SIMPLOT: Enhancing Chart Question Answering by Distilling Essentials

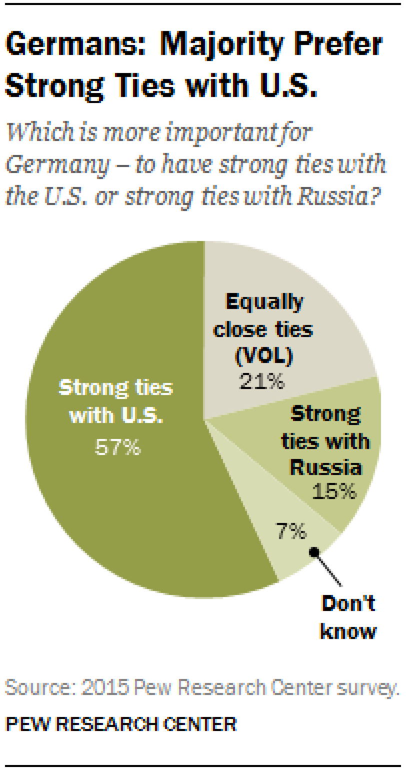
Wonjoong Kim*, Sangwu Park*, Yeonjun In, Seokwon Han, Chanyoung Park

KAIST

DSAIL Data Science & Artificial Intelligence

Introduction

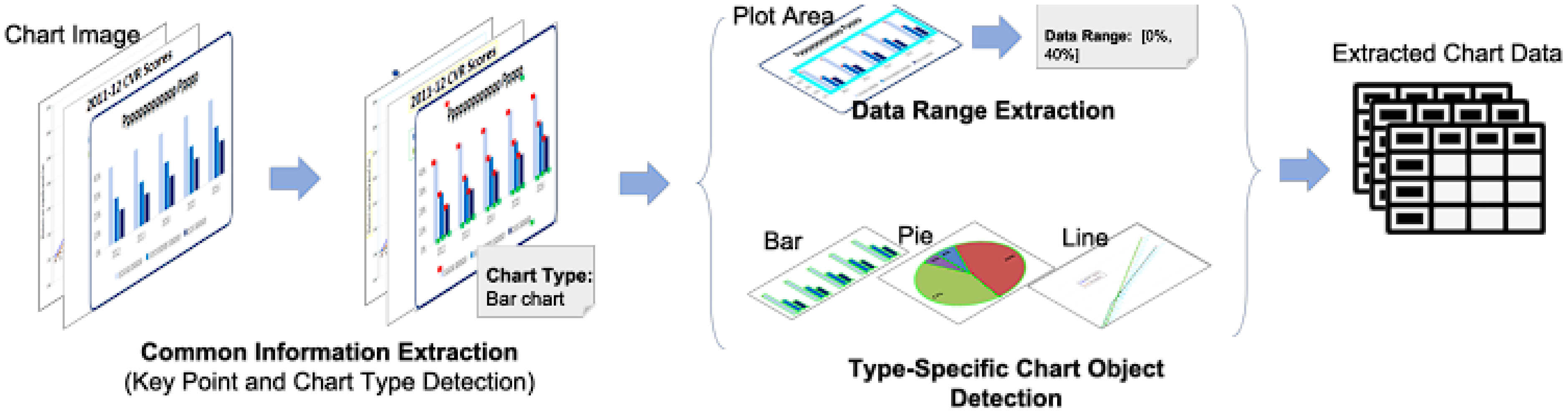
Recent advance in Vision Language Model & Limited research on Mathematical Reasoning



- Increased interest in **advanced reasoning models** from images
- Models still fall short in achieving sufficient performance for specific types of images, **such as charts**
- Charts have unique formats (e.g. columns, rows) requiring a **different learning approach** compared to traditional VQA models

Introduction

Existing chart reasoning methodology and limitation



Heuristic rule based

- Applicable only to charts with predefined formats
- New rules need to be added when a new format is introduced

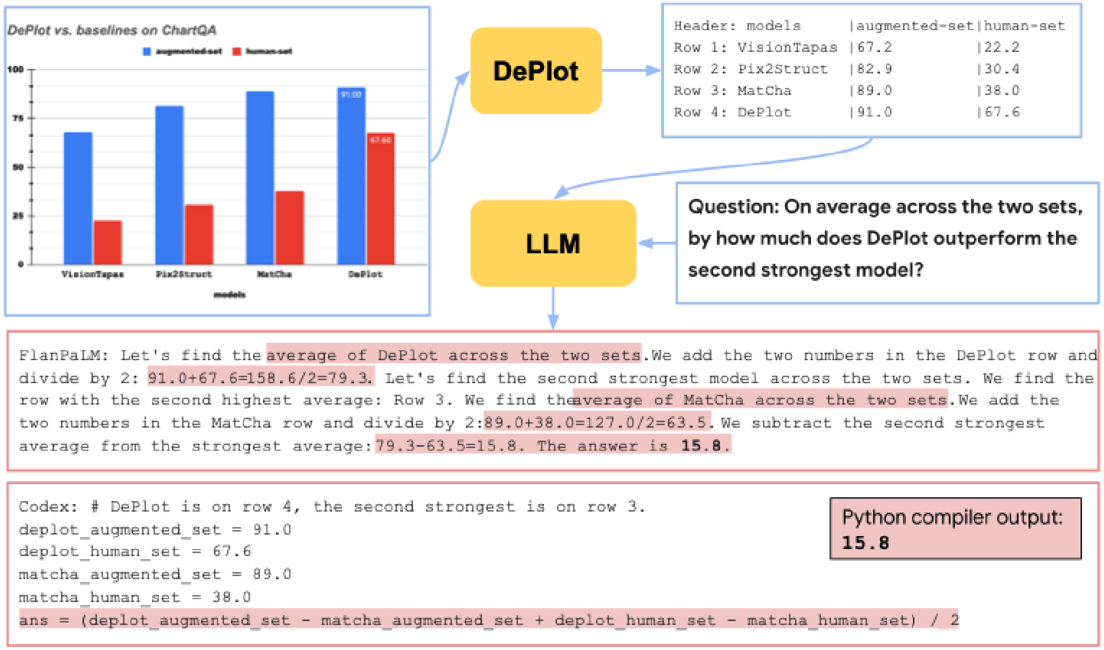
Using OCR / Key-point detection module

- Highly dependent on OCR / Key-point detection module, time consuming
- High annotating cost for dataset
- Most of research conduct only chart component detection, not reasoning

Luo, Junyu, et al. "Chartocr: Data extraction from charts images via a deep hybrid framework." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021.

Introduction

Existing chart reasoning methodology and limitation



Vision-Language Models

- To address the issues, end-to-end trained vision-language models are used
- However, each downstream task (e.g. QA, summarization) requires **separate fine-tuning, limiting scalability**

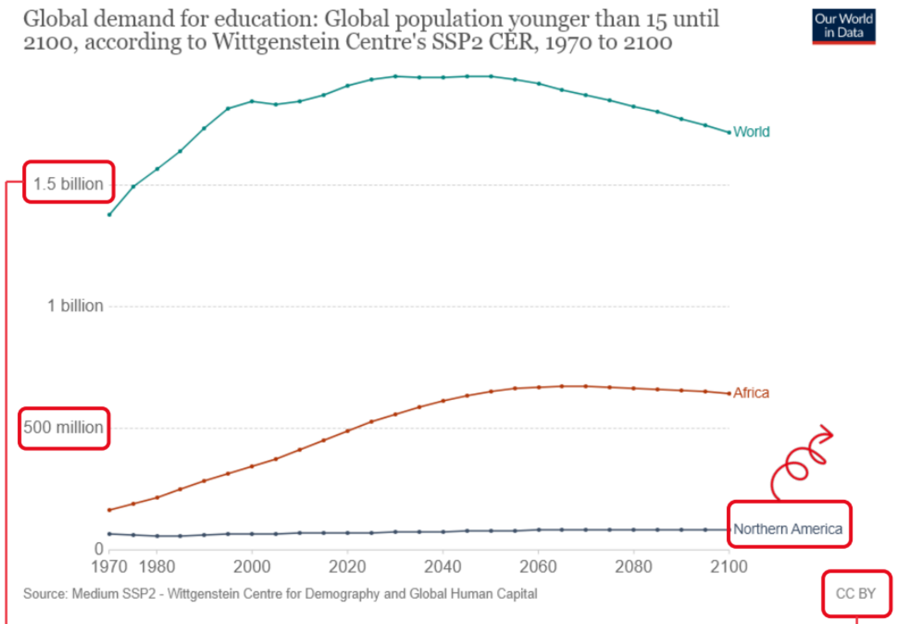
Vision-Language Models + LLM

- To address the above issues and apply the performance of LLM, a method has emerged where the **chart is first converted into a table** and then **reasoning with LLM**
- This enables **interpretability and high performance** in QA tasks

Liu, Fangyu, et al. "DePlot: One-shot visual language reasoning by plot-to-table translation." *Findings of the Association for Computational Linguistics: ACL 2023*. 2023.

Motivation

Limitation of SOTA method



(a) Input chart

Entity	1970	1980	...	2080	2100
World	*1.37*	*1.55*	...	*1.79*	*1.71*
Africa	190	200	...	650	637
CC BY	*1.3*	*1.3*	...	*1.3*	*1.3*

(b) Extracted table

Limitation of SOTA method

- Focusing only on image features to convert to a table, the extraction process cannot utilize **text information** (context)

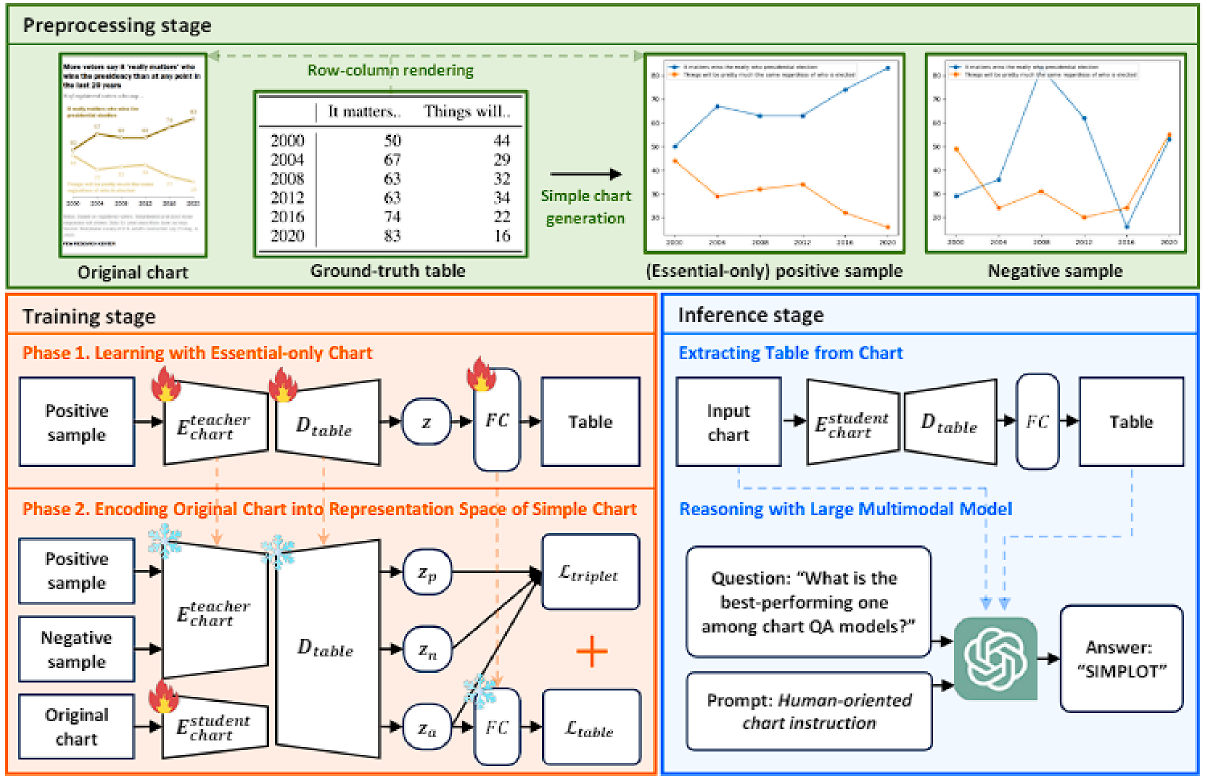
  ex) confusing **billion** and **million** leads to incorrect extraction of table values

- Real-world charts are highly complex, containing a mix of **unnecessary text and visual information**, making it difficult for models to interpret

  ex) fails in table extraction by recognizing 'CC BY' as a column

Method

Brief explanation of SIMPLOT

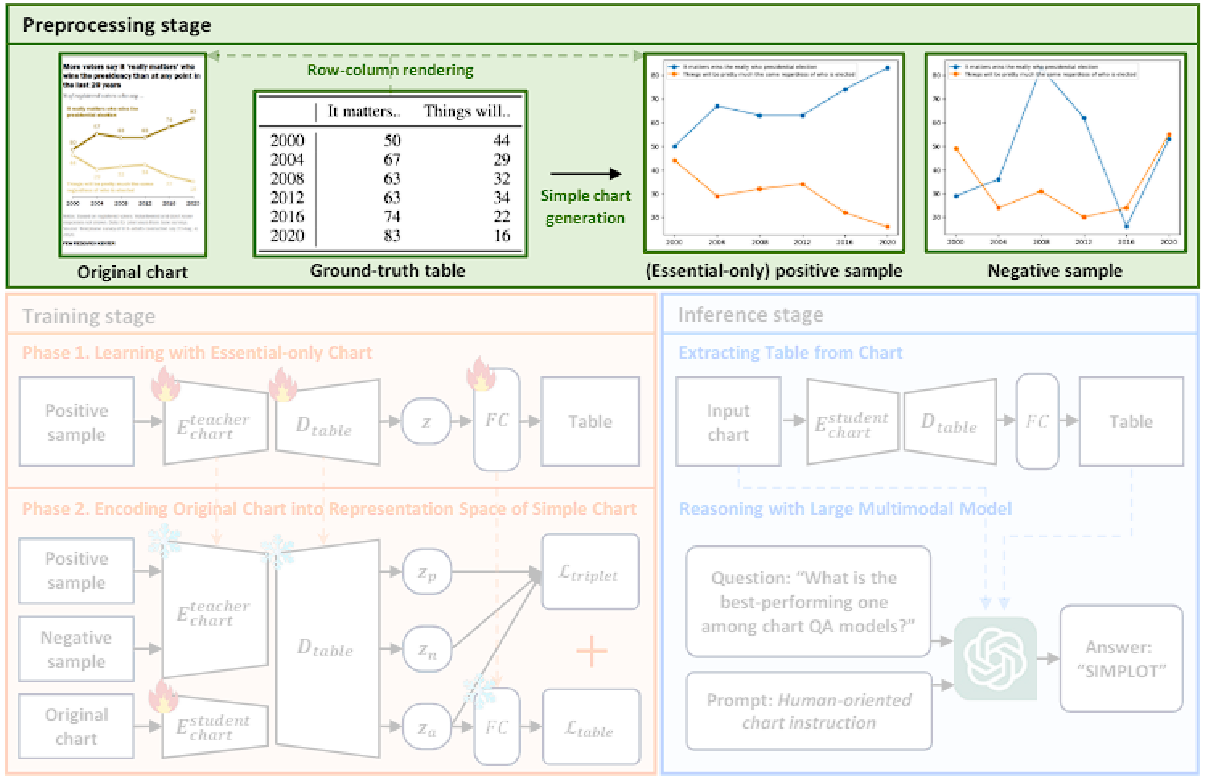


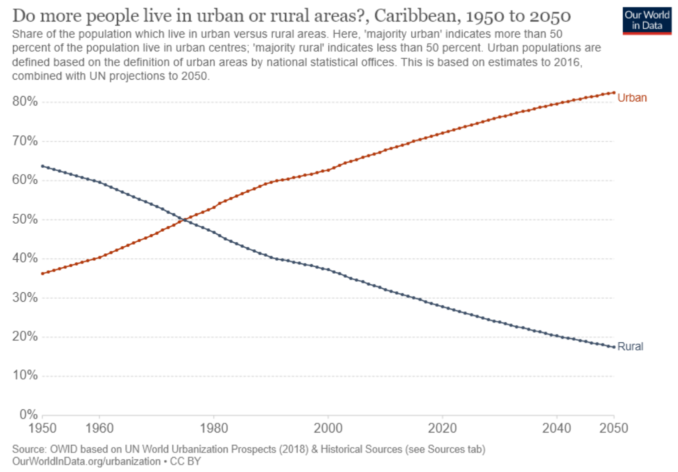
Proposed method (SIMPLOT)

- **Pre-extracting the columns and rows** of the image and rendering them helps the model's table extraction process
- Create **simple charts containing only the essential information** for reasoning, and train the model to **extract only the necessary details** from complex charts
- Enhance the chart reasoning performance of LLM by using prompts that **mimic how humans interpret charts**

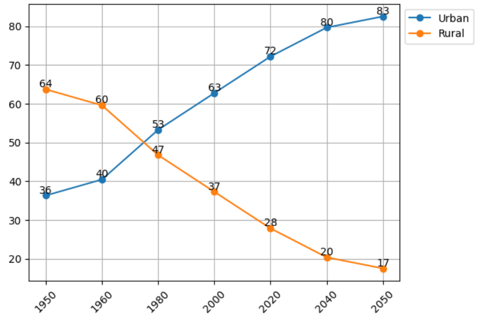
Method

Preprocessing Stage – Simple Chart Generation

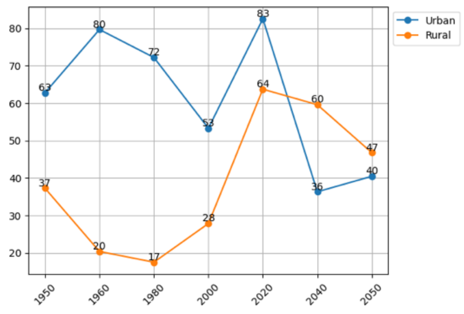




Ground-truth chart



Positive chart

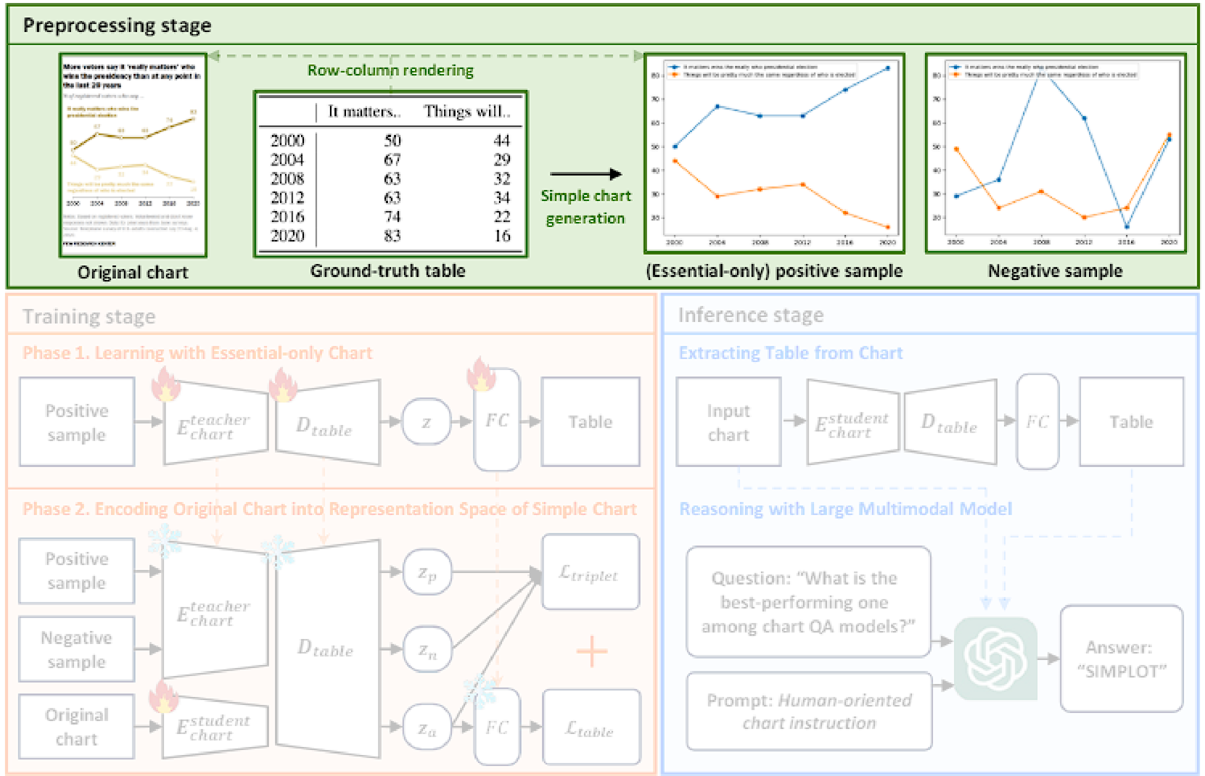


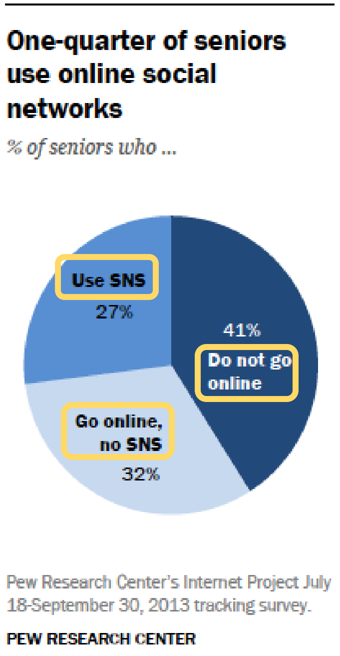
Negative chart

- Generate a simple chart with **essential component** extracted from real-world chart
- Generate **positive charts** and create **negative charts** by shuffling the values

Method

Preprocessing Stage – Row Column Rendering





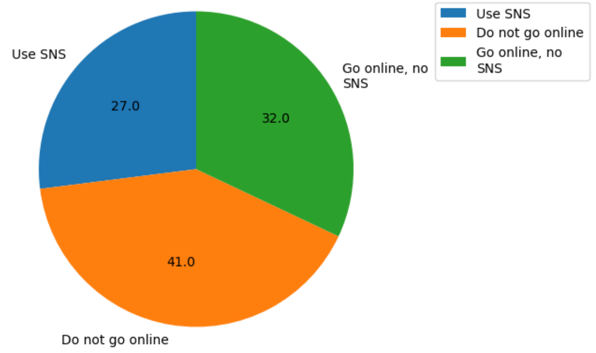
Original chart

Entity	Value
Use SNS	27
Do not go online	41
Go online, no SNS	32

Ground-truth table

Row-column rendering

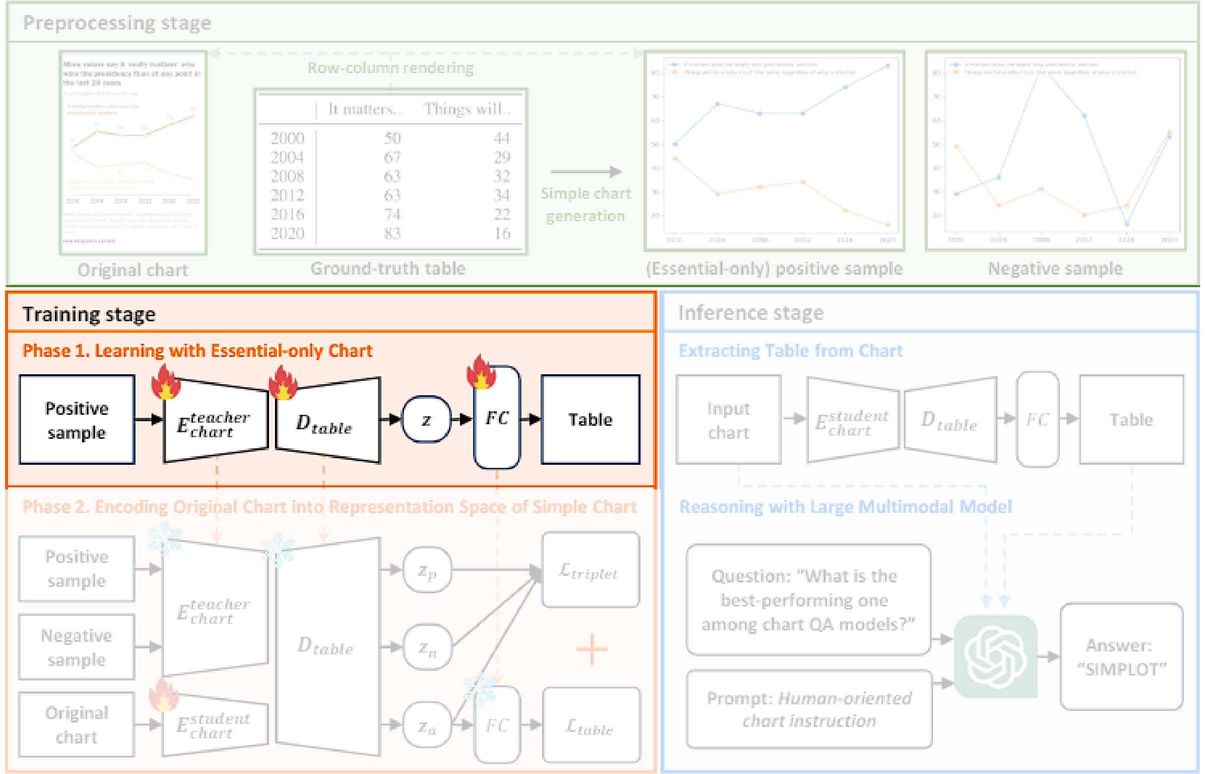
Generate data table of the figure below given the columns Value; and the rows Use SNS | Do not go online | Go Online, No SNS
A Number of columns are 2 and rows are 4

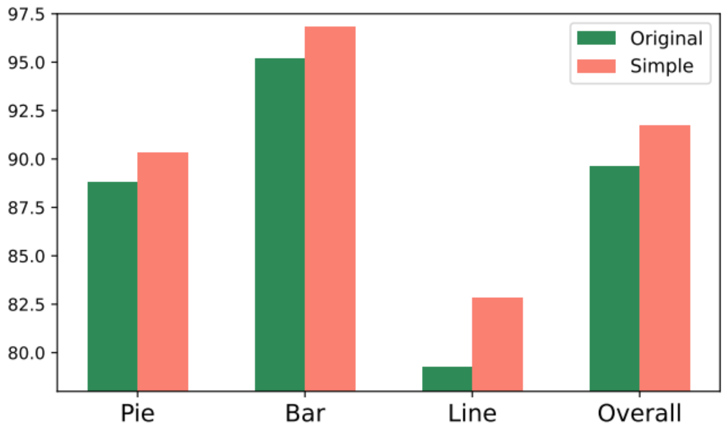


- Rendering **rows and columns** to effectively extract table

Method

Training Stage – Phase 1: Learning with Essential from Simple Chart



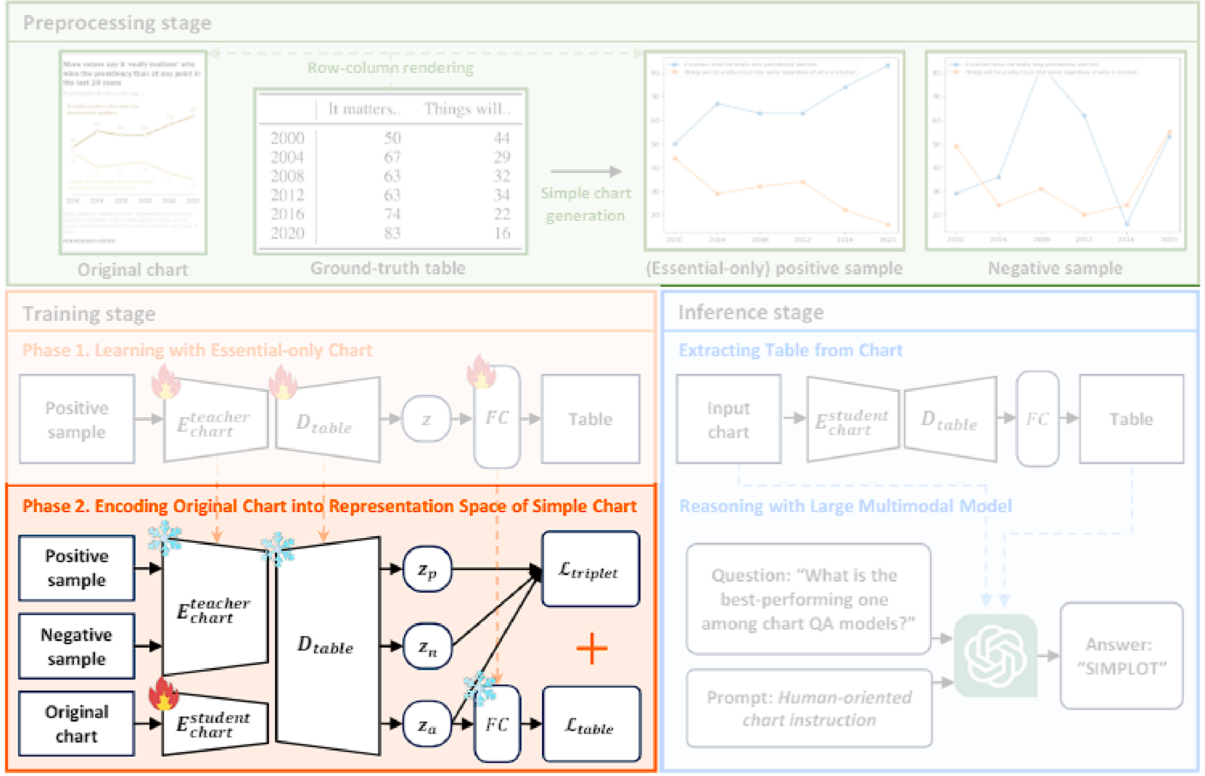


Comparison of table extraction performance using **original chart** vs **simple chart**

- Train the image encoder and text decoder to **generate the ground-truth** table from the generated **simple image**
- This trains the model to **extract only the essential information** from the chart
- The comparison of table extraction performance shows that **training with simple charts** results in better performance than using original charts

Method

Training Stage – Phase 2: Encoding Original Chart into Representation Space of Simple Chart



$$\mathcal{L}_{triplet}(A,P,N) = max\{d(z_a,z_p) - d(z_a,z_n) + m, 0\},$$

$$\text{where} \quad z_a = D_{table}(E^{student}_{chart}(A)), \quad z_p = D_{table}(E^{teacher}_{chart}(P)), \quad z_n = D_{table}(E^{teacher}_{chart}(N)).$$

Triplet Loss

$$T = [\hat{y}_1, \ldots, \hat{y}_N] = FC(z_a),$$

$$\mathcal{L}_{table} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\frac{\exp(\hat{y}_{i,c})}{\sum_{j=1}^{C}\exp(\hat{y}_{i,j})}\right)$$

Table Loss

$$\mathcal{L}_{final} = \lambda\mathcal{L}_{triplet} + (1-\lambda)\mathcal{L}_{table}.$$

Final Loss

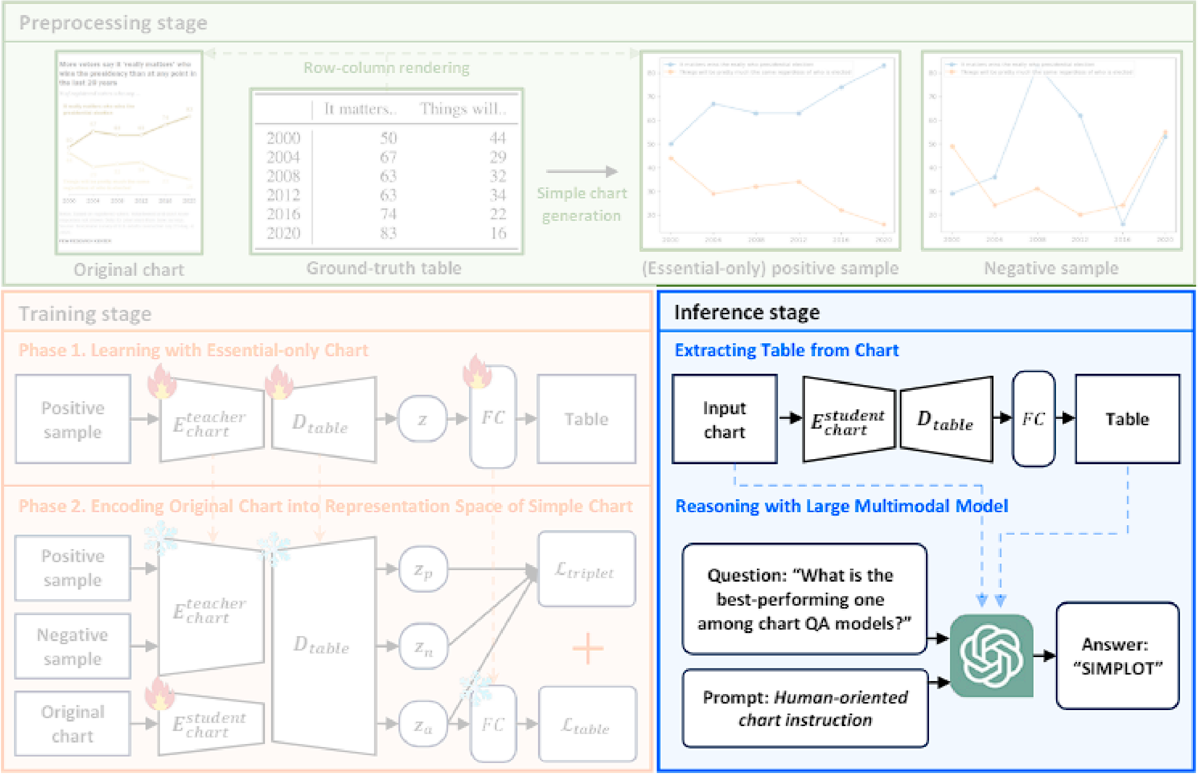
- Triplet loss: make the representation of the original chart similar to that of the simple chart representation

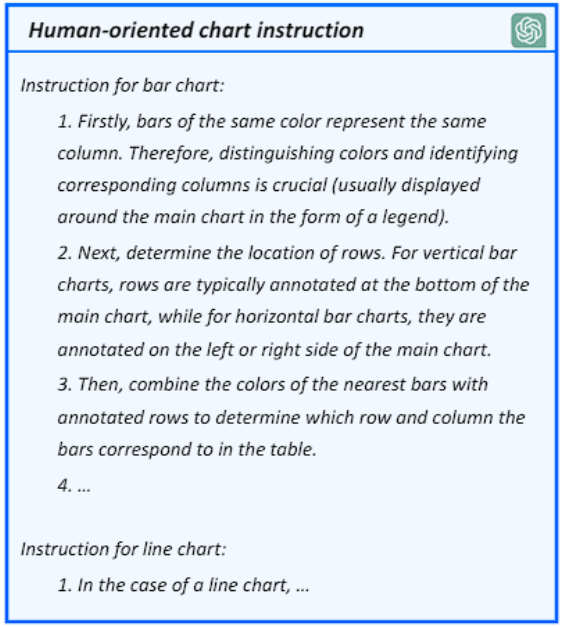
  → Extract only the representation of **essential information**

- Final loss: Triplet loss + Table loss

Method

Inference Stage – Reasoning with Extracted Table





Human-oriented chart instruction: prompt designed to mimic the human chart reasoning process

- Inference with **generated table and image**

  → Answer the question about **visual features** (position, color, etc.)

- Provide **chart-specific prompt** to enhance understanding of visual attributes and effectively align tables and charts

Experiments

Experimental Setting

- **Dataset**
  - ChartQA
  - PlotQA

split \ type	Pie	Bar	Line	QA pair
Train set	541	15,581	2,195	-
Validation set	48	837	171	-
Test set	78	1,230	211	2,500

ChartQA dataset statistics

split \ type	Dot line	Line	Bar	QA pair
Train set	26,010	25,897	105,163	-
Validation set	5,571	5,547	22,541	-
Test set	5,574	5,549	22,534	4,342,514

PlotQA dataset statistics

- **Compared Methods**
  - TaPas
  - V-TaPas
  - T5
  - VL-T5
  - PaLI
  - Mini-GPT
  - LLaVa
  - GPT-4V

  Vision Language Pretrained Models

  - ChartQA
  - ChartT5
  - Pix2Struct
  - MatCha
  - Unichart
  - ChartLlama

  Fully-supervised on QA task

  - Deplot
  - Unichart
  - **SIMPLOT**

  Utilize extracted table for QA

Experiments

Table Extraction Performance on ChartQA dataset

- Performance of chart to table extraction on the ChartQA dataset

  → Achieve state of the art performance across various chart types

  → Effectiveness of **Row-Column rendering, Simple chart**

Models	Chart type			Overall
	Pie	Bar	Line	
GPT-4V	<u>90.13</u>	91.53	71.51	84.24
UniChart	84.86	92.58	**85.16**	88.03
Deplot	88.82	<u>96.37</u>	82.25	<u>90.95</u>
SIMPLOT	**91.41**	**96.87**	<u>84.74</u>	**92.32**

Table Extraction Performance

Experiments

Chart Question Answering Performance

- Vision-language pretrained (VLP) models have limitations when handling charts

  → Demonstrating **the need for research targeting chart reasoning**

- Table extraction-based methods outperform supervised methods

  → **Table extraction and reasoning** through it are effective for QA

- Achieve SOTA performance among methods that utilize the extracted table

  → Effectiveness of **precise table extraction** and **proposed prompt**

- The performance on the Human type (complex questions) of ChartQA is overwhelming

  → Better performance as the **questions become more difficult**

	Models	Data type		
		Human	Augmented	Overall
VLP models	TaPas	28.72	53.84	41.28
	V-TaPas	29.60	61.44	45.52
	T5	25.12	56.96	41.04
	VL-T5	26.24	56.88	41.56
	PaLI	30.40	64.90	47.65
	Mini-GPT	8.40	15.60	12.00
	LLaVa	37.68	72.96	55.32
	GPT-4V	56.48	63.04	59.76
Supervised	ChartQA	40.08	63.60	51.84
	ChartT5	31.80	74.40	53.10
	Pix2Struct	30.50	81.60	56.05
	MatCha	38.20	90.20	64.20
	Unichart	43.92	88.56	66.24
	ChartLlama	48.96	90.36	69.66
	ChartAssistant	65.90	**93.90**	79.90
	ChartInstruct	45.52	87.76	66.64
Table	Deplot	62.71	78.63	70.67
	Unichart[2]	67.04	69.92	68.48
	SIMPLOT	**78.07**	88.42	**83.24**

QA Performance on ChartQA dataset

Models	Dot line	Line	Bar	Overall
GPT-4V	50.53	58.84	53.85	54.11
Unichart	58.78	53.26	60.10	58.74
Deplot	**66.66**	55.59	61.73	61.53
SIMPLOT	60.93	**65.57**	**73.84**	**70.32**

QA Performance on PlotQA dataset

Experiments

Ablation Study

- Confirm that **each component helps in accurately extracting the table**
- Proposed **prompt has a significant impact** on performance improvement and emphasized **the importance of task-specific prompts**

Row-col rendering	Simple chart	Prompt	RD_{F1}	RA
✗	✗	-	90.95	-
✓	✗	-	91.40	-
✗	✓	-	91.86	-
✓	✓	-	**92.32**	-
-	-	✗	-	79.79
-	-	✓	-	**83.24**

Ablation study for table extraction (upper) and QA (lower)

Experiments

Proposed Method is Model-agnostic

- SIMPLOT can enhance performance when combined with any model
- Confirm that combining with other models significantly improves both table extraction and question answering performance

→ **Prove the generality of the proposed method**

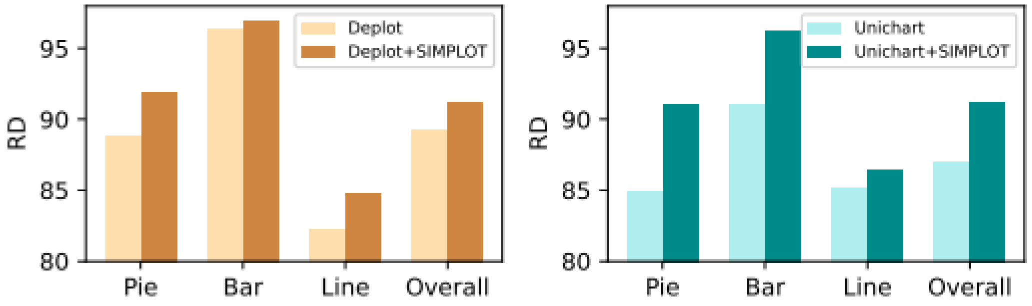
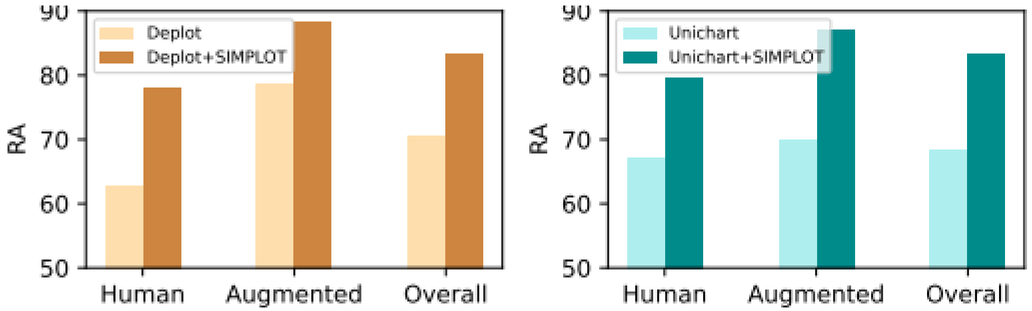


Table extraction performance of Deplot(left) and Unichart(right) with SIMPLOT applied



QA performance of Deplot(left) and Unichart(right) with SIMPLOT applied

Masry, Ahmed, et al. "Unichart: A universal vision-language pretrained model for chart comprehension and reasoning." *arXiv preprint arXiv:2305.14761* (2023).

Experiments

Further Analysis

- For a fair comparison, compared SIMPLOT without using prompts by using both the table and image generated by Deplot (left table)

  → **Accurately extracting the table improves QA performance**

- For a more strict comparison, compared the performance when applying the proposed prompt to Deplot as well (right table)

  → **Accurately extracting the table improves QA performance**

- Even if Deplot extracts the table inaccurately, there is a possibility of generating the correct answer as long as the question does not inquire about the extracted part.

  → For **harder questions that require more complex reasoning, a significant performance difference** was observed

Models	Human	Augmented	Overall
Unichart	67.04	69.92	68.48
Unichart + img.	75.04	**88.82**	81.93
Unichart + SIMPLOT w/o prompt	76.56	88.64	82.60
Unichart + SIMPLOT	**79.56**	87.18	**83.37**
Deplot	62.71	78.63	70.67
Deplot + img.	72.39	85.01	78.70
Deplot + SIMPLOT w/o prompt	73.91	85.67	79.79
Deplot + SIMPLOT	**76.70**	**88.42**	**82.56**

	Models	Human	Augmented	Overall
Easy	Deplot + img.	72.39	85.01	78.70
	Deplot + img. + prompt	77.75	88.30	83.03
	SIMPLOT	**78.07**	**88.42**	**83.24**
Hard	Deplot + img. + prompt	-	-	49.41
	SIMPLOT	-	-	**65.88**

Experiments

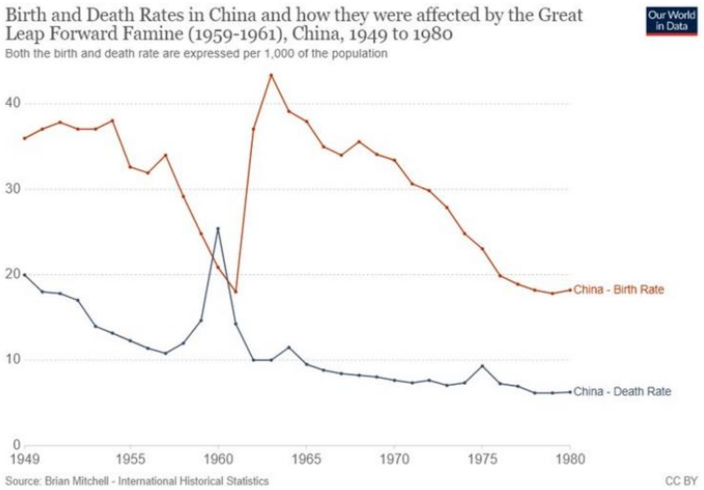
Example of Hard Question

	Models	Human	Augmented	Overall
Easy	Deplot + img.	72.39	85.01	78.70
	Deplot + img. + prompt	77.75	88.30	83.03
	SIMPLOT	**78.07**	**88.42**	**83.24**
Hard	Deplot + img. + prompt	-	-	49.41
	SIMPLOT	-	-	**65.88**

- Most of the existing questions are designed in a way that the answer can be derived by referencing just one row or column
- Using GPT, generate QA pairs that require referencing two or more rows and columns to answer

("Create a challenging question-answer pair that requires referencing at least two rows and two columns to solve.")

→ The **more complex the reasoning required**, the **more significant the performance improvement** from accurately extracted tables



Input Chart

Entity	1949	1955	1960	1965	1970	1975	1980
Birth Rate	37.8	nan	nan	nan	nan	nan	18.4
Death Rate	17.2	nan	nan	nan	nan	nan	6.5

Entity	1949	1955	1960	1965	1970	1975	1980
Birth Rate	35.9	32.5	25.3	37.7	33.5	24.3	18.2
Death Rate	18.2	12.2	14.8	9.6	7.4	9.4	6.2

Extracted table from Deplot (upper) & SIMPLOT (lower)

Question: What is the sum of the birth rate in China in 1955 and the death rate in China in 1965?? Correct Answer: 42.1

Deplot + img. + prompt

Explanation: For a line chart, the x-axis typically represents the rows (time periods in this case), and each line represents a different column (birth rate and death rate). The value of the line at a given year can be determined by the y-axis value that the line intersects. Since the question asks for specific years, we locate the points on the lines that correspond to the years 1955 and 1965 and sum the values of the birth rate in 1955 and the death rate in 1965.
Answer: 47

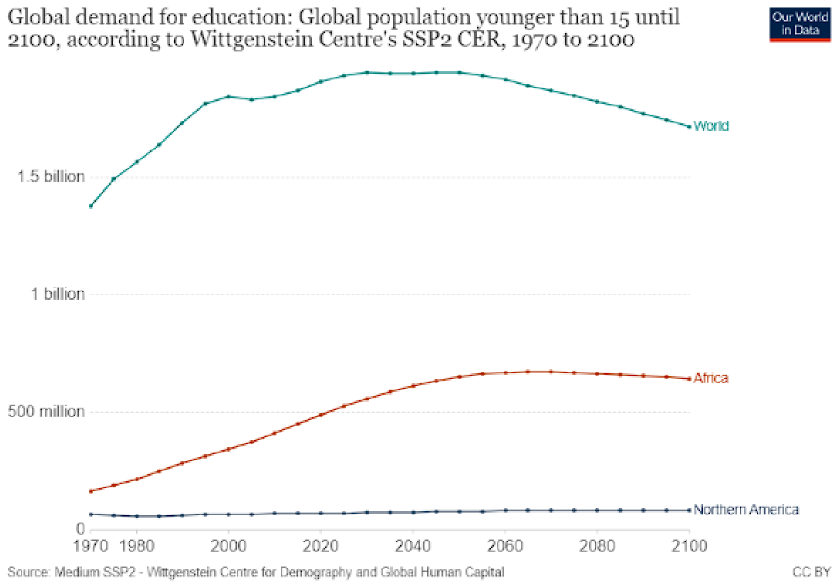
SIMPLOT

Explanation : To find the sum of the birth rate in China in 1955 and the death rate in China in 1965, we need to identify the values from the line chart. The x-axis represents the years, and the two lines represent the birth and death rates, respectively. We locate the year 1955 on the x-axis and find the corresponding point on the birth rate line, then do the same for the year 1965 on the death rate line. The y-axis provides the values for these points. Since the question is about specific years, we refer to the predicted table for the exact values for those years.
Answer: 42.1

QA explanation of Deplot + image + prompt & SIMPLOT

Experiments

Case Study – Table Extraction 1



Input Chart

Entity	1970	1980	2000	2020	2040	2060	2080	2100
World	1.37	1.55	1.83	1.93	1.93	1.89	1.79	1.71
Africa	190	200	350	479	604	665	650	637
Northern America	6	3.79	3.23	3.92	4.06	4.63	4.79	4.81

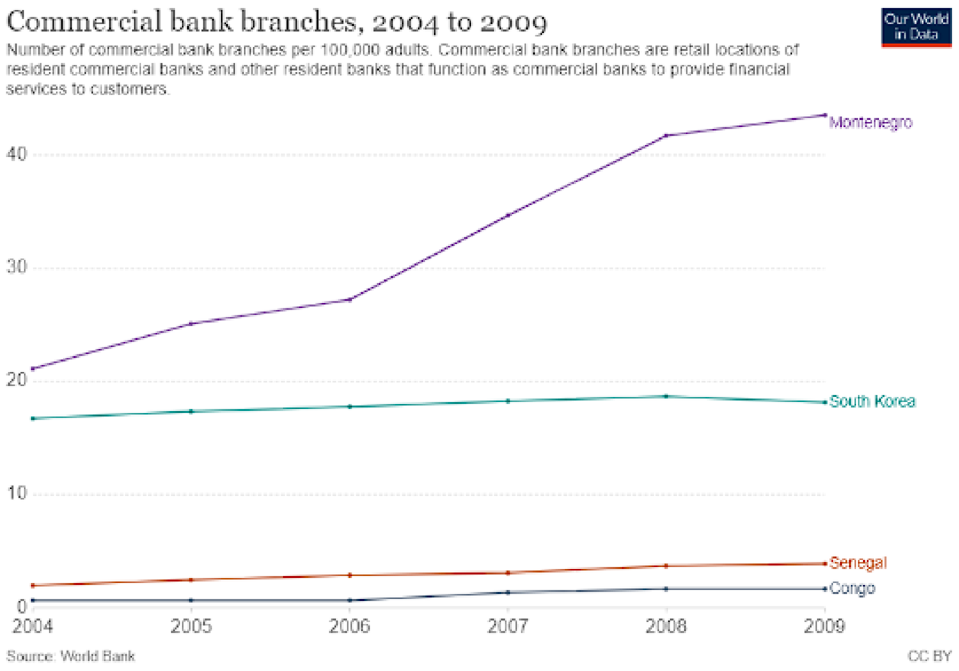
Entity	1970	1980	2000	2020	2040	2060	2080	2100
World	1388.75	1520.46	1830.13	1934.69	1952.49	1970.42	1712.03	1696.22
Africa	117.45	115.97	334.05	490.55	642.36	677.90	686.98	694.67
Northern America	0.06	0.05	0.059	0.05	0.05	0.06	0.06	0.06

Extracted table from Deplot (upper) & SIMPLOT (lower)

- Deplot confuses the unit of values, such as 'million/billion', leading to inaccurate table extraction
- SIMPLOT successfully generates a table closer to the ground truth by **incorporating textual information**

Experiments

Case Study – Table Extraction 2



Input Chart

Entity	Commercial bank branches, 2004 to 2009
South Korea	16.86
Congo	4.16
Senegal	3.12
Congo	3.7
Senegal	4.32
Como	4.0
CC BY	+ 1 missing row 4.19

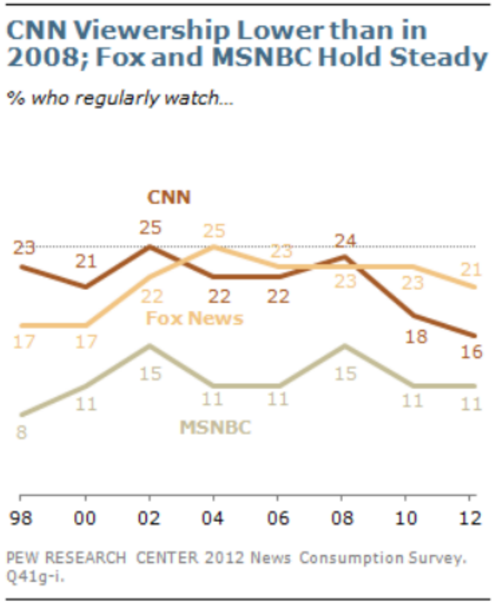
Entity	2004	2005	2006	2007	2008	2009
Montenegro	21.4	25.5	27.0	34.5	42.1	43.2
South Korea	16.8	17.0	17.9	18.3	18.7	18.2
Senegal	2.0	2.1	2.2	2.3	2.7	2.8
Congo	0.3	0.3	0.3	0.3	0.3	0.3

Extracted table from Deplot (upper) & SIMPLOT (lower)

- Deplot is heavily influenced by **unnecessary information** in the image, failing to extract the 'Montenegro' row and **adding inaccurate** rows like 'CC BY', resulting in table extraction failure
- SIMPLOT successfully generates a table closer to the ground truth while **extracting accurate row and columns** of chart

Experiments

Case Study – Chart Question Answering



Question: What is the % who regularly watch CNN in 02? Correct Answer: 25

SIMPLOT w/o prompt

Explanation: The percentage of people who regularly watch CNN in the year ’02 is found in the chart under the CNN column at the point where the year ’02 is marked. The chart shows that the value is 21%.
Answer: 21

SIMPLOT

Explanation: To answer the question, we must refer to the line chart. The x-axis represents the years, which correspond to the rows in the table. Each line represents a different news network, which corresponds to the columns in the table. The value for CNN in a specific year can be found by locating the point on the CNN line (red) that aligns with the year on the x-axis and then reading the value from the y-axis.
The chart shows the percentage of people who regularly watch CNN, Fox News, and MSNBC over various years. The red line represents CNN. In 2002, the red line is at the 25% mark on the y-axis.
Answer: 25

- Comparison of the QA explanation between SIMPLOT and the case without using a prompt to prove the effectiveness of proposed prompt
- While SIMPLOT without a prompt failed to derive the correct answer, SIMPLOT with the prompt effectively **mimicked the flow of human reasoning in chart interpretation**, leading to the correct answer

Thank you!

[Full Paper] https://arxiv.org/abs/2405.00021

[Source Code] https://github.com/sangwu99/Simplot

[Lab Homepage] https://dsail.kaist.ac.kr

[Email] wjkim@kaist.ac.kr
sangwu.park@kaist.ac.kr

KAIST DSAIL Data Science & Artificial Intelligence