

Final Report on Knowledge Extraction

Jing Wang, jwangke@connect.ust.hk

1. Introduction

Knowledge Extraction (KE) refers to the process of identifying and extracting structured information or knowledge from unstructured data sources, such as text, images, or videos. This information can include entities, relationships, concepts, or any structured data that can be used for further analysis. KE is a critical task in various applications, including knowledge graph construction, search engines, question answering, and domain-specific expert systems.

In recent years, Large Language Models (LLMs) such as GPT have emerged as powerful tools for KE due to their ability to understand and generate natural language with remarkable accuracy. However, LLMs face significant challenges when attempting to process highly complex unstructured data (like tables, equations, figures, citations) due to several inherent limitations in their design and capabilities. Their limitations particularly in accuracy and explicit knowledge representation highlight the need for hybrid approaches combining LLMs with rule-based systems, domain-specific fine-tuning, or structured databases.

This report evaluates the performance of several LLMs in table knowledge extraction and understanding. Recognizing the inherent limitations of LLMs in processing unstructured table data, the report introduces a preprocessing pipeline that utilizes Optical Character Recognition and rule-based methods to reconstruct tables into structured data formats, aiming to improve the accuracy and reliability of LLMs when handling tabular content.

2. Overview

Large Language Models (LLMs) like GPT-4 (ChatGPT), AskyourPDF, and Unstruct have demonstrated notable capabilities in extracting and understanding information from tables. However, complex tables often contain various advanced features such as multiple column items (columns that span more than one field), merged cells (span multiple rows or columns), parenthetical explanations, and superscripts with corresponding footnotes or annotations, all of which pose significant challenges for table extraction and understanding. These features demand a high level of precision and contextual awareness from extraction tools.

Here, I will first analyze their performance from two perspectives: locating and summarizing content from directories, and understanding complex table contents.

2.1 Ability of locating and summarizing

First, I uploaded the pdf file to GPT-4o. As shown in the right image, I asked GPT-4o two summary questions:

1. Summarize the ways to save money.
 2. The contents page has 9 subheadings from A to I. List these subheadings correctly and summarize the contents of each section.

The left image corresponds to the original PDF text. Based on GPT-4o's responses, we can see that it can locate the answers in the PDF and provide a relatively comprehensive summary of the original text.

YOU: uploaded 20231101-guide.pdf

 YOU: Answer based on the PDF document:

1. Summarize the ways to save money
 2. The contents page has 9 subheadings from A to I. List these subheadings correctly and summarize the contents of each section
 3. I have a Personal Integrated Account, how much does it cost me to have a cashier's check?

AI: #### 1. Ways to Save Money

- **Interbank Transfer:** Use HSBC Internet or Mobile Banking to waive the HK\$200 branch fee.
 - **Outward Telegraphic Transfer:** Fee reduced to HK\$50 via Internet or Mobile Banking.
 - **Cheque Charges:** Avoid insufficient funds charges by ensuring account balance or using overdraft protection.
 - **HSBC One Account:** No minimum balance requirement or below balance fee.
 - **Automatic Bill Payment:** Use Internet or Mobile Banking to avoid overdue charges.
 - **HSBC Jade and Premier Privileges:** Enjoy fee waivers or discounts on selected services.
 - **Credit Card Charges:** Waived annual fees for HSBC Jade/Premier customers.
 - **Paper Statement Fee:** Avoid the HK\$60 annual fee by opting for eStatements.

For the second question, involving locating and listing subheadings (as shown in the next two images of the original text), GPT-4o was able to answer all correctly. Additionally, I asked more questions that could be directly answered from the original text and requested that it specify the page numbers of the answers. GPT-4o managed to perform this task quite well.

Item	Page	Page
4.0 Merchant Credit Card	32	
1. Integrated account services	32	
2. Merchants services	32	
a. Merchants payment	32	
b. Merchants financing	32	
3. Credit card	34	
a. HSBC Prema-MasterCard Credit Card	34	
b. ATM card - repayment fee	34	
c. ATM withdrawal	34	
d. Overdraft	34	
e. Change of Multiple Plus	35	
f. Change of Multiple Plus Advance Debit	35	
g. Personal or business line of documents	35	
h. Personal or business line of documents	35	
i. Request for document copies	35	
4. Merchants financing	36	
5. General services - all accounts	36	
a. Cards change/hold/switch to miles	36	
b. Cards change/hold Hong Kong dollar rates	36	
c. Cards change/reload	36	
d. Dancing Internet - the registered mail	36	
e. ThruNet - the registered mail	36	
f. Request for document copies	36	
6. Merchants financing	36	
4. SPECIAL PRIVILEGES FOR HSBC JADE	36	
1. Integrated account services	36	
2. Merchants services	36	
a. Merchants payment	36	
b. Merchants financing	36	
3. Credit card	36	
a. HSBC Prema-MasterCard Credit Card	36	
b. ATM card - repayment fee	36	
c. ATM withdrawal	36	
d. Overdraft	36	
e. Change of Multiple Plus	36	
f. Change of Multiple Plus Advance Debit	36	
g. Personal or business line of documents	36	
h. Personal or business line of documents	36	
i. Request for document copies	36	
4. Merchants financing	36	
5. General services - all accounts	36	
a. Cards change/hold/switch to miles	36	
b. Cards change/hold Hong Kong dollar rates	36	
c. Cards change/reload	36	
d. Dancing Internet - the registered mail	36	
e. ThruNet - the registered mail	36	
f. Request for document copies	36	
4. SPECIAL PRIVILEGES FOR HSBC PREMIER	37	
1. Integrated account services	37	
2. Merchants services	37	
a. Merchants payment	37	
b. Merchants financing	37	
3. Credit card	37	
a. HSBC Prema-MasterCard Credit Card	37	
b. ATM card - repayment fee	37	
c. ATM withdrawal	37	
d. Overdraft	37	
e. Change of Multiple Plus	37	
f. Change of Multiple Plus Advance Debit	37	
g. Personal or business line of documents	37	
h. Personal or business line of documents	37	
i. Request for document copies	37	
4. Merchants financing	37	
5. General services - all accounts	37	
a. Cards change/hold/switch to miles	37	
b. Cards change/hold Hong Kong dollar rates	37	
c. Cards change/reload	37	
d. Dancing Internet - the registered mail	37	
e. ThruNet - the registered mail	37	
f. Request for document copies	37	
4. TERMINATION FOR HSBC ATM	38	
1. General services - all accounts	38	
a. ATM withdrawal	38	
b. ATM card - Hong Kong dollar rates	38	
c. ATM withdrawal	38	
d. ATM withdrawal	38	
e. ATM withdrawal	38	
f. ATM withdrawal	38	
g. ATM withdrawal	38	
h. ATM withdrawal	38	
i. ATM withdrawal	38	
2. General services - business services	38	

GPT-4o's responses:

2. Contents Page Subheadings and Summaries

- **A. General Services**
 - Covers fees for cashier's orders, coin services, bulk deposits, paper statements, and more.
- **B. Deposit Account/Services**
 - Details on stop payment orders, returned cheques, passbook replacements, and account closures.
- **C. Payment Services**
 - Information on autoPay services, remittances, global transfers, and interbank funds transfers.
- **D. Mortgage Services**
 - Fees related to mortgage applications, changes, and document requests.
- **E. Bundled Services**
 - Describes services for HSBC Jade, Premier, One, and Personal Integrated Accounts.
- **F. Loan Services – Personal Instalment Loan/Personal Tax Loan**
 - Charges for late payments, early repayment, and handling fees.
- **G. Cards**
 - Fees for personal credit cards, ATM cards, and other account cards.
- **H. Special Privileges for HSBC Jade**
 - Fee waivers and preferential offers for HSBC Jade customers.
- **I. Special Privileges for HSBC Premier**
 - Fee waivers and preferential offers for HSBC Premier customers.

2.2 Ability of table understanding

The third question above, "I have a Personal Integrated Account, how much does it cost me to have a cashier's check?", required GPT to locate the relevant table in the original text (Table A in the image below) and extract the answer from it.

A. General services

A1. General services – all accounts					
Item	Charge				
	Personal Customer	Personal Integrated Account	HSBC One	HSBC Premier	HSBC Jade
Cashier's order <ul style="list-style-type: none">• Issue/repurchase a cashier's order¹ (This charge will be waived if the repurchase of cashier's order follows the cheque clearing process)• Loss of a cashier's order• Additional fee for handling instructions not using Bank's standard form (where accepted by the Bank)	HK\$75	HK\$60	HK\$40	Waived	
	HK\$60 plus HK\$331 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued		HK\$60 plus HK\$331 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued	HK\$150	
Coin changing/withdrawal/exchange to notes Each bag of coins	HK\$2	HK\$2	HK\$1	Waived	
Coin paid-in (must be pre-sorted, per customer per day) <ul style="list-style-type: none">• Below 500 coins• 500 coins or more	Waived	2% of the full amount paid-in (minimum HK\$50)	2% of the full amount paid-in (minimum HK\$50)	1% of the full amount paid-in (minimum HK\$25)	Waived
Bulk cash deposit - Hong Kong dollar notes (per customer per day) <ul style="list-style-type: none">• Up to 200 pieces of notes• Over 200 pieces of notes	Waived	0.25% of the full amount deposited (minimum HK\$50)	0.25% of the full amount deposited (minimum HK\$50)	0.25% of the full amount deposited (minimum HK\$50)	Waived
Bulk cheque deposit (per customer per day) <ul style="list-style-type: none">• Up to 30 cheques• Over 30 cheques	Waived	HK\$1 per additional cheque ²	HK\$1 per additional cheque ²	Waived	Waived
Renminbi notes deposits/withdrawals	Waived				
Foreign currency notes deposits/withdrawals	Waived				
Gift cheque (per cheque) Via HSBC branches ³	HK\$10	HK\$8		Waived	
Paper statement⁴ To receive paper statements	HK\$60 per customer p.a.		HK\$60 per customer p.a.		
Safe deposit boxes Annual rental and other charges	1. Please refer to appropriate branch 2. Customer has to set up autopy from his/her HSBC account for the annual safe deposit box rental fee		1. Please refer to appropriate branch 2. Customer has to set up autopy from his/her HSBC Jade, HSBC Premier or HSBC One account for the annual safe deposit box rental fee		

¹ The following group of customers will be exempted from the fee: aged below 18 or 65 and above. Senior Citizen Card holders, recipients of Government's Comprehensive Social Security Allowance, recipients of Government's Disability Allowance and the physically disabled or visually impaired.

² Waived if the cheque was deposited through cheque deposit machines or other non branch channel.

³ Effective 1 January 2023, this standard charge applies to Personal Banking HK Dollar Current, SuperSave, HSBC Jade, HSBC One, Personal Integrated Account, University Student Account, Cash Card, Renminbi Savings, HK Dollar Statement Savings, Current/Statement Savings, Foreign Currency Current or any Credit Card account if account holder receive any paper statement in a full calendar year. The following group of customers will be exempt from the annual fee: aged below 18 or 65 and above, recipients of Government's Comprehensive Social Security Allowance, recipients of Government's Disability Allowance and the physically disabled or visually impaired.

As shown in the image, although GPT accurately identified the source table containing the answer, it encountered issues when asked to convert the table into text. Specifically:

1. It omitted text within parentheses,
2. Misaligned corresponding column items,
3. Failed to recognize superscripts, which prevented it from including important information tied to footnotes.

3. Cost for a Cashier's Check with a Personal Integrated Account

- **Cashier's Order Fee:** HK\$60 for issuing or repurchasing a cashier's order.

😊 YOU: Where did you find answers for the third question above?

🤖 AI: The answer to the third question is found in the section titled "A. General Services" under "A1. General services – all accounts" in the PDF document. It lists the charges for a cashier's order for different account types, including the Personal Integrated Account, which is HK\$60.

😊 YOU: Translate Cashier's order in A1 table to text

🤖 AI: In the A1 table under "General services – all accounts," the entry for "Cashier's order" includes the following details:

- **Issue/repurchase a cashier's order:** HK\$75 for Personal Customers, HK\$60 for Personal Integrated Accounts, HK\$40 for HSBC One, and waived for HSBC Premier and HSBC Jade customers.
- **Loss of a cashier's order:** HK\$60 plus HK\$33 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued.
- **Additional fee for handling instructions not using the Bank's standard form (where accepted by the Bank):** HK\$150.

This summarizes the costs associated with cashier's orders for different account types.

Even after I pointed out that it had missed the superscripts, GPT was still unable to correctly identify and include them in its response.

😊 YOU: You missed the footnotes when translating Item Cashier's order in A1 table to text

🤖 AI: Here's the translation of the "Cashier's order" item from the A1 table, including the footnotes:

- **Issue/repurchase a cashier's order:**
 - HK\$75 for Personal Customers
 - HK\$60 for Personal Integrated Accounts
 - HK\$40 for HSBC One
 - Waived for HSBC Premier and HSBC Jade customers.
- **Loss of a cashier's order:**
 - HK\$60 plus HK\$33 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued.
- **Additional fee for handling instructions not using the Bank's standard form (where accepted by the Bank):** HK\$150.

Footnotes:

1. This charge will be waived if the repurchase of the cashier's order follows the cheque clearing process.
2. The customer has to set up autopay from his/her HSBC account for the annual safe deposit box rental fee.

This includes all relevant details and footnotes regarding the cashier's order charges.

The following two images show the responses to the same question using the AskYourPDF plugin on the ChatGPT platform and the AskYourPDF official website. Their performance and issues are similar to those observed with GPT.

These issues demonstrate that even tools specifically designed for document parsing, like AskYourPDF, share similar limitations when handling complex table structures.

The top screenshot shows the AskYourPDF Research Assistant interface. It displays a PDF file titled "20231101-guide.pdf" and a summary of the document's contents. The summary includes:

Answer based on the PDF document:
1. Summarize the ways to save money
2. The contents page has 9 subheadings from A to I. List these subheadings correctly and summarize the contents of each section
3. I have a Personal Integrated Account, how much does it cost me to have a cashier's check?

The bottom screenshot shows a web browser displaying the PDF "20231101-guide.pdf". The PDF is titled "Bank tariff guide for HSBC Wealth and Personal Banking Customers" and is effective from 1 November 2023. It contains sections on "Issuance/Repurchase Charges" and "Loss of Cashier's Order".

2.3 Analyzing Unstruct

Unstruct is one of most popular knowledge extraction tools, which enables user to add documents to no-code Prompt Studio and do prompt engineering to extract required fields and ask questions based on the documents. Unstruct official website has another free tool called LLMWhisperer that presents data from complex documents to LLMs in a way that they can best understand.

As shown in the following two images, LLMWhisperer converts Table A from the PDF mentioned above into texts. These texts mostly reconstruct the original table's content and relative positioning correctly, with the following minor errors:

The title "personal integrated account" is misaligned.

Some multi-line segments contain recognition errors, including "HK\$60 plus HK\$331~on circulars issued", "2% of the~HK\$50)", and "1% of the~HK\$25)".

The second instance of "2. customer has to~rental fee" is not correctly placed under the "HSBC Jade" column header.

However, all superscripts indicating footnotes are successfully recognized.

A. General services	Personal Customer	Personal Account Integrated	Customer	Personal Account Integrated	HSBC One	HSBC Premier	HSBC Jade
A1. General services - all accounts							
Item							
Cashier's order							
- Issue/repurchase a cashier's order (This charge will be waived if the repurchase of cashier's order follows the cheque clearing process)							
- Loss of a cashier's order							
- Additional fee for handling instructions not using Bank's standard form (where accepted by the Bank)							
Coins changing/withdrawal/exchange to notes							
Each bag of coins							
Coins paid-in							
(must be pre-sorted, per customer per day)							
- Below 500 coins							
. 500 coins or more							
Bulk cash deposit - Hong Kong dollar notes (per customer per day)							
- Up to 200 pieces of notes							
Over 200 pieces of notes							
Bulk cheque deposit (per customer per day)							
- Up to 30 cheques							
. Over 30 cheques							
Renminbi notes deposits/withdrawals							
Foreign currency notes deposits/withdrawals							
Gift cheque (per cheque)							
Via HSBC branches1							
Paper statement3							
To receive paper statements							
Safe deposit boxes							
Annual rental and other charges							
1. The following group of customers will be exempted from the fee: aged below 18 or 65 and above, Senior Citizens Card holders, recipients of Government's Comprehensive Social Security Allowance, recipients of Government's Disability Allowance and the visually impaired.							
2. Waived if the cheques are deposited through cheque deposit machines or other non branch counter channels.							
3. Effective 1 January 2023, this standard charge applies to Personal Banking HK Dollar Current, SuperSaver, HSBC One, Personal Integrated Account, Cash Card, Remittance Savings, HK Dollar One integrated Account, University Student Account, Cash Card, Remittance Savings, HK Dollar Statement Savings, Foreign Currency Current or any Credit Card account if account holders receive any paper statement in a calendar year. The following group of customers will be exempted from the annual fee: aged below 18 or 65 and above, recipients of Government's Disability Allowance and the physically disabled or visually impaired.							

The following two images show working pipeline on the PDF with the same prompts mentioned above:

Left in trial: 14 days / 662.4K LLM Tokens

SinglePass Extraction

Document Parser Combined Output

PDF View Raw View Summary View 20231101-guide.pdf Manage Documents

A. General services

A1. General services - all accounts

Item

Cashier's order

- Issue/repurchase a cashier's order
(This charge will be waived if the repurchase of cashier's order follows the cheque clearing process)
- Loss of a cashier's order
- Additional fee for handling instructions not using Bank's standard form (where accepted by the Bank)
- Coins changing/withdrawal/exchange to notes
- Each bag of coins
- Coins paid-in
(must be pre-sorted, per customer per day)
- Below 500 coins
- . 500 coins or more
- Bulk cash deposit - Hong Kong dollar notes (per customer per day)
- Up to 200 pieces of notes
- Over 200 pieces of notes
- Bulk cheque deposit (per customer per day)
- Up to 30 cheques
- . Over 30 cheques
- Renminbi notes deposits/withdrawals
- Foreign currency notes deposits/withdrawals
- Gift cheque (per cheque)
- Via HSBC branches1
- Paper statement3
- To receive paper statements
- Safe deposit boxes
- Annual rental and other charges

Left in trial: 14 days / 662.4K LLM Tokens

SinglePass Extraction

Document Parser Combined Output

PDF View Raw View Summary View 20231101-guide.pdf Manage Documents

A. General services

A1. General services - all accounts

Item

Cashier's order

- Issue/repurchase a cashier's order
(This charge will be waived if the repurchase of cashier's order follows the cheque clearing process)
- Loss of a cashier's order
- Additional fee for handling instructions not using Bank's standard form (where accepted by the Bank)
- Coins changing/withdrawal/exchange to notes
- Each bag of coins
- Coins paid-in
(must be pre-sorted, per customer per day)
- Below 500 coins
- . 500 coins or more
- Bulk cash deposit - Hong Kong dollar notes (per customer per day)
- Up to 200 pieces of notes
- Over 200 pieces of notes
- Bulk cheque deposit (per customer per day)
- Up to 30 cheques
- . Over 30 cheques
- Renminbi notes deposits/withdrawals
- Foreign currency notes deposits/withdrawals
- Gift cheque (per cheque)
- Via HSBC branches1
- Paper statement3
- To receive paper statements
- Safe deposit boxes
- Annual rental and other charges

The raw view window, similar to LLMWhisperer, converts Table A into texts. However, raw view performs better:

It does not have typographical errors and successfully recognizes superscripts.

It still has minor issues with long text segments not fully aligning with the corresponding column headers (i.e., these segments should span multiple column headers but are instead too narrow compared to the original table).

On the other hand, Unstract still performs poorly in answering questions, exhibiting the same issues as ChatGPT and AskyourPDF:

Although it can locate the answers in the PDF and provide a relatively comprehensive summary of the original text, it fails to match column items to their respective column headers, leading to misalignment and missing entries.

It still does not capture footnotes indicated by superscripts.

However, Unstract correctly recognizes text within parentheses.

3. Improvement

To assist LLMs in accurately understanding table content and answering related questions, I manually developed code that takes a screenshot of a table as input and converts it into a CSV file based on the relative positional relationships of all text coordinates in the table.

By uploading this CSV file to the LLM, the model is able to successfully complete the task, as the structured CSV format eliminates ambiguities caused by the complexity of the original table layout. This method bridges the gap between complex table structures and LLM processing capabilities, significantly improving accuracy and reliability.

The code utilizes the pytesseract and cv2 libraries and consists of four main steps:

1. Extract and Save Text with Coordinates from the Image
2. Merge Text into Blocks Based on Row Distance
3. Merge Text Blocks into Cells Based on Column Distance
4. Reconstruct and Output the Original Table

3.1 Extract Text with Coordinates from Image

Using pytesseract, the text and their bounding box coordinates are extracted from the input image. These raw coordinates are saved for further processing. The specific workflow of this code is:

1. Read the Image
2. Convert the Image to Grayscale and Apply Binarization (enhance text regions and improve OCR accuracy)
3. Extract Text and its Bounding Box Coordinates Using Tesseract OCR (filter out blank or irrelevant text)
4. Save Extracted Text and Coordinates to a File

The code is as follows:

```
'# pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'

image_path = '/Users/yingmanyouyu/Desktop/biao.png'
output_path='/Users/yingmanyouyu/Desktop/1.csv'
image = cv2.imread(image_path)

gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
_, binary = cv2.threshold(gray, 150, 255, cv2.THRESH_BINARY)

custom_config = r'--oem 3 --psm 6'
results = pytesseract.image_to_data(binary, config=custom_config, output_type=pytesseract.Output.DICT)

data = pd.DataFrame({
    "text": results["text"],
    "left": results["left"],
    "top": results["top"],
    "width": results["width"],
    "height": results["height"]
})

data = data[data["text"].str.strip() != ""]

data["right"] = data["left"] + data["width"]
data["bottom"] = data["top"] + data["height"]

data["center_x"] = data["left"] + data["width"] / 2
data["center_y"] = data["top"] + data["height"] / 2

data.to_csv(output_path, index=False)
print(f"{output_path}")

for i, row in data.iterrows():
    x1, y1, x2, y2 = row["left"], row["top"], row["right"], row["bottom"]
    cv2.rectangle(image, (x1, y1), (x2, y2), (0, 255, 0), 2)

cv2.imshow("Bounding Boxes", image)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

The output 1.csv:

A	B	C	D	E	F	G	H	I
text	left	top	width	height	right	bottom	center_x	center_y
o_o	2612	0	34	7	2646	7	2629	3.5
A.	11	10	42	35	53	45	32	27.5
General	74	9	161	37	235	46	154.5	27.5
services	256	20	172	26	428	46	342	33
A1.	19	82	39	18	58	100	38.5	91
General	66	82	86	19	152	101	109	91.5
services	160	82	91	19	251	101	205.5	91.5
—	257	92	14	3	271	95	264	93.5
all	278	82	24	19	302	101	290	91.5
accounts	310	84	103	17	413	101	361.5	92.5
Item	27	127	52	19	79	146	53	136.5
Charge	1776	123	84	23	1860	146	1818	134.5
Personal	1056	172	99	19	1155	191	1105.5	181.5
Customer	1165	172	117	19	1282	191	1223.5	181.5
Personal	1390	156	96	51	1486	207	1438	181.5
Integrated	1500	160	122	43	1622	203	1561	181.5
	1659	154	1	57	1660	211	1659.5	182.5
HSBC	1780	172	68	19	1848	191	1814	181.5
One	1857	172	46	19	1903	191	1880	181.5
	2023	154	1	57	2024	211	2023.5	182.5
HSBC	2087	172	67	19	2154	191	2120.5	181.5
Premier	2164	172	93	18	2257	190	2210.5	181
	2318	154	1	57	2319	211	2318.5	182.5
HSBC	2420	172	67	19	2487	191	2453.5	181.5
Jade	2495	172	57	19	2552	191	2523.5	181.5
Cashier's	11	221	110	18	121	239	66	230
order	131	221	63	18	194	239	162.5	230
€	12	263	9	9	21	272	16.5	267.5
Issue/repur	32	258	183	23	215	281	123.5	269.5
a	224	263	10	13	234	276	229	269.5
cashier's	244	258	94	18	338	276	291	267
order'	346	254	62	22	408	276	377	265
(This	32	282	48	22	80	304	56	293

I also used cv2 to visualize the bounding boxes drawn by Tesseract in order to inspect the OCR extraction results and manually correct any errors, as shown in the image below.

General services

General services - all accounts		Charge			
		Personal Customer	Personal Integrated Account	HSBC One	HSBC Premier
tier's order					
we/repurchase a cashier's order		HK\$75		HK\$60	HK\$40
its charge will be waived if the repurchase of cashier's order follows the cheque clearing process					Waived
ss of a cashier's order		HK\$60 plus HK\$31 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued		HK\$60 plus HK\$31 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued	
ditional fee for handling instructions not using Bank's standard form (where adopted by the Bank)		HK\$150		HK\$150	
is changing/withdrawal/exchange to notes					
bag of coins		HK\$2		HK\$2	
is stand-in					
st be pre-sorted per customer per day					
low HK\$ coins		Waived		Waived	
1 coins or more		2% of the full amount paid-in (minimum HK\$50)	2% of the full amount paid-in (minimum HK\$50)	2% of the full amount paid-in (minimum HK\$25)	2% of the full amount paid-in (minimum HK\$25)
i cash deposit - Hong Kong dollar notes (per customer per day)					
to 200 pieces of notes		Waived		Waived	
at 200 pieces of notes		0.25% of the full amount deposited (minimum HK\$50)		0.25% of the full amount deposited (minimum HK\$50)	
i cheque deposit (per customer per day)					
to 80 cheques		Waived		Waived	
or 80 cheques		HK\$1 per additional cheque		HK\$1 per additional cheque	
minil notes deposits/withdrawals		Waived		Waived	
ign currency notes deposits/withdrawals		Waived		Waived	
cheque (per cheque)					
ISBU branches		HK\$10		HK\$8	
or statement					
cheque paper Statements		HK\$60 per customer p.a.		HK\$60 per customer p.a.	
i deposit boxes					
or rental and other charges		Please refer to appropriate branch Customer has to set up egypt from his/her HSBC Jade, HSBC Premier or HSBC One account for the annual safe deposit box rental fee		Please refer to appropriate branch Customer has to set up egypt from his/her HSBC Jade, HSBC Premier or HSBC One account for the annual safe deposit box rental fee	

Here are a few Tesseract recognition defects that require manual corrections:

Incorrect symbols (e.g., "o_o"),

Missing text (e.g., the word "Account" after "Integrated," or missing row values such as HK\$2, HK\$2, HK\$1, Waived from the "notes" section),

Incorrect recognition of superscripts (e.g., "order'"),

Poor quality of boxes leading to incorrect alignment, such as the box of next line being higher than that of the previous one, or boxes covering two lines of text leading to incorrect recognition (e.g., "additional" being misrecognized as "aatonied"),

Issues with column separators, like the vertical bar "|" being recognized in some places but not in others,

Variations in the height of bounding boxes for certain characters, which causes issues in subsequent processing (e.g., the missing vertical bar | at the right of the "Personal Customer" entry resulted in the HK\$75 and HK\$10 values not being matched to any columns below).

Additionally, I calculated the minimum distance between row and column items (57 and 8 for Table A) and separately handled the correctly recognized footnotes for easier processing later.

3.2 Merge Text into Blocks Based on Row Distance

Text elements that are close to each other in the row direction are merged into text blocks. The merged blocks and their updated bounding coordinates are saved.

Starting from the leftmost character, I continuously merge it with characters located on the same line and with a distance smaller than the row item distance. The specific criteria for merging are: if the boundary distance between two characters is less than 57 and their height ranges overlap.

The code is as follows:

```
df = pd.read_csv("2.csv")
texts = []
visited = set()

for i in range(len(df)):
    if i in visited:
        continue

    current_text = [df.loc[i, "text"]]
    current_top = df.loc[i, "top"]
    current_bottom = df.loc[i, "bottom"]
    visited.add(i)

    current_left = df.loc[i, "left"]
    current_right = df.loc[i, "right"]

    for j in range(i + 1, len(df)):
        if j in visited:
            continue

        next_left = df.loc[j, "left"]
        next_right = df.loc[j, "right"]
        next_top = df.loc[j, "top"]
        next_bottom = df.loc[j, "bottom"]

        if (0 < next_left - current_right < 57) and ((next_top == current_bottom and next_top >= current_top) or (current_top == next_bottom <= current_bottom) or (next_top == current_top and next_bottom <= current_bottom) or (next_top <= current_top and next_bottom >= current_bottom)):

            current_text.append("-")
            current_text.append(df.loc[j, "text"])

            current_top = min(current_top, next_top)
            current_bottom = max(current_bottom, next_bottom)
            current_left = min(current_left, next_left)
            current_right = max(current_right, next_right)
            visited.add(j)

    texts.append(
        {
            "text": "-".join(current_text),
            "top": current_top,
            "bottom": current_bottom,
            "left": current_left,
            "right": current_right
        }
    )

output_file = "merged_texts.csv"
result_df = pd.DataFrame(texts)
result_df.to_csv(output_file, index=False, encoding="utf-8")

print(f"Output file: {output_file}")
```

3.3 Merge Text Blocks into Cells Based on Column Distance

Text blocks are then grouped into cells by merging them column-wise. The resulting cells, along with their final boundary coordinates, are saved.

The specific operation involves starting from the topmost text block and continuously merging it with adjacent blocks that are within a distance greater than -4 (due to poor quality of the bounding boxes causing the box for the next line to be higher than the previous one) and less than 8. Additionally, the width range of the two blocks must overlap for them to be merged.

The code is as follows:

```
df = pd.read_csv("merged_texts.csv")  
  
merged_texts = []  
visited = set()  
  
for i in range(len(df)):  
    if i in visited:  
        continue  
  
    current_texts = [df.loc[i, "text"]]  
    current_top = df.loc[i, "top"]  
    current_bottom = df.loc[i, "bottom"]  
    current_left = df.loc[i, "left"]  
    current_right = df.loc[i, "right"]  
    visited.add(i)  
  
    for j in range(i + 1, len(df)):  
        if j in visited:  
            continue  
  
        next_top = df.loc[j, "top"]  
        next_bottom = df.loc[j, "bottom"]  
        next_left = df.loc[j, "left"]  
        next_right = df.loc[j, "right"]  
  
        if (-4 < next_top - current_bottom < 8) and ((next_left <= current_right and next_left >= current_left) or (current_left <= next_right <= current_right) or (next_left >= current_left and next_right <= current_right) or (next_left <= current_left and next_right >= current_right)):  
            current_texts.append(df.loc[j, "text"])  
            current_top = min(current_top, next_top)  
            current_bottom = max(current_bottom, next_bottom)  
            current_left = min(current_left, next_left)  
            current_right = max(current_right, next_right)  
            visited.add(j)  
  
    merged_texts.append(  
        {"text": "\n".join(current_texts),  
         "top": current_top,  
         "bottom": current_bottom,  
         "left": current_left,  
         "right": current_right  
    })  
  
output_file = "final_merged_texts.csv"  
result_df = pd.DataFrame(merged_texts)  
result_df.to_csv(output_file, index=False, encoding="utf-8")  
print(f"(output_file)")
```

3.4 Reconstruct and Output the Original Table

Here, I manually selected "Item," "Personal Customer," "Personal Integrated Account," "HSBC One," "HSBC Premier," and "HSBC Jade" as the column headers.

Using the relative positions of the cells, the table is reconstructed by determining the row and column items. The reconstructed table is then output in CSV format, preserving the original structure.

The specific operation involves considering all cells that have overlapping width ranges with the aforementioned column headers as belonging to that column (a single item may belong to multiple columns, so it will be displayed multiple times in the output table). Additionally, cells with overlapping height ranges are treated as part of the same row.

Furthermore, I also merge the vertical bar (|) following the column headers with the header itself to expand the width range, ensuring a better match for the full column items (e.g., the vertical bar after "HSBC One").

The code is as follows:

```

df = pd.read_csv("final_merged_texts.csv")

columns = [
    {"name": "Item", "top": 127, "bottom": 146, "left": 27, "right": 79},
    {"name": "Personal Customer", "top": 172, "bottom": 191, "left": 1056, "right": 1282},
    {"name": "Personal Integrated Account", "top": 154, "bottom": 211, "left": 1390, "right": 1660},
    {"name": "HSBC One", "top": 172, "bottom": 191, "left": 1780, "right": 2024},
    {"name": "HSBC Premier", "top": 172, "bottom": 191, "left": 2887, "right": 2319},
    {"name": "HSBC Jade", "top": 172, "bottom": 191, "left": 2420, "right": 2552},
]

if 'column' not in df.columns:
    df['column'] = None
for index, row in df.iterrows():
    if index in range(0, 9):
        continue
    left, right = row["left"], row["right"]
    col_id = assigned_column = False
    for col_index, col in enumerate(columns):
        if (left <= col["right"] and left >= col["left"]) or (col["left"] <= right <= col["right"]) or (left >= col["left"] and right <= col["right"]) or (left <= col["left"] and right >= col["right"]):
            col_id.append(col_index)
            assigned_column = True
    df.at[index, "column"] = col_id if assigned_column else None
    print(f"<row{text}> {col_id}")
if not assigned_column:
    print(f"<row{text}> ")
rows = []
visited = set()

for i in range(9, len(df)):
    if i in visited:
        continue
    current_row = [""] * len(columns)
    assigned_columns = df.loc[i, "column"]

    if assigned_columns:
        for col in assigned_columns:
            current_row[col] = df.loc[i, "text"]

    current_top = df.loc[i, "top"]
    current_bottom = df.loc[i, "bottom"]
    visited.add(i)

    for j in range(i + 1, len(df)):
        if j in visited:
            continue
        next_top = df.loc[j, "top"]
        next_bottom = df.loc[j, "bottom"]

        if not (next_top > current_bottom or next_bottom < current_top):
            assigned_columns = df.loc[j, "column"]

            if assigned_columns:
                for col in assigned_columns:
                    current_row[col] = df.loc[j, "text"]

            current_top = min(current_top, next_top)
            current_bottom = max(current_bottom, next_bottom)
            visited.add(j)

    rows.append(current_row)

output_df = pd.DataFrame(rows, columns=[col["name"] for col in columns])
output_file = "organized_table.csv"
output_df.to_csv(output_file, index=False, encoding="utf-8")
print(f"<{output_file}>")

```

The output table is:

Item	Personal C	Personal In	HSBC One	HSBC Pre	HSBC Jade
Cashier's order					
Issue/repurchase a cashier's order	1 HK\$60	HK\$40	Waived		
Loss of a co	HK\$60 plus	HK\$60 plus	HK\$60 plus	HK\$331 collected on behalf of Hong Kong Interbank Clearing Limited on circulars issued	
Additional	HK\$150		HK\$150		
Coins changing/withdrawal/exchange to notes each bag of coins (HK\$2, HK\$2, HK\$1, Waived)					
Coins paid-in (must be pre-sorted, per customer per day)					
Below 500 c	Waived		Waived		
500 coins or 2% of the fi 2% of the fi 1% of the fi Waived					
Bulk cash deposit - Hong Kong dollar notes (per customer per day)					
Up to 200 p	Waived		Waived		
Over 200 p	0.25%	of th 0.25%	of th 0.25%	of th 0.25% of the full amount deposited (minimum HK\$50)	
Bulk cheque deposit (per customer per day)					
Up to 30 ch	Waived		Waived		
Over 30 che	HK\$1 per a	HK\$1 per a	HK\$1 per a	HK\$1 per additional cheque	
Renminbi notes deposits/withdrawals		Waived	(Waived)		
Foreign currency notes deposits/withdrawals		Waived	(Waived)		
Gift cheque (per cheque) Via HSBC HK\$8		Waived			
Paper statement3					
To receive 1 HK\$60 per	HK\$60 per	HK\$60 per	HK\$60 per	HK\$60 per customer p.a.	
Safe deposit boxes					
Annual ren 1 Please ref 1 Please ref 1 Please ref 1 Please refer to appropriate branch					
2 Customer 2 Customer 2 Customer 2 Customer 2 Customer has to set up autopay from his/her HSBC Jade, HSBC Premier or HSBC One account for the annual safe deposit box rental					

The output table successfully includes and corresponds to all text content accurately. However, due to the complexity or irregularity of table items, some entries are missing, and such issues are difficult to resolve with my code. For example, in Table A, the HK\$75 for "Issue/repurchase a cashier's order" falls exactly between "Personal Customer" and "Personal Integrated Account," making it impossible to match with any column headers (this is also due to Tesseract's instability, which failed to recognize the vertical bar | between the two, preventing the accurate representation of the column header's width

range). Additionally, the HK\$150 for "Additional fee for handling instructions not using Bank's standard form" in another row is only matched with "Personal Customer" and "HSBC Premier" because their actual width ranges only overlap with those two columns. In reality, it should belong to five column items.

While humans can accurately determine the row and column relationships by using the background color or surrounding divider lines of an item, Tesseract fails to accurately extract boundary lines, leading to missing situations where the width ranges of column items does not strictly cover the corresponding column header ranges, as mentioned above.

4. Limitations and Future Works

As mentioned earlier, in my workflow, Tesseract can make unpredictable small errors, such as low-quality boxes, misidentified characters and superscripts, or omitting important characters like divider lines. Low-quality boxes may result in the inability to merge characters that should belong to the same text block or create trouble when accurately calculating row and column item distances. The latter two errors can cause larger issues in the final table reconstruction and LLM table understanding, such as failing to accurately identify the width ranges of column headers, leading to missing entries, particularly in situations where the width ranges of column items does not strictly cover the corresponding column header ranges. Moreover, the failure to properly represent superscripts makes it difficult for the LLM to match the corresponding footnotes. In fact, I have tried using cv2's automatic line detection and cell segmentation functions, and mapping rows and columns based on the coordinates of the cells, but the result was worse. Improvements can focus on enhancing OCR's recognition and bounding box accuracy, as well as its ability to detect table divider lines or different background colors. It would also be beneficial to continue experimenting with the performance of other related open-source libraries.

Additionally, I manually calculated row and column item distances, manually locate footnotes, and manually identify column headers. This is because the distance between two items and the distance between two characters in a text block have no inherent relationship. Also, due to the possibility of multiple column headers (columns that span more than one field), it is very challenging to automate the identification of corresponding distance values and column headers. This remains a key challenge for future work.

5. Acknowledgement

I would like to thank Professor Lin for his careful guidance in this Independent Project, where I learned the big picture of knowledge extraction and table comprehension by LLM. His guidance not only gave me a deeper understanding in this field, but also inspired my research.

6. Appendix

Minutes of the 1st Project Meeting

Date Wednesday, September 11,
2024

Time 2:30 pm

Place Room 3557

Present Jing Wang

Prof Fangzhen Lin

Note-taker Jing Wang

Discussion items

The professor introduced the project and assigned tasks:

Survey apps/tools that can accept a PDF file and answer questions based on it, including GPT4 (ChatGPT), Askyourpdf (askyourpdf.com), unstruct and many apps and plugins on OpenAI platform.

For these tools, test their abilities in answering questions using the HSBC fee document that includes tables, footnotes and other features. To do this, compile a sample questions such as “I have a Personal Integrated Account, how much does it cost me to have a cashier’s check?”, and analyse the answers by the tools.

Survey apps/tools that can convert a PDF file to a text file, and translate a table in the HSBC document to a text format like the following (Table A1 in the HSBC document).

I talked about some tools that I found before can convert PDF into texts, jsons and htmls.

Meeting adjournment and next meeting

The meeting adjourned at 3:00 pm. The next meeting will be held in October.

Minutes of the 2nd Project Meeting

Date Tuesday, October 8, 2024

Time 2:00 pm

Place Room 3557

Present Jing Wang

Prof Fangzhen Lin

Note-taker Jing Wang

Discussion items

I reported on my project progress:

I tested the ability of GPT-4o, Askyourpdf (askyourpdf.com), askyourpdf plugin on OpenAI platform to answer questions based on the HSBC fee document. I found that all these tools performed well in summarizing the full text and identifying key topics, but only partially answered the questions correctly. In particular, these tools made mistakes when they needed to extract information from complex tables such as A1 or convert them into text.

Then, the professor assigned the next task: use OCR and relevant techniques to read the characters in the table, determine the row and column items of the table by their relative coordinate positions, and then input the reconstructed table into the LLM to answer the question.

Meeting adjournment and next meeting

The meeting adjourned at 2:30 pm. The next meeting will be held in November.

Minutes of the 3rd Project Meeting

Date Saturday, November 9, 2024

Time 2:30 pm

Place Room 3557

Present Jing Wang

Prof Fangzhen Lin

Note-taker Jing Wang

Discussion items

I reported on my project progress:

I used CV2's automatic detection of table divider lines and used it to split the text read by Tesseract OCR into cells, then determined the row-column relationships based on the cell coordinates. However, the result was poor.

Therefore, the professor suggested globally reading each character and its coordinates, and gradually merging the text and finding the row-column relationships based on these positional relationships.

Meeting adjournment and next meeting

The meeting adjourned at 2:45 pm. The next meeting will be held in December.

Minutes of the 4th Project Meeting

Date Tuesday, December 10, 2024

Time 11:00 am

Place Room 3557

Present Jing Wang

Prof Fangzhen Lin

Note-taker Jing Wang

Discussion items

I summarized the entire work of this report, including the method of reconstructing tables using OCR, the various complexities of real-world tables, and the limitations of this approach. Together with the professor, I analyzed potential future improvements and other possible methods.

Meeting adjournment and next meeting

The meeting adjourned at 11:30 am.