IndoMPF: Dataset and Tabular-Graph Numerical Reasoning Baseline Model for Generating Notes to Government Financial Statements

REZA DARMAWAN
Master of Data Science
Faculty of Information Technology

DERRY WIJAYA
Supervisor
Faculty of Information Technology

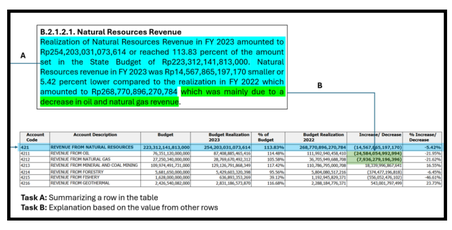
BACKGROUND

Notes to financial statements accompanies a financial statement to enhance its understandability and to avoid misunderstanding.

The current development of Language Model that allows for various text-generation tasks brings the potential of automating Notes to Financial Statement generation.

INHERENT COMPLEXITY

Construction of notes to financial statements contain complexity, beyond accurate summarization and explanation of certain facts from source table.



B.2.1.2.1. Natural Resources Revenue

Realization of Natural Resources Revenue in FY 2023 amounted to Rp254,203,031,073,614 or reached 113.83 percent of the amount set in the State Budget of Rp223,312,141,813,000. Natural Resources revenue in FY 2023 was Rp14,567,865,197,170 smaller or 5.42 percent lower compared to the realization in FY 2022 which amounted to Rp268,770,896,270,784 which was mainly due to a decrease in oil and natural gas revenue.

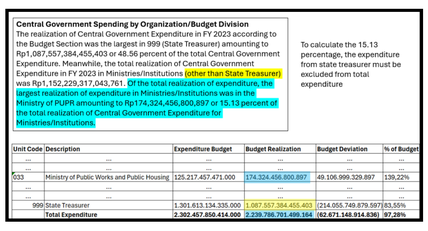
A B

Account Code	Account Description	Budget	Budget Realization 2023	% of Budget	Budget Realization 2022	Increase/ Decrease	% Increase/ Decrease
421	REVENUE FROM NATURAL RESOURCES	223,312,141,813,000	254,203,031,073,614	113.83%	268,770,896,270,784	(14,567,865,197,170)	-5.42%
4211	REVENUE FROM OIL	95,351,135,000,000	87,468,885,465,455	114.68%	111,592,249,858,455	(24,084,004,992,994)	-21.60%
4212	REVENUE FROM NATURAL GAS	37,256,340,000,000	39,769,670,402,557	105.58%	36,395,649,688,790	(3,626,279,286,246)	-21.60%
4213	REVENUE FROM MINERAL AND COAL MINING	100,574,491,772,000	126,126,795,888,349	113.42%	103,786,995,000,788	18,339,900,887,561	18.59%
4214	REVENUE FROM FORESTRY	3,681,650,000,000	5,429,601,326,598	95.96%	5,804,086,517,216	(374,477,190,618)	-6.45%
4215	REVENUE FROM FISHERIES	1,450,000,000,000	1,409,277,000,000	97.24%	1,409,277,000,000	0	0.00%
4216	REVENUE FROM GEOTHERMAL	2,458,540,000,000	2,831,086,577,070	115.68%	2,308,184,776,375	522,901,797,490	22.73%

Task A: Summarizing a row in the table
Task B: Explanation based on the value from other rows

"What should be included?"

Understanding contextual relationships among rows to execute comparative reasoning



Central Government Spending by Organization/Budget Division

The realization of Central Government Expenditure in FY 2023 according to the Budget Section was the largest in 999 (State Treasurer) amounting to Rp1,087,557,384,455,403 or 48.56 percent of the total Central Government Expenditure. Meanwhile, the total realization of Central Government Expenditure in FY 2023 in Ministries/Institutions (other than State Treasurer) was Rp1,152,229,317,043,761. Of the total realization of expenditure, the largest realization of expenditure in Ministries/Institutions was in the Ministry of PUPR amounting to Rp174,324,456,800,897 or 15.13 percent of the total realization of Central Government Expenditure for Ministries/Institutions.

To calculate the 15.13 percentage, the expenditure from state treasurer must be excluded from total expenditure

Unit Code	Description	Expenditure Budget	Budget Realization	Budget Deviation	% of Budget
...	...	...	...	...	...
033	Ministry of Public Works and Public Housing	125.217.457.471.000	174.324.456.800.897	49.106.999.329.897	139,22%
...	...	...	...	...	...
999	State Treasurer	1.301.613.134.335.000	1.087.557.384.455.403	(214.055.749.879.597)	83,55%
	Total Expenditure	**2.302.457.850.414.000**	**2.239.786.701.499.164**	**(62.671.148.914.836)**	**97,28%**

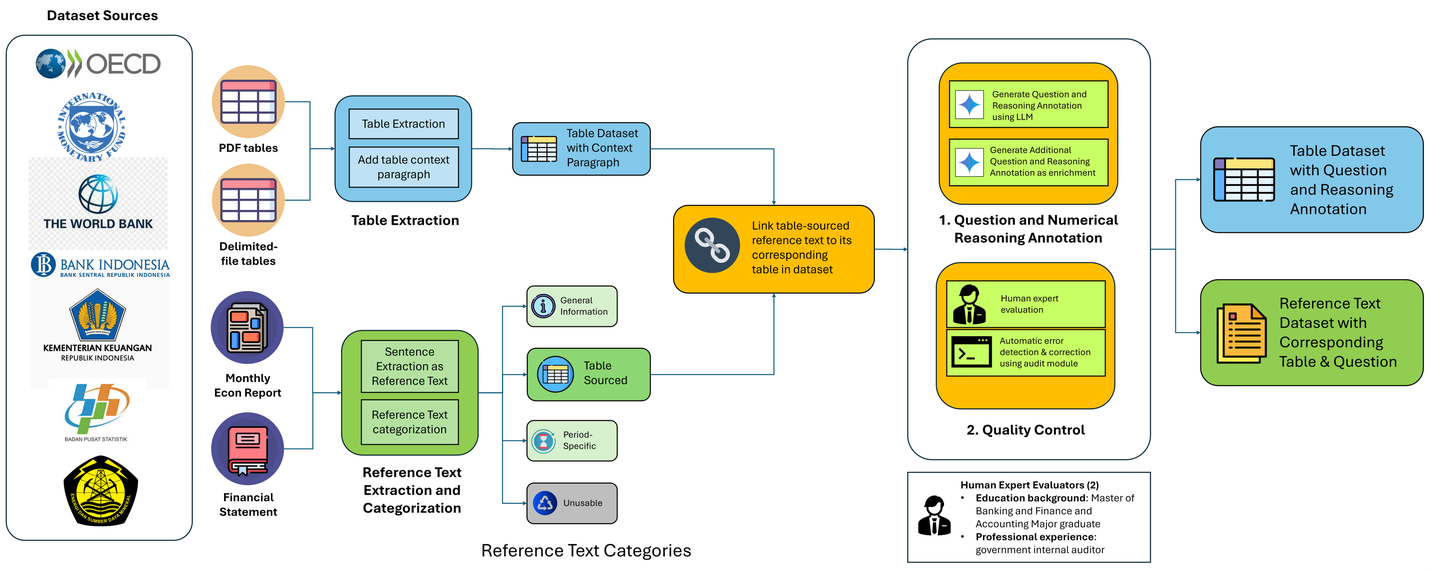
"What should be excluded?"

Exclusion of certain rows when performing numerical operation.

OBJECTIVE

- Construction of table question answering dataset with numerical reasoning with specialized domain in macroeconomic and public finance
- Fine tuning of embedding model using constructed domain-specific dataset for table retrieval task.
- Development of baseline model for table question-answering with numerical reasoning

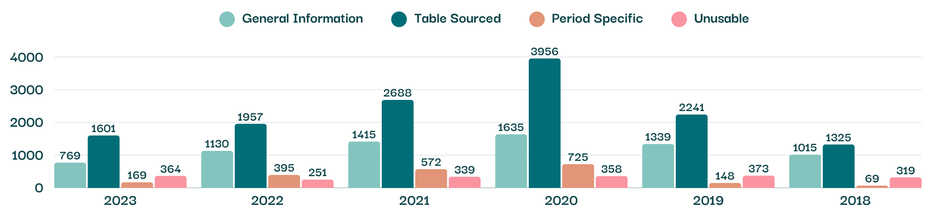
DATASET CONSTRUCTION



REFERENCE TEXT DATASET CHARACTERISTICS

Reference Text Dataset with Corresponding Table & Question

- General information: information generally relevant for all fiscal year.
- Table-sourced: information can be derived and obtained from the source table.
- Period-specific: Information that is only relevant for a fiscal year
- Unusable: Other sentences extracted but is irrelevant to use.



RETRIEVAL MODEL

EMBEDDING MODEL FINETUNING

Base embedding model: Gemma

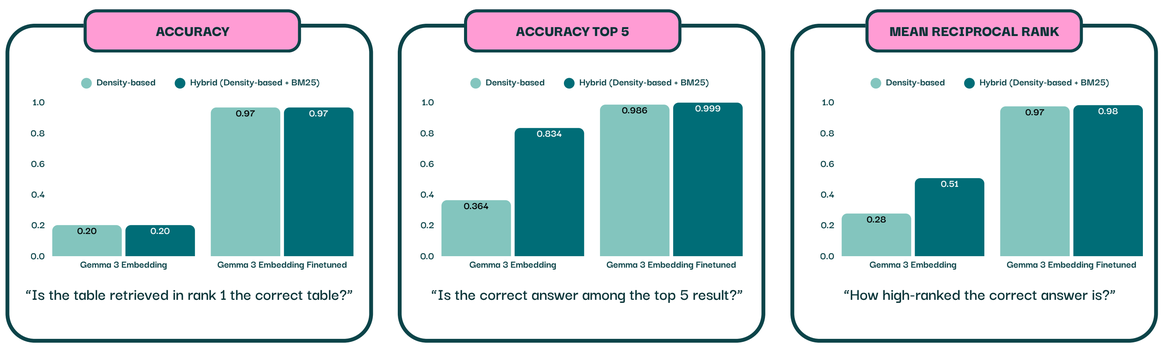
Finetuning Dataset: Table Dataset with Question and Reasoning Annotation

Finetuning Strategy

Multiple Negative Ranking

Contrastive learning technique to train embedding model to distinguish relevant and irrelevant text.

"Which table can be used to answer a question?"



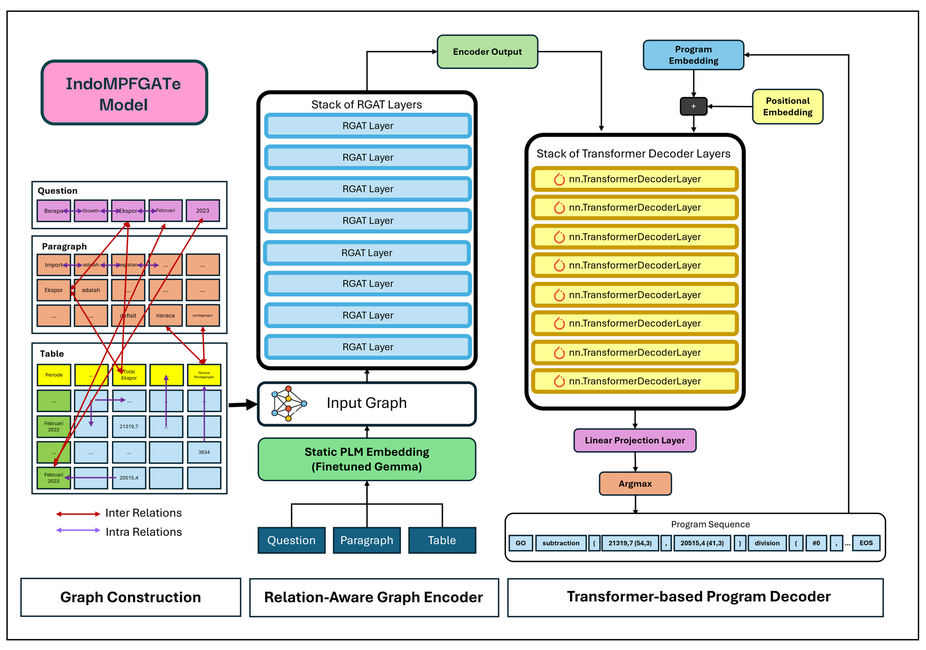
- **Finetuning was the Primary Success Factor**: The use of Multiple Negatives Ranking (MNR) loss for embedding model finetuning objective was exceptionally effective, leading to a nearly five-fold increase in top-1 accuracy. This confirms that for specialized retrieval tasks, finetuning is not just beneficial—it is essential for achieving state-of-the-art performance.
- **Hybrid Retrieval is a Powerful Secondary Enhancement**: The addition of BM25 provides a crucial boost to recall, ensuring that even if the dense model misses, the keyword-based search can often find the correct document. It is a low-cost, high-reward addition to the system.
- **The Combined System is Production-Ready**: The final strategy, combining the MNR-finetuned embedding model with a hybrid (dense + BM25) retrieval system, has achieved near-perfect results on the test set. With an Accuracy@1 of 96.7% and Accuracy@5 of 99.9%, the model is already approaching their practical ceiling.

INDOMPFGATE: BASELINE TABLE QUESTION-ANSWERING MODEL

Graph Representation of Input

Table, question, and explanatory paragraph are represented as graph, to better represent structural information of the table and effectively capture the relation between table and texts.

Source x	Source y	Relation	Description
Intra-Relation			
Q_{word}	Q_{word}	question to next word	y is the next word of x in the question.
P_{word}	P_{word}	paragraph to next word	y is the next word of x in the paragraph.
P_{node}	P_{word}	paragraph to paragraph word	y is the word in the paragraph x .
T_{cell}	T_{cell}	table cell to next cell in row	x is adjacent to cell y in the same table row.
T_{cell}	T_{cell}	table cell to next cell in col	x is adjacent to cell y in the same table column.
T_{cell}	T_{cell}	table cell in the same row	x and y are cells in the same table row.
T_{cell}	T_{cell}	table cell in the same col	x and y are cells in the same table column.
T_{cell}	$T_{rowhead}$	table cell to row header	x is in a row where cell y is the row header.
T_{cell}	$T_{colhead}$	table cell to col header	x is in a column where cell y is the column header.
Inter-Relation			
Q_{word}	P_{word}	question to paragraph word	x and y is the common word of question & paragraph.
Q_{word}	P_{node}	question to paragraph	Paragraph y contains word x in the question.
T_{cell}	Q_{word}	table to question	Table cell x contains question word y.
T_{cell}	P_{node}	table to paragraph	Paragraph y contains cell x.
T_{cell}	P_{word}	table to paragraph word	Table cell x contains paragraph word y.



Graph Encoder

- Aggregates information about nodes and edges of graph.
- Relation awareness is incorporated through embedding of edges.

Transformer Decoder

- Autoregressive decoder that generates the program token by token.
- During prediction, masking is applied to ensure program generated is grammatically correct.

Results Epoch 85

Validation F1	Validation EM	Test F1	Test EM
0.0045	0.0038	0.0009	0.0008
	Can only answer 1 question correct.		Can only answer 1 question correct.

Baseline model underperformance is possibly due to architectural choice of static PLM embedding, to compensate for limited computing resource. Other tabular graph question-answering model includes PLM as learnable weight during model training.

TABLE QA DATASET CHARACTERISTICS

Table Dataset with Question and Reasoning Annotation

IndoMPF table question-answering dataset can contain multi-step numerical reasoning, using a diverse type of operators.

Periode	Ekspor Migas	Ekspor Non Migas	Total Ekspor	Impor Migas	Impor Non Migas	Total Impor	Neraca Perdagangan
.....							
Januari 2022	903.5	18291.3	19194.8	2229.2	15981.9	18211.1	963
Februari 2022	1029.6	19485.8	20515.4	2902.3	13736.2	16638.5	3834
Maret 2022	1493.3	25138.5	26631.8	3492.2	18470.2	21962.4	4535
April 2022	1466.6	25904.4	27371	3813.9	15943.6	19757.5	7565
Mei 2022	1498	20036.2	21534.2	3353.7	15255.6	18609.3	2900
.....							
Tahunan 2022	15998.2	276188.5	292186.7	40416.4	197026.6	237443	54456
Januari 2023	1487.9	20832.8	22320.7	2906.1	15536.8	18442.9	3880.9
Februari 2023	1186.5	20133.2	21319.7	2406.5	13512.7	15919.2	5402.1
Maret 2023	1338.2	22075.7	23413.9	3014.8	17573.3	20588.1	2827.9
April 2023	1258.7	18021.7	19280.4	2955	12392.6	15347.6	3936.5
Mei 2023	1308.6	20397.9	21706.5	3135.1	18144.5	21279.6	427.2
.....							
November 2024	1314.4	22798.3	24112.7	2570.3	17061.4	19631.7	4366.5
Desember 2024	1539.4	22060.2	23599.6	3296.8	17925.4	21222.2	2238.9
Tahunan 2024	15876.9	250652.3	266529.2	36275.1	197384.6	233659.7	31043.5

Question
Berapa persentase pertumbuhan total ekspor secara tahunan (yoy) pada Februari 2023? (Answer: 3.92)

Calculations

$$\frac{(21319.7 - 20515.4)}{20515.4} \times 100 = 3.92$$

Program

retrieve (21319.7) | retrieve (20515.4) | subtraction (#0, #1) | division (#2, #1) | multiplication (#3, const_100)

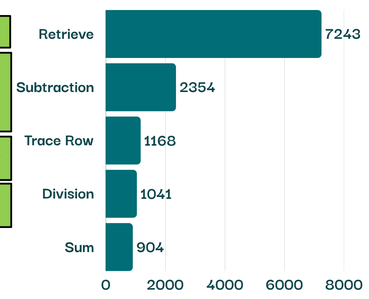
Context Paragraph
Ekspor adalah Kegiatan mengeluarkan barang dan jasa dari wilayah suatu negara, baik bersifat komersial maupun bukan komersial (.....)

Available Operators in the IndoMPF Dataset

SINGLE OPERATION	**unique** (can have more than 2 operands) Getting the unique values among operands.	**retrieve** (1 operand) Retrieve certain cell from the table	**addition** (2 operands) The total between two operands.
	subtraction (2 operands) Subtraction, the order of operands represent the order of operation	**multiplication** (2 operands) Multiplication of 2 operands	**division** (2 operands) Division, with the first operand become nominator, and the other the denominator
	exponential (2 operands) Power operation, with the first operand being the base and the second is the exponent	**greater** (2 operands) Compare two values to find the greater, yields true or false.	**smaller** (2 operands) Compare two values to find the smaller, yields true or false.
	equal (2 operands) Compare two values to evaluate equality, yields true or false.		
AGGREGATE	**sum** (can have more than 2 operands) To calculate the total of values given.	**average** (1 operand) To calculate the average of values given	**count** (2 operands) to identify the number of values given
	sum_if (2 operands) to identify the sum or total of elements value that satisfy criteria to certain predefined value. The first operand is the value that will be compared with the other value.	**count_if** (2 operands) to identify the number of elements that satisfy that the value satisfy criteria to certain predefined value. The first operand is the value that will be compared with the other value.	**filter_if** (2 operands) to identify which elements that satisfy criteria to certain predefined value. The first operand is the value that will be compared with the other value.
	minimum_n (can have > 2 operands) to identify the number of "n" minimum of values given. The first operand is a constant that represents the number of "n" minimum to take. It can return more than one value depends on "n" of the first operand.	**maximum_n** (can have > 2 operands) to identify the number of "n" maximum of values given. The first operand is a constant that represents the number of "n" maximum to take. It can return more than one value depends on "n" of the first operand.	**minimum** (can have > 2 operands) To identify the minimum of values given. Only returns one value.
			maximum (can have > 2 operands) to identify the minimum of values given. Only returns one value.
TRACE	**trace_column** (can have > 2 operands) Get the value of other row, e.g. header row. The first operand is the index of the row to be traced along the column. Therefore the "source" of this operand is "constant". The second operand is the value to be traced.	**trace_row** (can have > 2 operands) Get the value of other column in the same row. The first operand is the index of the column to be traced along the row. Therefore the "source" of this operand is "constant". The second operand is the value to be traced.	

Total QA	Total Table
7007	156

Top 5 Most Used Operators in the IndoMPF Dataset

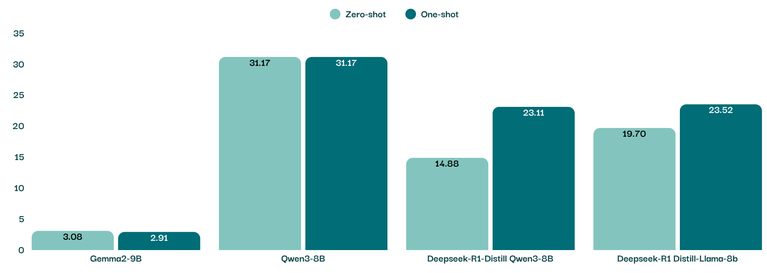


Comparison with Existing Finance Table QA Dataset

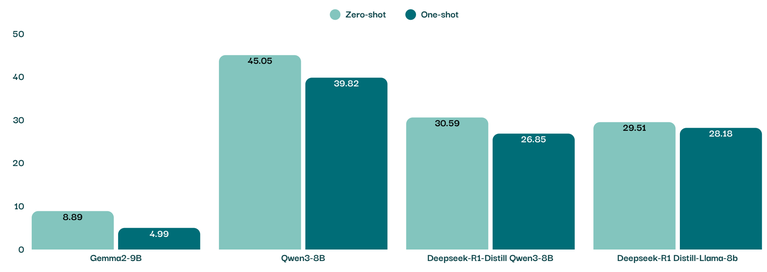
Dataset	# of Question	# of Operators	Type of Numerical Reasoning	Language	Theme
TAT-QA (Zhu et al 2021)	16.552	10	Single Operation	English	Finance
FinQA (Chen et al 2021)	8.281	10	Multi-step	English	Finance
Multihiertt (Zhao et al 2021)	10.440	4	Multi-step	English	Finance
IndoMPF (Ours)	7.007	37	Multi-step	Bahasa Indonesia	Macroeconomic and Public Finance

LLM BENCHMARK

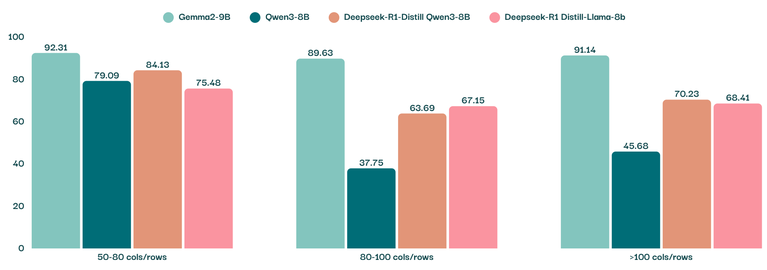
Exact Match - Percentage of Correctly Generated Program



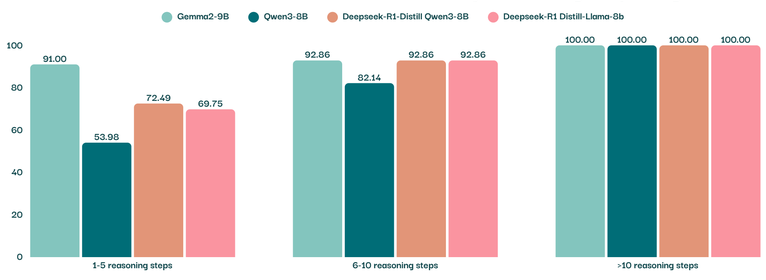
Correct Answer - Percentage of Correctly Answered Questions



Pitfalls Identification - % Incorrect Answer by Table Dimension



Pitfalls Identification - % Incorrect Answer by Reasoning Step Length



- Open source mid-sized LLMs appear to struggle with answering IndoMPF dataset.
- Finetuning is required to enhance the medium-sized LLM ability to answer the question.

FUTURE WORKS

- Utilization of reference text dataset to fine tune text generation model.
- Utilization of reference text dataset to create knowledge graph to be retrieved in text generation.
- Utilization of table question-answering dataset to finetune reasoning model.
- Utilization non-static PLM Embedding for baseline model.

Project Github Page | Dataset Webpage | Grab this poster!