Enhancing Numerical Reasoning with the Guidance of Reliable Reasoning Processes

ACL 2024

Dingzirui Wang, Longxu Dou, Xuanliang Zhang, Qingfu Zhu, Wanxiang Che

Harbin Institute of Technology

Part 1.

Background

- **Numerical reasoning**
  - Generate answers to numerical questions with given evidence
  - The evidence is employed to support reasoning in cases
  - Reasoning processes of the current methods:
    Reasoning processes during fine-tuning can make the prediction more accurate and explainable
    - Such processes could contain information that does not support the answer

ENCORE (Our Method)

Table

Year	Data Group (Col1)	Mobileye (Col2)	…	Total (Col8)
2018 (Row1)	$5,421	$10,278	…	$24,389
2019 (Row2)	$5,424	$10,290	…	$24,513

Text
During the third quarter of 2018, we made an organizational change to combine … approximately $480 million of goodwill was reallocated.

Question
What is the percentage change of total goodwill from 2018 to 2019?

Answer
(24,513 - 24,389) / 24,389

LLM Generation

Rationale
Find the difference between the total amount in 2018 and 2017, which is divided by 2017. **Additionally, the text mentions a change in goodwill allocation in the third quarter of 2018, which should be considered.**

Formula Decomposition

Operand
{Col8, Row2}, {Col8, Row1}, {Col8, Row1}
Operator
(x1 - x2) / x3
Located Formula
({Col8, Row2} - {Col8, Row1}) / {Col8, Row1}

Part 1.

Background

- **Numerical reasoning**
  - Task
    - Generate single or multiple formulas as answers based on the given question
  - Input
    - Textual evidence, Tabular evidence
  - Output
    - One formula consists of operators and operands $(\text{e.g.}, 2 + 1 \times 3)$

TAT-QA

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.

Extracted Question & Answer

Reasoning	Question	Answer	Scale	Derivation
Set of spans (11.94%)	Which were the bottom 2 revenue items for 2017?	**LinkedIn, Other**	-	-

Calculated Question & Answer

Reasoning	Question	Answer	Scale	Derivation
Subtraction (16.17%)	How much of the total revenue in 2018 did not come from devices?	105,226	million	**110,360 - 5,134**

Fengbin Zhu et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 2021. ACL.

Part 1.

Background

- **Numerical reasoning**
  - Questions with simple calculations (e.g., DROP (Dua et al., 2019))
    - Directly generate the formulas
  - More complex calculations (e.g., GSM8K (Cobbe et al., 2021))
    - Employ LLMs with in-context learning with a few samples without fine-tuning

DROP

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million

GSM8K

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Dheeru Dua et al. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 2019. AACL.
Karl Cobbe et al. Training verifiers to solve math word problems. 2021.

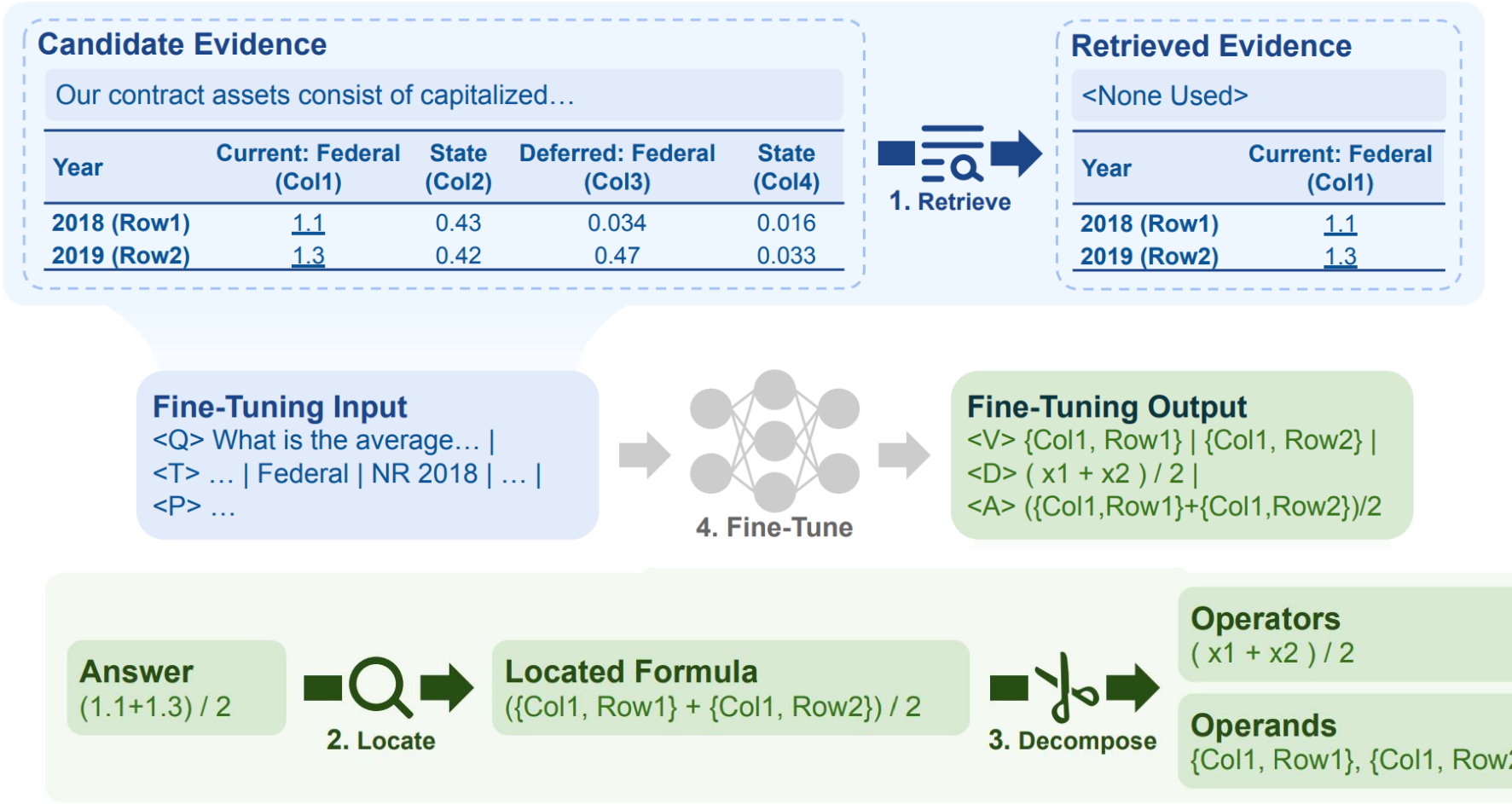
Introduction

- **ENCORE: Enhancing NumeriCal reasOning with Reliable procEsses**
  - Reliable reasoning process
    - Reasoning process does not contain answer-unrelated information
    - Decompose answer formulas for reliable reasoning process
  - Learning difficulty
    - Lack enough data to enable the model to learn the reasoning process generation adequately
    - Our method generates only one single reasoning process, not multiple process

  -> Overcome learning difficulty
    - Present a series of pre-training tasks to help models learn the process generation with synthesized data

Introduction

- **Reasoning Process**
  - 1.Retrieve - 2.Locate - 3.Decompose - 4.Fine-tune



Part 2.

Introduction

- **Reasoning Process**
  - 1.Retrieve - 2.Locate - 3.Decompose - 4.Fine-tune

FinQA

Solve the following questions with the programs.
The program consists of a sequence of operations.
Each operation takes a list of arguments.
There are 6 mathematical operations: add, $subtract$, $multiply$, $divide$, $greater$, exp.
And 4 table aggregation operations: $table-max$, $table-min$, $table-sum$, $table-average$.
The mathematical operations take arguments of either numbers from the given text and table, or a numerical result from a previous step.
The table operations take arguments of table row names.
We use the special token #n to denote the result from the nth step.
The given information is enough to solve the question.

Read the following text and table, and then answer a question:
- | september 24 2005 | september 25 2004 | september 27 2003
beginning allowance balance | $ 47 | $ 49 | $ 51
charged to costs and expenses | 8 | 3 | 4
deductions (a) | -9 (9) | -5 (5) | -6 (6)
ending allowance balance | $ 46 | $ 47 | $ 49
Question: what was the highest ending allowance balance, in millions?
Entities: ending allowance balance
Formula: table_max(x0, none)
Answer: table_max(ending allowance balance, none)

Part 2.

Introduction

- **Reasoning Process**
  - 1.Retrieve - 2.Locate - 3.Decompose - 4.Fine-tune

TAT-QA

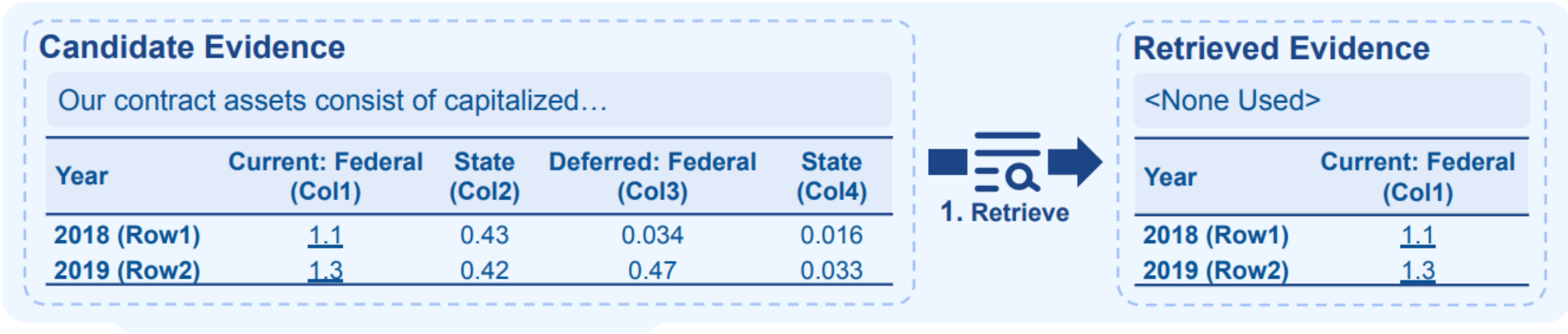
Answer the given question based on the given evidence.
You should generate an formula to answer the arithmetic question.
When answering the question, you should firstly generate the used entities.
Then you generate the formula structure.
Finally you generate the answer formula based on the entities and the formula structure.

GSM8K

Answer the given question.
You firstly generate the used values, which must be mentioned in the question.
Then you generate the formula structure.
Finally you generate the answer formula based on the values and the formula structure.
You only need to generate the formula without any other words, not to calculate the answer.

Method

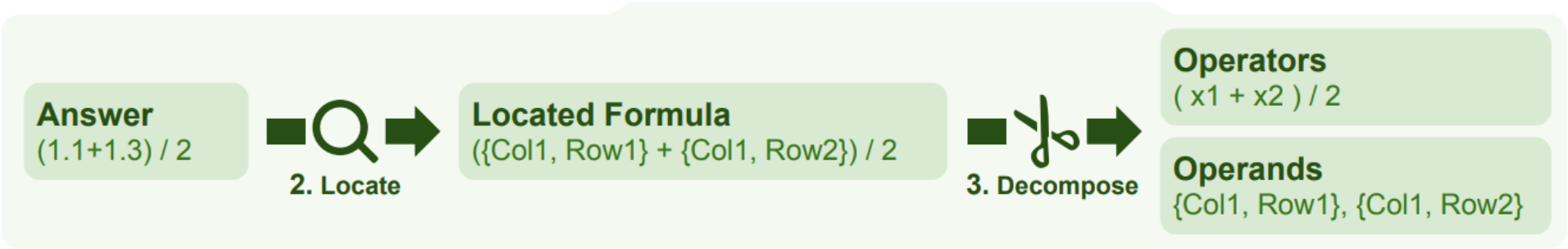
- **1. Retrieve**
  - Binary classification model is trained with the ground evidence annotated by the dataset (e.g., Relevance: Y/N)
  - Sort each text paragraph and table column with the correlation confidence
  - Select the top-k evidence as the retrieval result
  - Concatenate such evidence with the question as the model input



Method

- **2. Locate**
  - Located formula
    - Substitute values in the answer by locating their respective headers in the table
  - Lower the difficulty of specific value memory and table understanding
    - Model only needs to recall the headers associated with relevant cells
    - Enhance the reasoning performance
  - Limitation:

    Use string matching to locate the cell corresponding to each value in the formula
    - Could not handle the cells with duplicated same value



Part 3.

Method

• 2. Locate

Question

I want to know the balance sum from 2018 to 2020

Tabular evidence

Year	Outcome (Col_1)	Income (Col_2)
2018 (Row_1)	18,967	113,246
2019 (Row_2)	19,766	120.523
2020 (Row_3)	21,355	125,843
2021 (Row_4)	22,312	130,725

Answer formula grammar

Rules
Formula → Formula Operator Formula
Formula → (Formula)
Formula → Operand

Gold Answer

$$(113,246 - 18,967) + (120,523 - 19,786) + (125,843 - 21,355)$$

Answer Parsing

(Operand Operator Operand) Operator
(Operand Operator Operand) Operator
(Operand Operator Operand)

Get Formula Structure

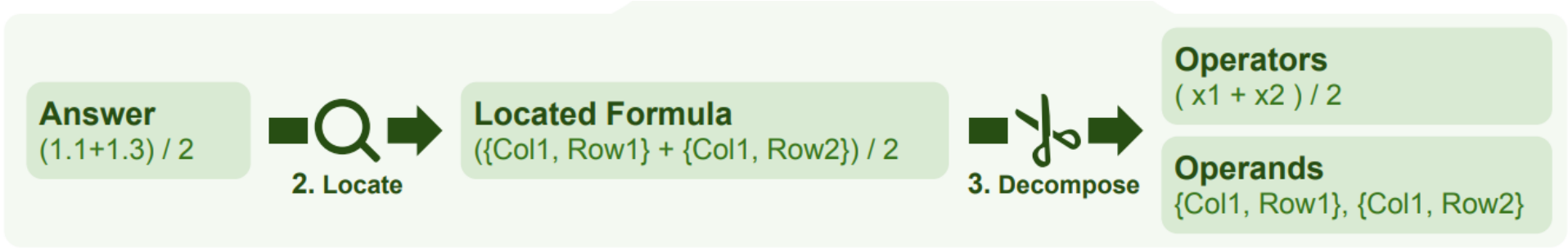
$$(x_1 - x_2) + (x_3 - x_4) + (x_5 - x_6)$$

Employ String match method

$$(\{Col_2, Row_1\} - \{Col_1, Row_1\}) + (\{Col_2, Row_2\} - \{Col_1, Row_2\}) + (\{Col_2, Row_3\} - \{Col_1, Row_3\})$$

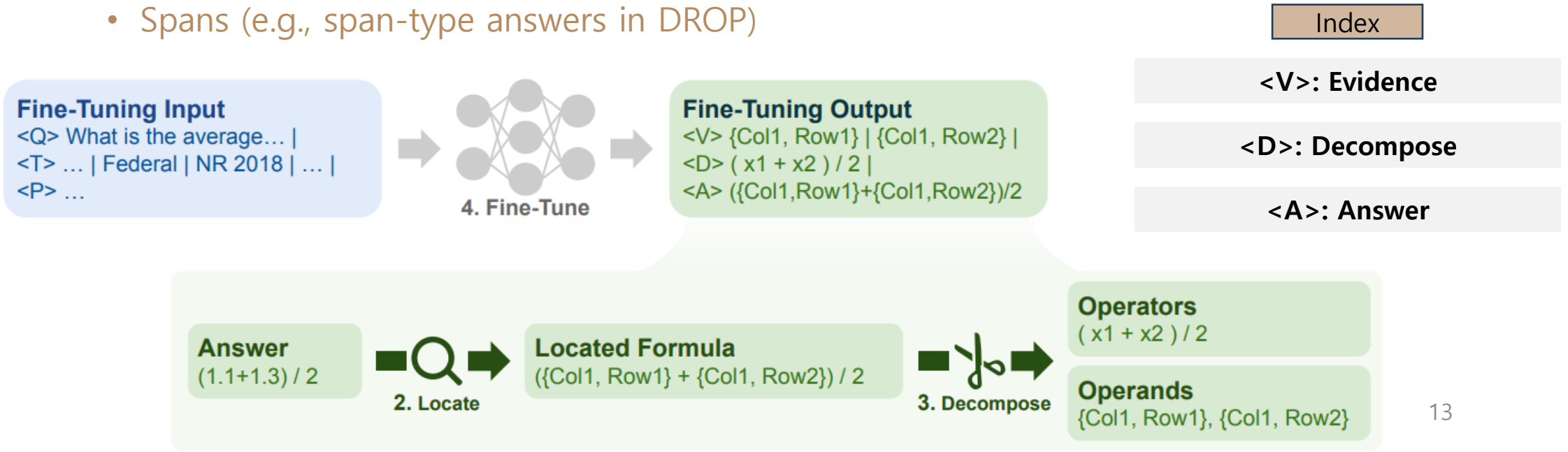
Method

- **3. Decompose**
  - Current methods
    - Ask models to generate formulas in one step
    - Extraction from semantics of the question:
      (e.g., "ratio" in the question leads to the division operator)
  - Reduce the complexity of reasoning through multi-step generation
    - Previous works use multi-step reasoning generation which has high complexity
    - Some formulas are too complex to generate correctly in one step
  - Decompose the formula into operators and operands
    - Operators and operands are used to ask the model to generate before located formulas



Method

- **4. Fine-tune**
  - Input for fine-tuning model
    - Located formulas and corresponding operators and operands
    - Question and the retrieved evidence
  - Tags like the form "<A>" to distinguish different parts
  - Add other information in the output sequence for specific datasets
    - Value scales (e.g., "billion" and "percentage" in TAT-QA)
    - Spans (e.g., span-type answers in DROP)



Part 3.

Method

- **Pre-Training**
  - Overcome the limit of the training data scale
    - Synthesize questions, answers, and reasoning processes based on different templates
    - Pre-train the model with all these data as the multi-task training
  - Primarily design pre-training tasks for tabular evidence rather than textual evidence
    - Most current pre-trained language models are trained on textual data
    - Direct linearization of the tabular evidence during input disrupts the structural information

Pre-Training Task	Question	Answer
Table Location Prediction	What is **{ Col_i , Row_j }** ? *What is **{ Col3 , Row2 }** ?*	The cell value of the headers. ***0.47***
Table Calculation Prediction	What is the max/min/sum/average of **Col_i** ? *What is the sum of **Current : Federal** ?*	The formula of the column. ***{ Col1 , Row1 } + { Col1, Row2 }***
Hierarchical Table Prediction	What is the **{ Col_i , Row_j }** belong to ? *What is the **{ Col2 , Row2 }** belong to ?*	The first-level header of the cell. ***Current***

→ **Enhance Operator generation**

→ **Enhance Operand extraction**

Experiment

- **Experiment Setup**
  - Evaluation Metrics
    - Exatch Match and F1 Score (for arithmetic question)
    - Program Accuracy (e.g., Operand, Operator) and Execution Accuracy (e.g., Evidence)
  - Hyper-Parameter Settings
    - Select three negative examples for every positive instance
    - Retrieve the top 5 as candidate evidence for every question
    - To lower the difficulty of table understanding, we mark the numerical order of each column in the table

Baseline	Model
Retrieval	$\text{BERT}_{\text{BASE}}$
Generation	$\text{BART}_{\text{LARGE}}$, T5_{3B}

Dataset	Domain	Evidence	Answer
FinQA	Finance	Hybrid	Formula
ConvFinQA	Finance	Hybrid	Formula
TAT-QA	Finance	Hybrid	Span
MathQA	MWP	Text	Choice
DROP	Wikipedia	Text	Span
GSM8K	MWP	Text	Formula

Part 4.

Experiment

- **Experimental Result**
  - Published SOTA results are achieved by APOLLO (Sun et al., 2022)
  - From Low quality of the synthesized formulas
    - Misleads the model into erroneous reasoning and poor generation performance (e.g., DROP)
  - Significant improvement on datasets with both textual & tabular evidence (e.g., FinQA, ConvFinQA)

Method	FinQA		ConvFinQA		TAT-QA		MathQA		DROP	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Published SOTA	69.7	68.0	**76.5**	76.0	N/A†	**76.8**	N/A†	**83.0**	N/A†	**90.0**
$\text{BART}_{\text{LARGE}}$	62.5	58.8	67.4	71.5	68.5	–	77.4	78.0	68.6	67.4
+ ENCORE	64.0	62.3	68.9	74.4	71.0	–	77.7	78.8	69.2	68.4
Δ	+1.5	+3.5	+1.5	+2.9	+2.5	–	+0.3	+0.8	+0.6	+1.0
T5_{3B}	66.9	65.0	73.3	79.6	73.8	–	78.1	78.6	77.3	77.1
+ ENCORE	**71.6**	**69.4**	**76.0**	**79.8**	**75.6**	**71.5**	**80.0**	**80.6**	**77.6**	**77.1**
Δ	+4.7	+4.4	+2.7	+0.2	+1.8	–	+1.9	+2.0	+0.3	+0.0

Jia Sun et al. 2022. Apollo: An optimized training approach for long-form numerical reasoning. LREC. 2024.

Part 4.

Experiment

- **Ablation Study**
  - Reasoning Process
    - Extracting operands has the most apparent impact on model performance
    - The model regards the values in the answer as part of the formula structure
  - Pre-Training
    - Pretraining improve model performance by improving numerical reasoning capabilities
    - Effectiveness of those tasks proves that ability to handle operator generation of the baseline is weak

Reasoning Process on TAT-QA

Setting	EM	Arithmetic EM
ENCORE	74.1	78.6
w/o. Operand	72.7 (−1.4)	75.5 (−3.1)
w/o. Located Formula	73.9 (−0.2)	77.3 (−1.6)
w/o. Operator	73.1 (−1.0)	78.3 (−0.3)

Pre-Training on TAT-QA

Setting	EM	Arithmetic EM
ENCORE	75.7	81.2
w/o. Table Location	75.2 (−0.5)	79.8 (−1.4)
w/o. Table Calculation	74.6 (−1.1)	79.4 (−1.8)
w/o. Hierarchical Table	75.3 (−0.4)	80.5 (−0.7)
w/o. All	74.1 (−1.6)	78.6 (−2.6)

Part 4.

Experiment

- **Analysis**
  - SLLM VS LLM
    - Fine-tuned model with ENCORE outperforms baseline using gpt-3.5-turbo
  - Applicability to various answer formats
    - Evaluate them on the unified and divided datasets respectively
    - Adapt ENCORE to the unified setting by merging multiple dataset

Performance on TAT-QA

Method	Arithmetic EM
BART$_{\text{LARGE}}$	73.7
+ gpt-3.5-turbo†	74.7
+ ENCORE	**78.6**

Execution accuracy on TAT-QA

Dataset	BART$_{\text{LARGE}}$		+ ENCORE	
	Single	Unified	Single	Unified
MathQA	79.3	**82.7**	79.5	**84.4**
TAT-QA*	73.7	**79.5**	78.6	**79.8**
FinQA	63.1	**66.1**	65.0	**68.0**
Mixture†	-	**79.5**	-	**81.0**

Experiment

- **Analysis**
  - Compare our method with in-context learning methods
    - Generate directly and with Chain-of-Thought (CoT) (Wei et al., 2022)
    - Use 3-shot / 8-shot prompting (TAT-QA, FinQA / GSM8K)

Execution accuracy of in-context learning

Method	TAT-QA*	FinQA	GSM8K
code-davinci-002	36.2	12.8	19.3
+ CoT	45.4	19.8	60.3
+ ENCORE	**46.0**	**35.1**	**66.3**
gpt-3.5-turbo	25.2	9.9	7.9
+ CoT	38.2	28.2	63.1
+ ENCORE	**55.2**	**39.8**	**71.3**
Llama2-70b	18.5	13.7	16.0
+ CoT	41.9	21.6	54.4
+ ENCORE	**49.2**	**35.1**	**55.3**

Experiment

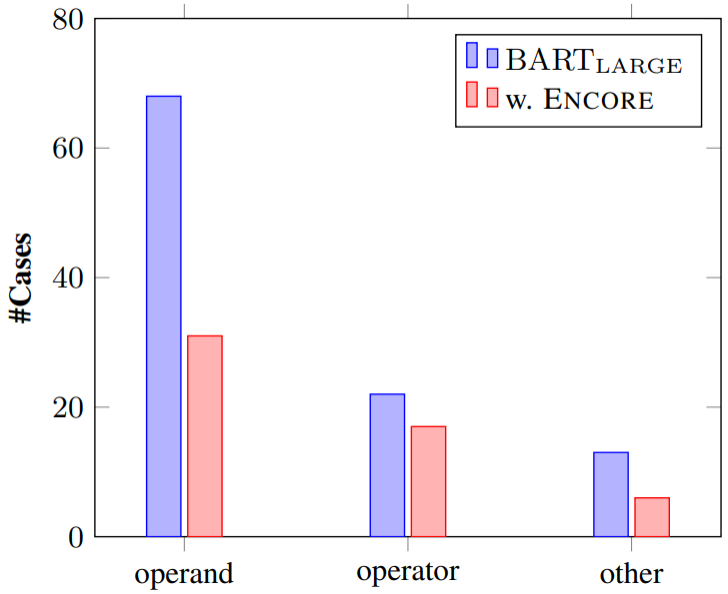
- **Analysis**
  - Increases the performance of table-source and hybrid-source questions
    - Enhance the numerical reasoning ability
  - Suffer from performance degradation on the text-source and span-type
    - Focuse on improving the numerical reasoning ability
    - There are no fixed rules for annotating span-type answers

The exact match on TAT-QA (Without pre-training)

Method	Type				Source			Total
	Span	Spans	Arithmetic	Count	Text	Table	Hybrid	
$\text{BART}_{\text{LARGE}}$	73.0	77.0	73.7	37.5	58.9	73.1	82.4	72.6
+ ENCORE	71.6	77.9	78.6	46.9	56.3	76.2	84.6	74.1
Δ	−1.4	+0.9	+4.9	+9.4	−2.6	+3.1	+1.8	+1.5

Experiment

- **Analysis**
  - Randomly select 256 numerical questions and then analyze them manually
  - Effectiveness of our method
    - Improve the model performance on both operator generation and operand extraction
  - Main error: Operand extraction error
    - Operand extraction and located formula make the model not need to memorize specific values
    - Lower the difficulty of reasoning



The number of error cases of numerical reasoning questions on TAT-QA

Error Type

- Mistakes in the operand extraction
- Mistakes in the operator generation
- Other

Part 5.

Conclusion

- **ENCORE: Enhancing NumeriCal reasOning with Reliable procEsses**
  - Reliable reasoning processes
    - Decompose answer formulas for reliable reasoning process
    - Present a series of pre-training tasks to help models learn the process generation with synthesized data
  - Experimental Result
    - Significant performance improvement over all five experimental datasets
    - Superior by about 10% in comparison to gpt-3.5-turbo
  - Limitation
    - Directly check whether each operand appears in the evidence, which could lead mistake
      -> Adapt better grounding methods
    - It is required that the training data be labeled with formulas, demanding high label overhead
      -> Employ LLMs to synthesize formulas

Appendix

- **Detect the question-related value in the table**
  - Baseline model generates a wrong number, which doesn't exist in the table
  - ENCORE extracts the value according to located formula

Question

What was the percentage change in the amount for **Appliances** in **2019** from **2018**?

Output

Baseline: … | <D> | <V> | <A> (680 - **754**) / **754**

Encore: … | <D> (x1 - x2) / x3 | <V> {Col10, Row1} | {Col10, Row2} | {Col10, Row2} | <A> ({Col10, Row1} - {Col10, Row2}) / {Col10, Row2}

Evidence

Year	…	Data and devices (Col9)	Appliances (Col10)	Total (Col11)	…
2019 (Row1)	…	993	680	13,448	…
2018 (Row2)	…	1,068	774	13,988	…

Part 6.

Appendix

- **Model Comparison**

MathQA

Method	Exe	Prog
FinQANet (Chen et al., 2021)	-	79.2
ELASTIC (Zhang and Moshfeghi, 2022)	-	**83.0**
BART (Lewis et al., 2020)	79.6	78.0
BART w. ENCORE	80.5	78.8
T5 (Raffel et al., 2020)	81.8	78.6
T5 w. ENCORE	82.9	80.6

DROP

Method	Dev		Test	
	EM	F_1	EM	F_1
Discriminative Methods				
NumNet+ (Ran et al., 2019)	81.1	84.4	81.5	84.8
QDGAT (Chen et al., 2020)	**84.1**	**87.1**	**84.5**	**87.6**
Generative Methods				
GPT-3.5* (OpenAI, 2023)	-	-	-	64.1
GPT-4* (OpenAI, 2023)	-	-	-	80.9
BART (Lewis et al., 2020)	68.6	71.7	67.4	70.7
BART w. ENCORE	69.2	72.2	68.4	71.4

Part 6.

Appendix

- **Model Comparison**

TAT-QA

Method	Dev		Test	
	EM	F_1	EM	F_1
TagOp (Zhu et al., 2021a)	55.2	62.7	50.1	58.0
TaCube (Zhou et al., 2022a)	57.1	65.6	-	-
KIQA (Nararatwong et al., 2022)	-	-	58.2	67.4
UniRPG (Zhou et al., 2022b)	70.2	77.9	67.4	75.5
RegHNT (Lei et al., 2022)	73.6	81.3	70.3	78.0
AeNER (Yarullin and Isaev, 2023)	**78.5**	**86.0**	**75.0**	**83.2**
BART (Lewis et al., 2020)	65.7	73.0	-	-
BART w. ENCORE	71.0	78.9	-	-
T5 (Raffel et al., 2020)	73.8	80.9	-	-
T5 w. ENCORE	75.8	82.8	71.5	79.5

FinQA/ConvFinQA

Model	FinQA		ConvFinQA	
	Exe	Prog	Exe	Prog
FinQANet (Chen et al., 2021)	61.2	58.9	68.9	68.2
DyRRen (Li et al., 2022)	63.3	61.3	-	-
APOLLO (Sun et al., 2022)	68.0	**65.6**	76.0	74.6
TabT5† (Andrejczuk et al., 2022)	**70.8**	**68.0**	-	-
BART (Lewis et al., 2020)	58.8	54.4	71.5	69.5
BART w. ENCORE	62.3	57.2	74.4	72.2
T5 (Raffel et al., 2020)	65.0	58.3	79.6	77.3
T5 w. ENCORE	**69.4**	63.7	**79.8**	**77.9**

Part 6.

Appendix

- **FinQA (Question-Answering pairs over Financial reports)**
  - Expert-annotated dataset that contains 8,281 financial QA pairs
  - Eleven finance professionals collectively constructed FINQA based on the earnings reports of S&P 500 companies (Zheng et al., 2021)

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)

Zhiyu Chen et al. Finqa: A dataset of numerical reasoning over financial data. EMNLP. 2021.

Part 6.

Appendix

- **ConvFinQA**
  - Conversational question answering over financial reports
  - Simulate conversation by decomposition and concatenation of the multihop questions from the FinQA (Chen et al., 2021)

Financial report:

	balance at beginning of year	acquisition of hittite	goodwill adjustment related to other acquisitions (2)	foreign currency translation adjustment	balance at end of year
2016	$1,636,526	2014	44046	-1456 (1456)	$1,679,116
2015	$1,642,438	-1105 (1105)	3663	-8470 (8470)	$1,636,526

… notes to consolidated financial statements 2014 (continued) depreciation expense for property , plant and equipment was $ 134.5 million , $ 130.1 million and $ 114.1 million in fiscal 2016 , 2015 and 2014 , respectively …

Original questions from FinQA:

Question 1: what is the percentage change of balance of good will from 2014 to 2015?
Answer 1: subtract(1636526, 1642438), divide(#0, 1642438)
Question 2: what is the percentage change of balance of good will from 2015 to 2016?
Answer 2: subtract(1679116, 1636526), divide(#0, 1636526)

Simulation from question 1

Q1: What's the balance of goodwill by the end of 2014?
A1: 1642438
Q2: and 2015?
A2: 1636526
Q3: what was the change in the balance of goodwill of these 2 years?
A3: subtract(1636526, 1642438) = #0
Q4: how much does this change represent, in percentage, in relation to to that balance in 2014?
A4: divide(#0, 1642438)

Simulation from question 2

Q5: (skip)
A5: 1679116
Q6: (skip)
A6: 1636526
Q7: (skip)
A7: subtract(1679116, 1636526)

Q8: and over the subsequent year, what is that percentage?
A8: subtract(1679116, 1636526), divide(#0, 1636526)

Zhiyu Chen et al. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. ACL. 2022.

Part 6.

Appendix

- **TAT-QA (Tabular And Textual dataset for Question Answering)**
  - Real-world financial reports composed of a table, at least two paragraphs

	Table	Text	Table-text	Total
Span	1,801	3,496	1,842	7,139
Spans	777	258	1,037	2,072
Counting	106	5	266	377
Arithmetic	4,747	143	2,074	6,964
Total	7,431	3,902	5,219	16,552

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.

Extracted Question & Answer

Reasoning	Question	Answer	Scale	Derivation
Set of spans (11.94%)	Which were the bottom 2 revenue items for 2017?	**LinkedIn, Other**	-	-

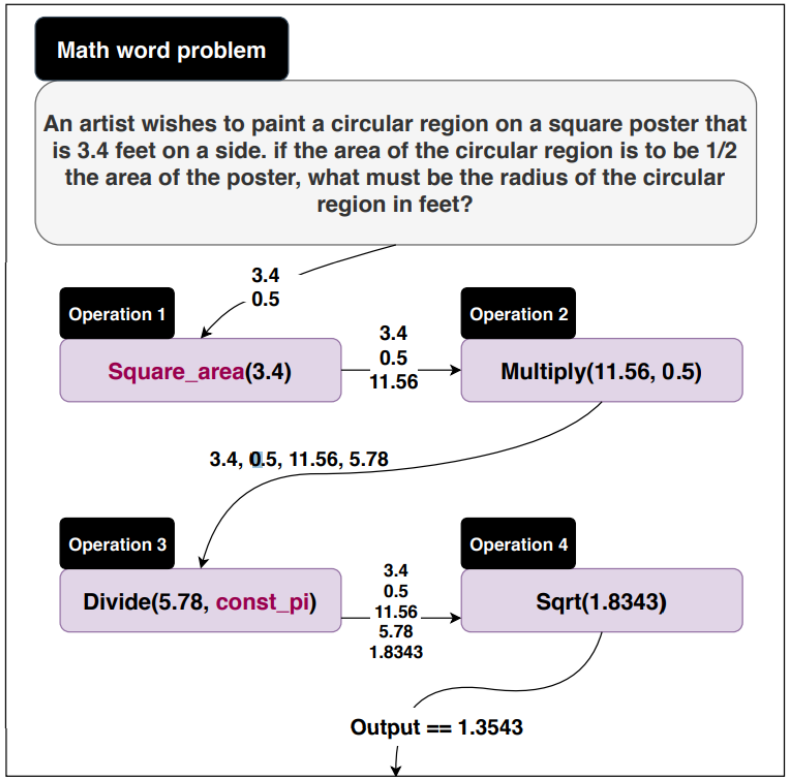
Calculated Question & Answer

Reasoning	Question	Answer	Scale	Derivation
Subtraction (16.17%)	How much of the total revenue in 2018 did not come from devices?	105,226	million	**110,360 - 5,134**

Fengbin Zhu et al. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 2021. ACL.

Appendix

- **MathQA**
  - 37k English multiple-choice math word problems covering multiple math domain categories
  - It requires logical reasoning about implied actions and relations between entities



Problem : A rectangular field is to be fenced on three sides leaving a side of 20 feet uncovered. if the area of the field is 10 sq. feet, how many feet of fencing will be required?
Operations : `divide(10,20), multiply(#0, const_2), add(20, #1)`

Problem : How long does a train 110m long running at the speed of 72 km/hr takes to cross a bridge 132m length?
Operations : `add(110, 132), multiply(72, const_0.2778), divide(#0, #1), floor(#2)`

Koncel-Kedziorski et al. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 2019. of AACL.

Part 6.

Appendix

- **DROP (Discrete Reasoning Over the content of Paragraphs)**
  - Questions that require discrete reasoning
    - Complex questions commonly found in the paragraph's semantic parsing literature: e.g., addition, sorting, counting
    - Combine distributed representations with symbolic reasoning: e.g., Find multiple occurrences of an event described in a question

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. . . . Carolina closed out the half with **Kasay nailing a 44-yard field goal**. . . . In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater

Dheeru Dua et al. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 2019. AACL.

Part 6.

Appendix

- **GSM8K**
  - 8.5k grade school math problems created by human problem writers
  - Problems take between 2 and 8 steps to solve, and solutions for final answer
  - Require a sequence of elementary calculations using basic arithmetic operations

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

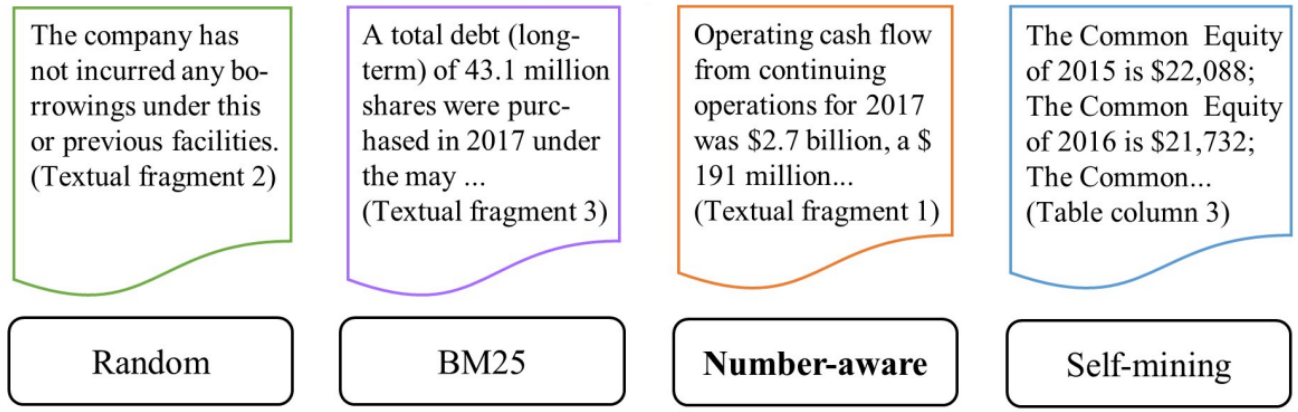
Karl Cobbe et al. Training verifiers to solve math word problems. 2021.

Part 6.

Appendix

- **APOLLO: An Optimized Training Approach for Long-form Numerical Reasoning**
  - Retriever
    - Number-aware negative sampling approach: Prioritize and differentiate between numerical facts
  - Generator
    - Target program augmentation
    - Consistency-based reinforcement learning: Explore the space of consistent programs & improve execution accuracy

Negative Sampling Strategies



Consistency-based Reinforcement Learning

$$R(G_g, G_T) = \begin{cases} -2, & \text{U.E.P} \\ -1, & \text{E.P but wrong answer} \\ +1, & \text{E.P and right answer} \end{cases}$$

$$\nabla L_{\Theta}^{\mathrm{RL}} = -\nabla_{\Theta}\left(\mathbb{E}_{w \sim p_w}\left[R(G_g, G_T)\right]\right)$$

$$\approx -R(G_g, G_T)\,\nabla_{\Theta} \sum_{t} \left(\log p_w\left(w_t; \Theta\right)\right)$$

Jia Sun et al. 2022. Apollo: An optimized training approach for long-form numerical reasoning. LREC. 2024.