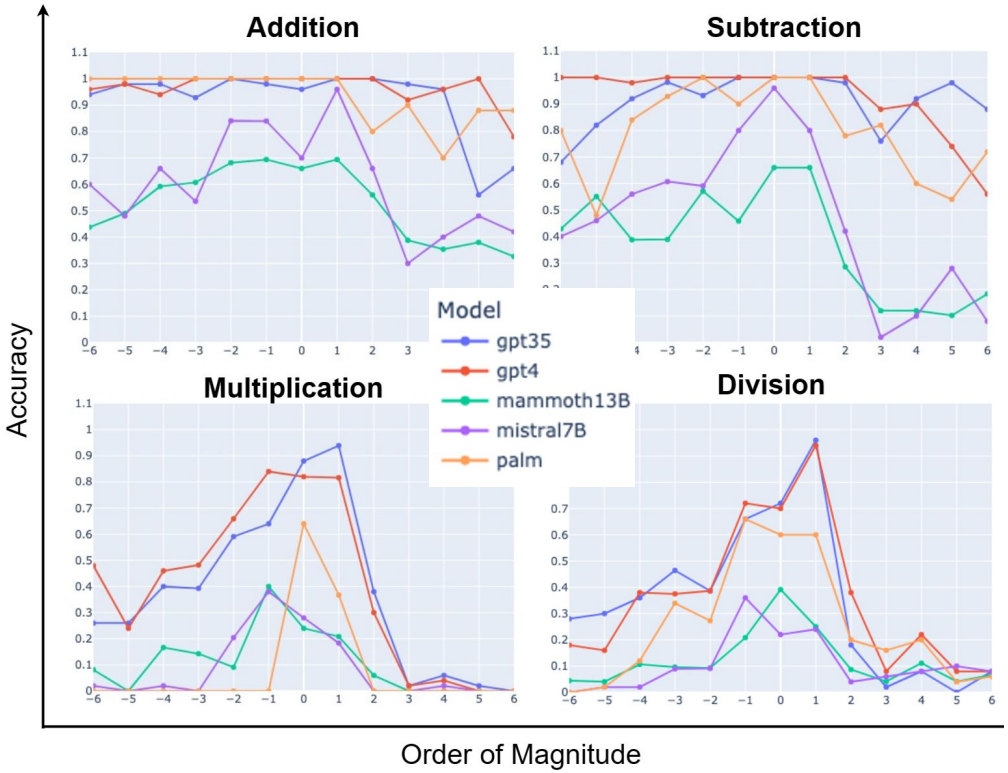
***Gold Evidences***:
The revenue from external customers from UK in 2019 is $83.2.
The revenue from external customers from UK in 2018 is $73.5.

Step 1: Subtract 73.5 from 83.2.
Response 1: 83.2 - 73.5 = 9.7
Step 2: Divide #1 by 73.5.
Response 2: #1 / 73.5 = 0.132
Step 3: Multiply #2 by 100.
Response 3: #2 * 100 = 13.2

The final answer is 13.2% ✓

EEDP performs **precise calculations** because of fine-grained simpler steps better than few-shot CoT.

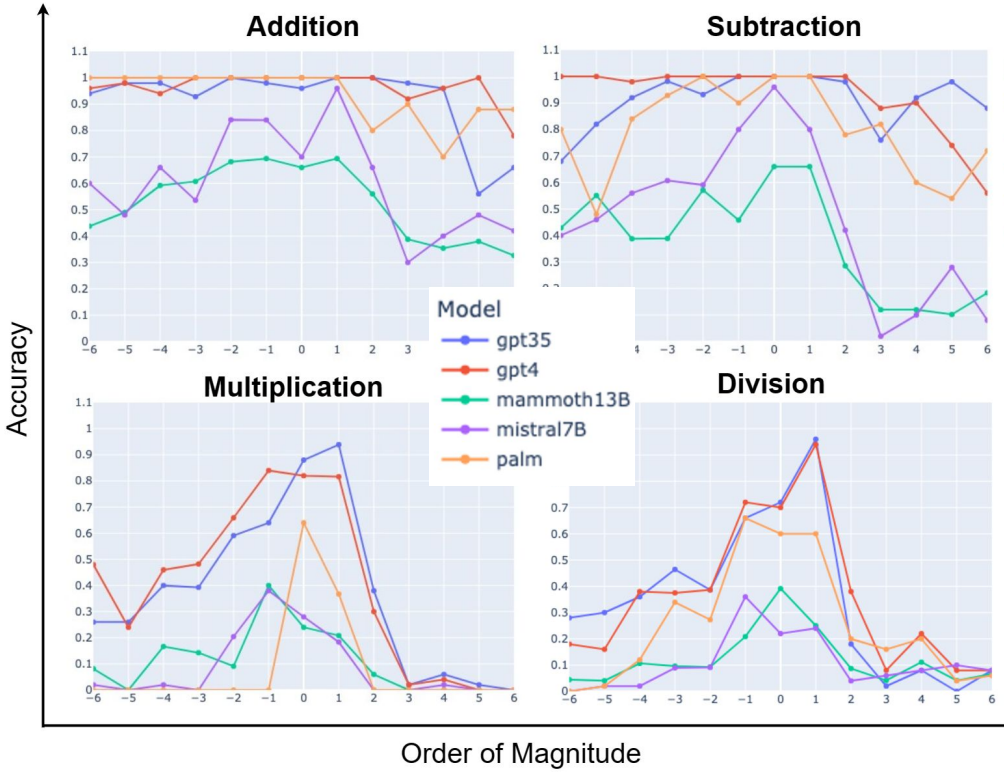
# Scaling of Arithmetic Operations




10^-5 to 10^5, power of 10s

Simple arithmetic operations, no multi-step hard reasoning

# Scaling of Arithmetic Operations



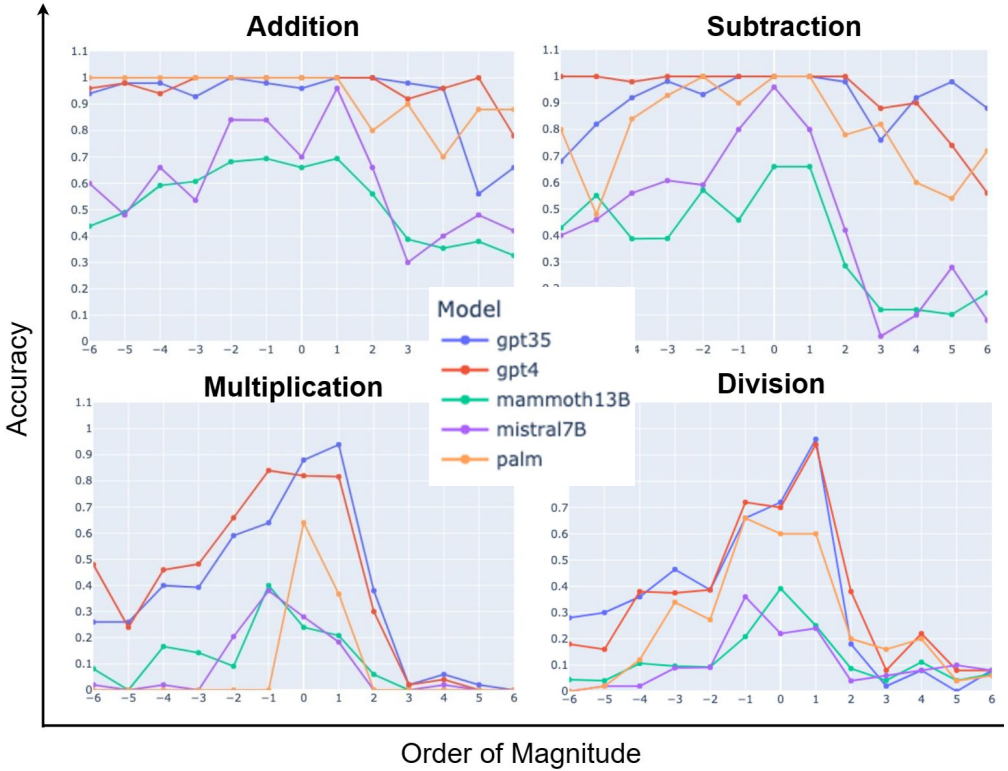

10^-5 to 10^5, power of 10s

Simple arithmetic operations, no multi-step hard reasoning

**Hardness:**

Div(“/”) = Multi(“X”) > Sub (“-”) = Add (“+”)

# Scaling of Arithmetic Operations




10^-5 to 10^5, power of 10s

Simple arithmetic operations, no multi-step hard reasoning

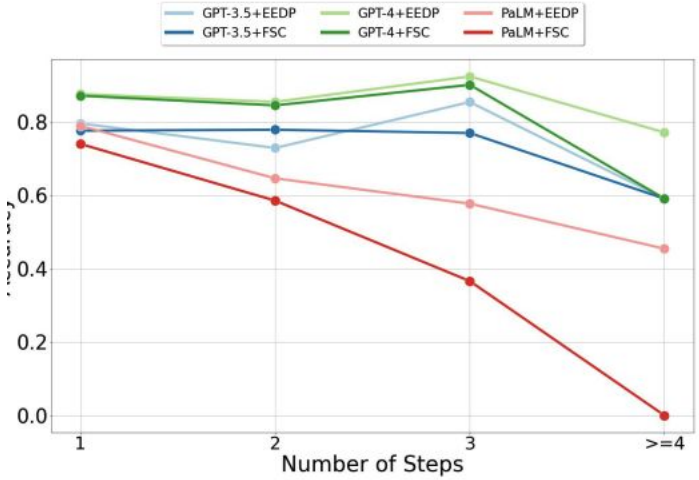
**Hardness:**

Div(“/”) = Multi(“X”) > Sub (“-”) = Add (“+”)

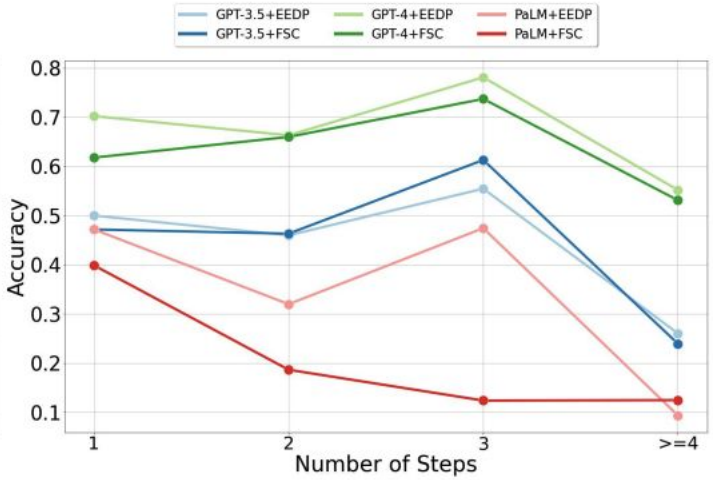
**Scaling:**

Bell curve for multi (“X”) , Div (“/”)

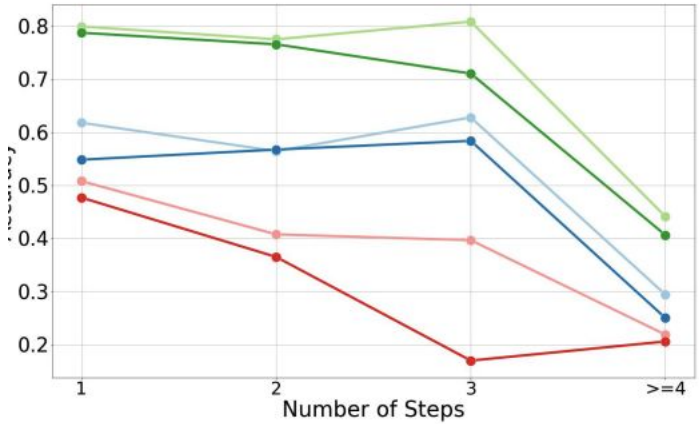
+ve power more harder esp. “+”, “-”



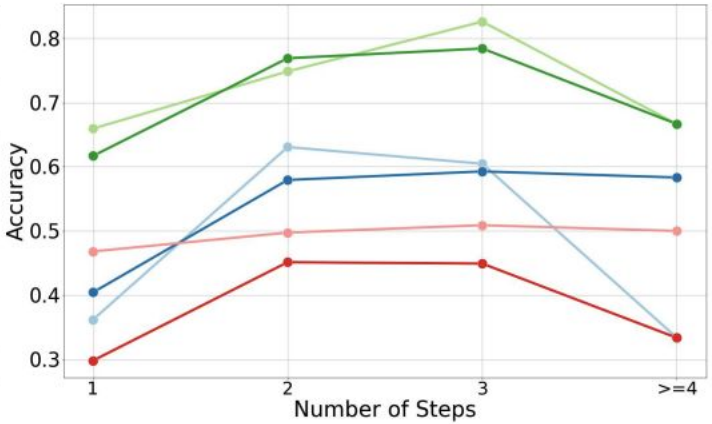

(a) TATQA




(b) MultiHiertt




(c) FinQA

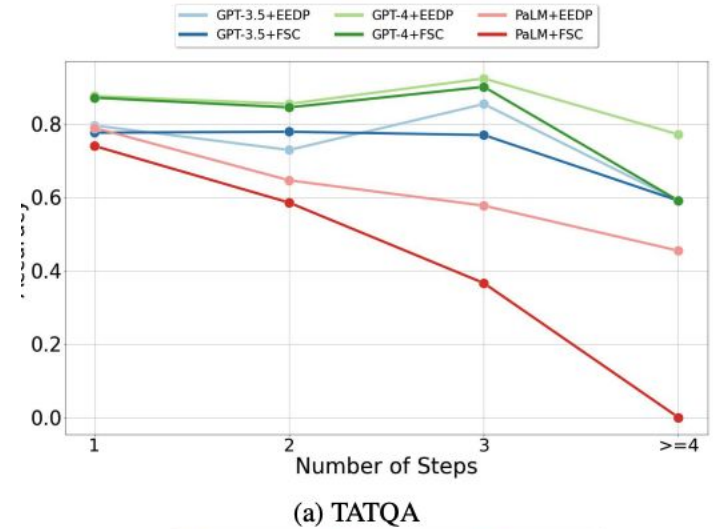



(d) ConvFinQA

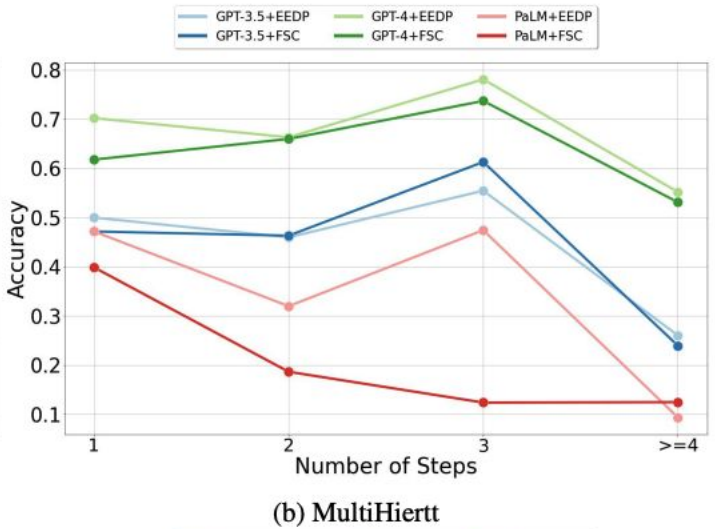
# #Reasoning

#Reasoning Step ↑
Performance ↓

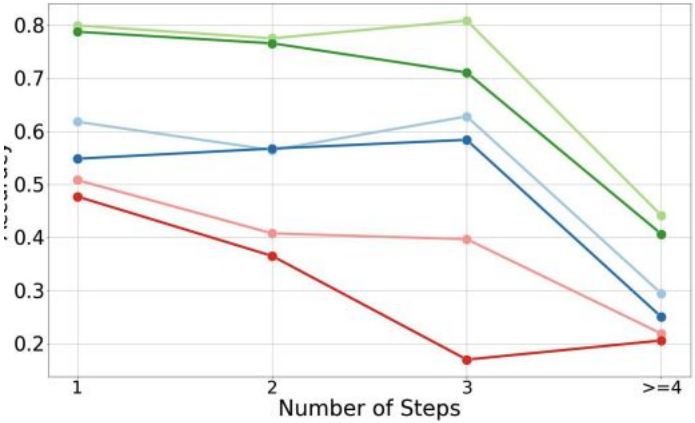
**Control Study**
Variance → different examples



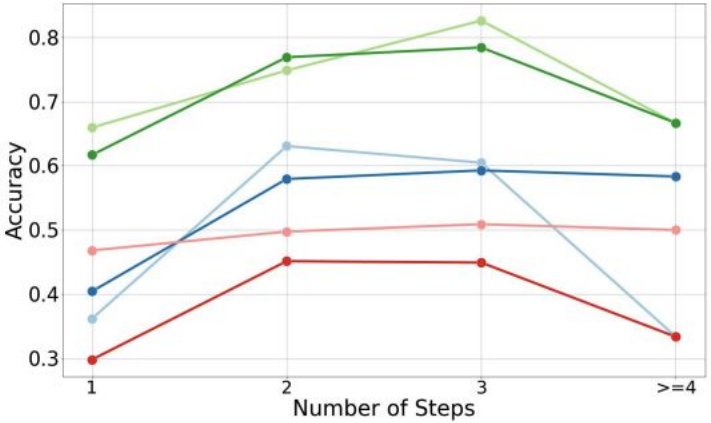

(a) TATQA




(b) MultiHiertt




(c) FinQA




(d) ConvFinQA

## #Reasoning

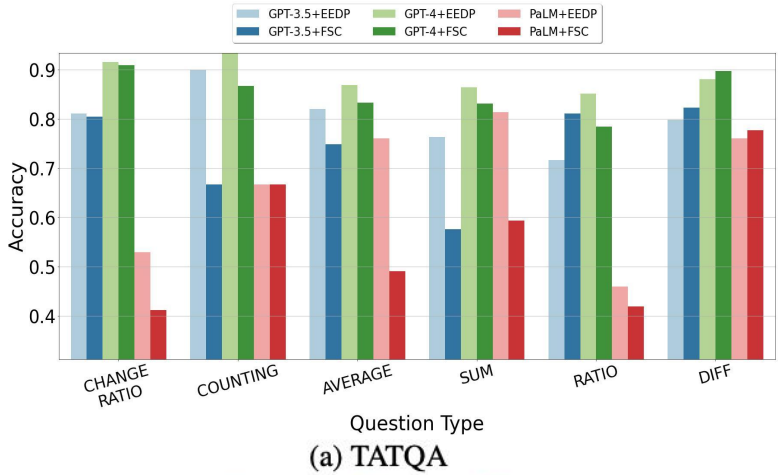
#Reasoning Step ↑
Performance ↓

**Control Study**
Variance → different examples

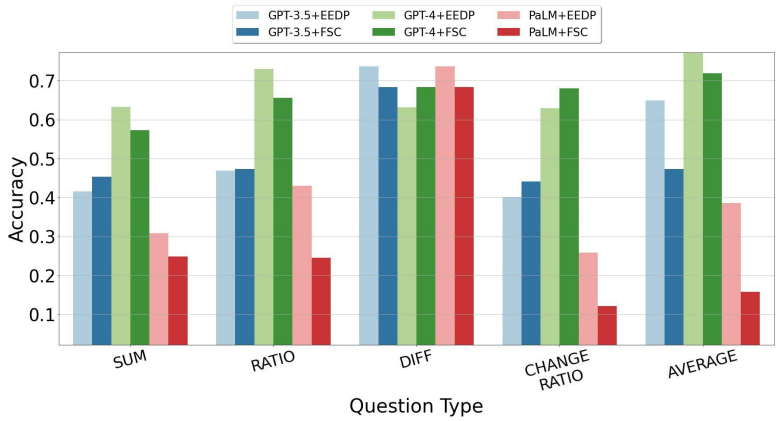
## Exception ConvFinQA

#Example already pre-splits

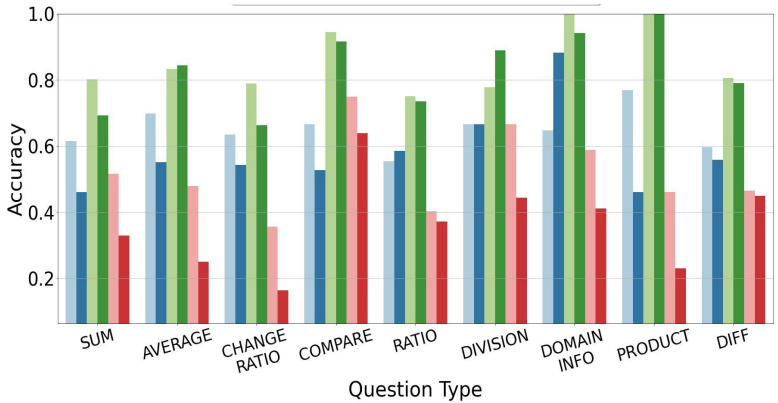
Non split examples are harder



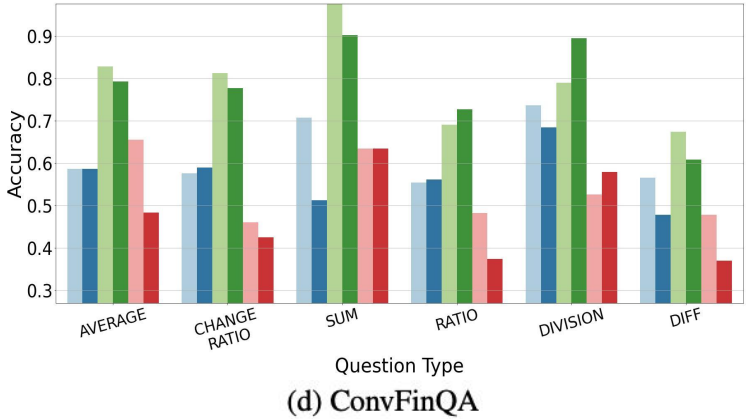

(a) TATQA




(b) MultiHiertt




(c) FinQA




(d) ConvFinQA

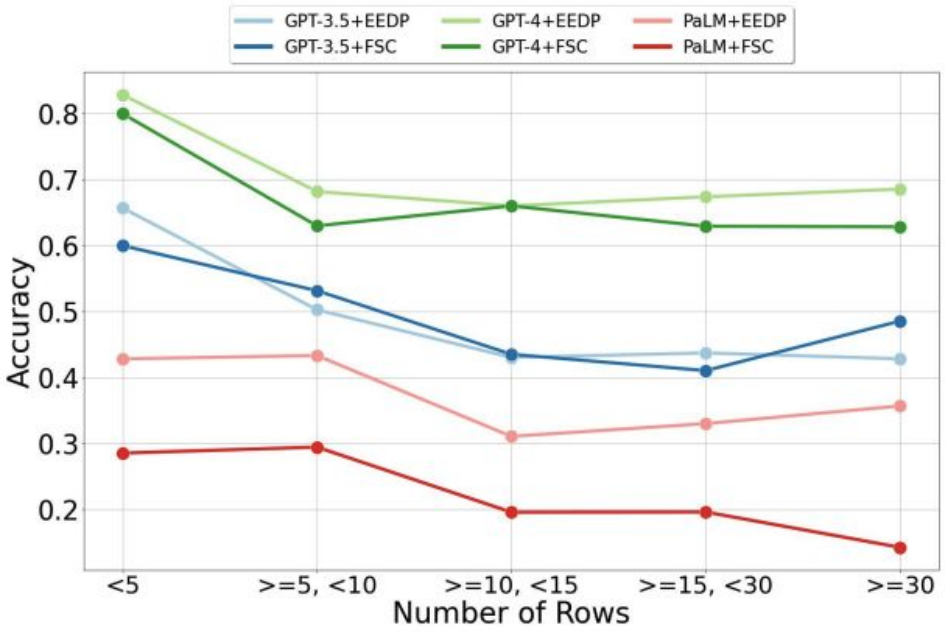
**Harder (GPT-4)**

Change Ratio, Division, Average

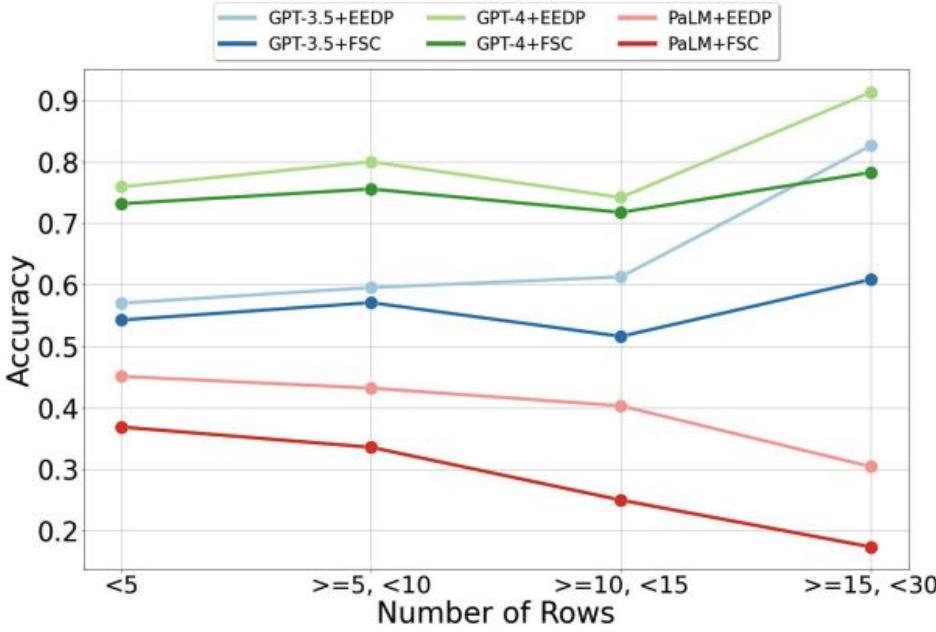
**EEDP more stable**

Lighter > **Darker**

# Number of Rows




(a) MultiHiertt

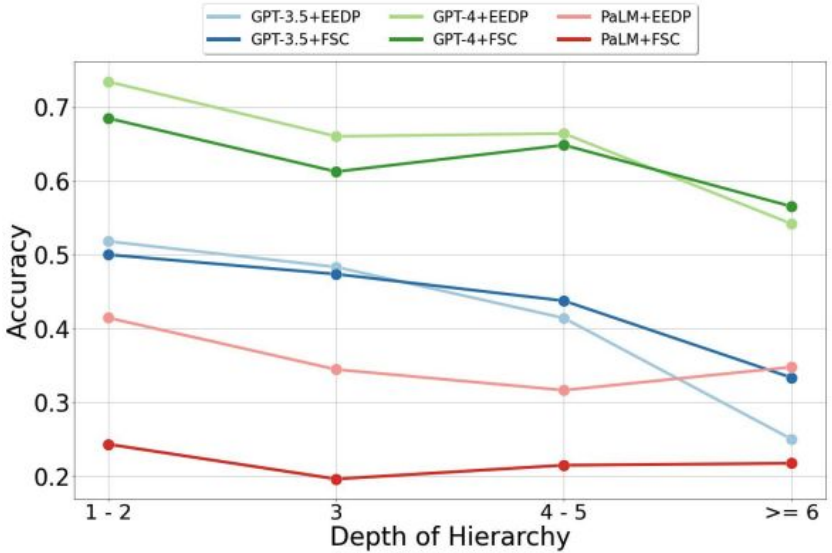



(b) FinQA

Model performance drops as the #rows increase - bracket continent substantial #examples

## Hierarchical Depth

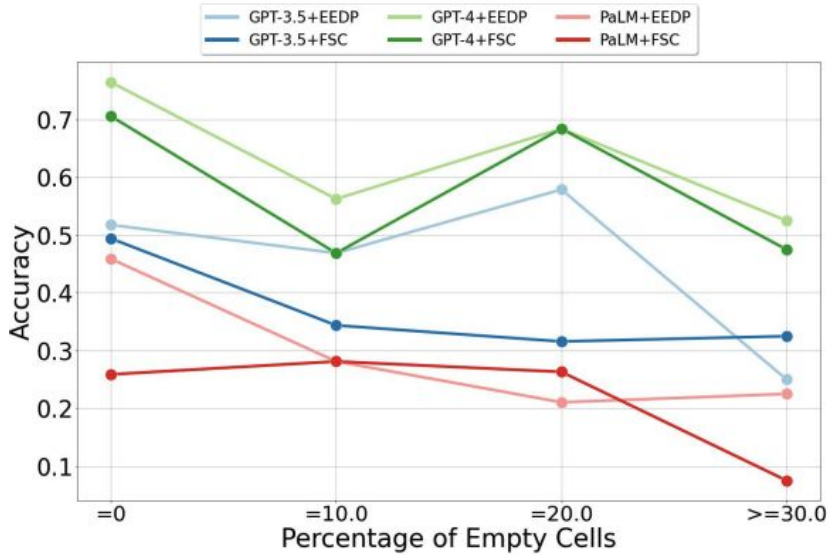
Use **MultiHeirtt** Dataset




Performance ↓ as Depth ↑

## %Empty Cells

Use **MultiHeirtt** Dataset




Performance ↓ as #Empty Cell ↑

# Take Away

Mathematical or Numerical Reasoning challenging for LLM

# Take Away

Mathematical or Numerical Reasoning challenging for LLM

Our proposed approach EEDP based on decomposition strategy outperform other prompting baselines

# Take Away

Mathematical or Numerical Reasoning challenging for LLM

Our proposed approach EEDP based on decomposition strategy outperform other prompting baselines

Model performance is not uniform, especially performance drop ↓ when

*Harder Operations Types:* Division, Multiplication

*Harder Reasoning Question:* Changing Ratio, Average, Proportions etc.

↑ increase in reasoning steps, operand values scales up/down

↑ increase hierarchical depth, in % of empty cells, in # of table rows

# Take Away

Mathematical or Numerical Reasoning challenging for LLM

Our proposed approach EEDP based on decomposition strategy outperform other prompting baselines

Model performance is not uniform, especially performance drop ⬇ when

*Harder Operations Types:* Division, Multiplication

*Harder Reasoning Question:* Changing Ratio, Average, Proportions etc.

⬆ increase in reasoning steps, operand values scales up/down

⬆ increase hierarchical depth, in % of empty cells, in # of table rows

*Overall, our study provide area of improvement, datasets (other domains), models (better open source), reasoning (harder operand, question) and other factors.*

# Tentative Next Steps

- Few datasets exist for numerical reasoning outside the financial sector. Expand into e-commerce, sports, healthcare, and scientific data domains.

# Tentative Next Steps

- Few datasets exist for numerical reasoning outside the financial sector. Expand into e-commerce, sports, healthcare, and scientific data domains.
- LLM performance in numerical reasoning is influenced by multiple factors.

  Isolating single factors is difficult; controlled studies with synthetic and counterfactual datasets might provide clearer insights.

  Getting error type analysis based on each step of EEDP would be interesting aspects. Also for correct examples see is model Right for the Wrong Reason or Hallucinating.

# Tentative Next Steps

- Few datasets exist for numerical reasoning outside the financial sector. Expand into e-commerce, sports, healthcare, and scientific data domains.
- LLM performance in numerical reasoning is influenced by multiple factors.

  Isolating single factors is difficult; controlled studies with synthetic and counterfactual datasets might provide clearer insights.

  Getting error type analysis based on each step of EEDP would be interesting aspects. Also for correct examples see is model Right for the Wrong Reason or Hallucinating.
- Bigger API models like GPT-4 and PaLM 2-540B outperform open-source counterparts.
- There's a need for bigger and effective open source cost effective models for diverse semi-structured data handling. Comparison with other prompting techniques as required.

**Suggestions / Feedback**

# Related Work

Wenhu Chen, *Large Language Models are few(1)-shot Table Reasoners*, EACL 2023

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa, *Large Language Models are Zero-Shot Reasoners*, NeurIPS 2022

Singha. et.al., *Tabular Representation, Noisy Operators, and Impacts on Table Structure Understanding Tasks in LLMs*, Preprint 2023

Mubashara et.al. *Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data*, EMNLP 2023

- Most closed work but studied only coarse grain reasoning on synthetic and counterfactual level