RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Extracting Financial Data From Unstructured Sources: Leveraging Large Language Models

Huaxia Li[1]

Haoyun (Harry) Gao[1]

Chengzhang Wu[2]

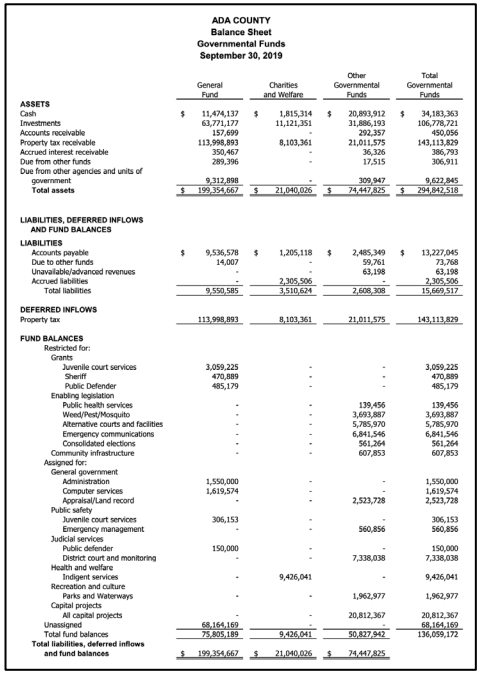
Miklos Vasarhelyi[1]

[1] Rutgers, The State University of New Jersey

[2] Stockton University

Motivation

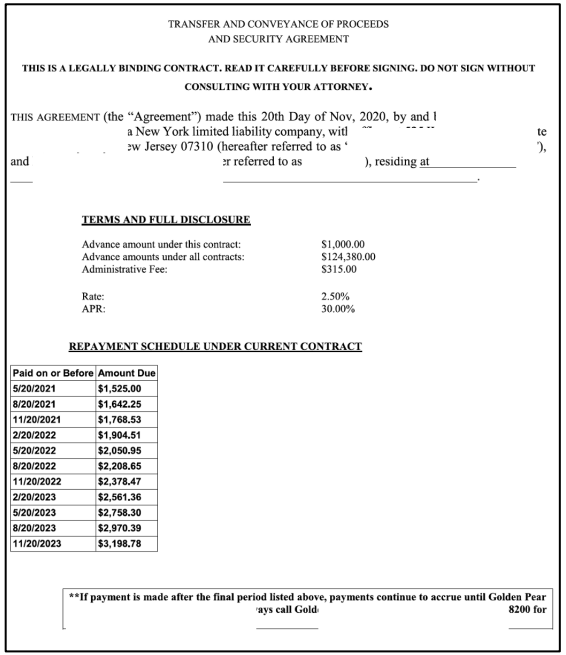
A large amount of accounting data is stored in unstructured formats, such as PDF files.



ADA COUNTY
Balance Sheet
Governmental Funds
September 30, 2019

	General Fund	Charities and Welfare	Other Governmental Funds	Total Governmental Funds
ASSETS				
Cash	$ 11,474,137	$ 1,815,314	$ 20,893,912	$ 34,183,363
Investments	63,771,177	11,121,351	31,886,193	106,778,721
Accounts receivable	157,699	-	292,357	450,056
Property tax receivable	113,998,893	8,103,361	21,011,575	143,113,829
Accrued interest receivable	350,467	-	36,326	386,793
Due from other funds	289,396	-	17,515	306,911
Due from other agencies and units of government	9,312,898	-	309,947	9,622,845
Total assets	$ 199,354,667	$ 21,040,026	$ 74,447,825	$ 294,842,518
LIABILITIES, DEFERRED INFLOWS AND FUND BALANCES				
LIABILITIES				
Accounts payable	$ 9,536,578	$ 1,205,118	$ 2,485,349	$ 13,227,045
Due to other funds	14,007	-	59,761	73,768
Unavailable/advanced revenues	-	-	63,198	63,198
Accrued liabilities	-	2,305,506	-	2,305,506
Total liabilities	9,550,585	3,510,624	2,608,308	15,669,517
DEFERRED INFLOWS				
Property tax	113,998,893	8,103,361	21,011,575	143,113,829
FUND BALANCES				
Restricted for:				
Grants				
Juvenile court services	3,059,225	-	-	3,059,225
Sheriff	470,889	-	-	470,889
Public Defender	485,179	-	-	485,179
Enabling legislation				
Public health services	-	-	139,456	139,456
Weed/Pest/Mosquito	-	-	3,693,887	3,693,887
Alternative courts and facilities	-	-	5,785,970	5,785,970
Emergency communications	-	-	6,841,546	6,841,546
Consolidated elections	-	-	561,264	561,264
Community infrastructure	-	-	607,853	607,853
Assigned for:				
General government				
Administration	1,550,000	-	-	1,550,000
Computer services	1,619,574	-	-	1,619,574
Appraisal/Land record	-	-	2,523,728	2,523,728
Public safety				
Juvenile court services	306,153	-	-	306,153
Emergency management	-	-	560,856	560,856
Judicial services				
Public defender	150,000	-	-	150,000
District court and monitoring	-	-	7,338,038	7,338,038
Health and welfare				
Indigent services	-	9,426,041	-	9,426,041
Recreation and culture				
Parks and Waterways	-	-	1,962,977	1,962,977
Capital projects				
All capital projects	-	-	20,812,367	20,812,367
Unassigned	68,164,169	-	-	68,164,169
Total fund balances	75,805,189	9,426,041	50,827,942	136,059,172
Total liabilities, deferred inflows and fund balances	$ 199,354,667	$ 21,040,026	$ 74,447,825	

Statements



TRANSFER AND CONVEYANCE OF PROCEEDS
AND SECURITY AGREEMENT

THIS IS A LEGALLY BINDING CONTRACT. READ IT CAREFULLY BEFORE SIGNING. DO NOT SIGN WITHOUT CONSULTING WITH YOUR ATTORNEY.

THIS AGREEMENT (the "Agreement") made this 20th Day of Nov, 2020, by and l
a New York limited liability company, witl te
ew Jersey 07310 (hereafter referred to as '),
and r referred to as), residing at

TERMS AND FULL DISCLOSURE

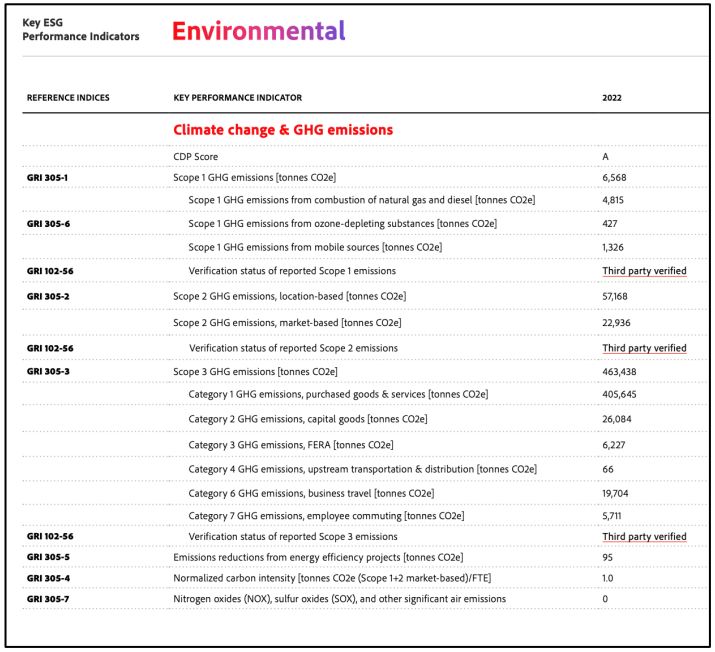
Advance amount under this contract:	$1,000.00
Advance amounts under all contracts:	$124,380.00
Administrative Fee:	$315.00
Rate:	2.50%
APR:	30.00%

REPAYMENT SCHEDULE UNDER CURRENT CONTRACT

Paid on or Before	Amount Due
5/20/2021	$1,525.00
8/20/2021	$1,642.25
11/20/2021	$1,768.53
2/20/2022	$1,904.51
5/20/2022	$2,050.95
8/20/2022	$2,208.65
11/20/2022	$2,378.47
2/20/2023	$2,561.36
5/20/2023	$2,758.30
8/20/2023	$2,970.39
11/20/2023	$3,198.78

**If payment is made after the final period listed above, payments continue to accrue until Golden Pear ays call Gold 8200 for

Contract



Key ESG Performance Indicators **Environmental**

REFERENCE INDICES	KEY PERFORMANCE INDICATOR	2022
	Climate change & GHG emissions	
	CDP Score	A
GRI 305-1	Scope 1 GHG emissions [tonnes CO2e]	6,568
	Scope 1 GHG emissions from combustion of natural gas and diesel [tonnes CO2e]	4,815
GRI 305-6	Scope 1 GHG emissions from ozone-depleting substances [tonnes CO2e]	427
	Scope 1 GHG emissions from mobile sources [tonnes CO2e]	1,326
GRI 102-56	Verification status of reported Scope 1 emissions	Third party verified
GRI 305-2	Scope 2 GHG emissions, location-based [tonnes CO2e]	57,168
	Scope 2 GHG emissions, market-based [tonnes CO2e]	22,936
GRI 102-56	Verification status of reported Scope 2 emissions	Third party verified
GRI 305-3	Scope 3 GHG emissions [tonnes CO2e]	463,438
	Category 1 GHG emissions, purchased goods & services [tonnes CO2e]	405,645
	Category 2 GHG emissions, capital goods [tonnes CO2e]	26,084
	Category 3 GHG emissions, FERA [tonnes CO2e]	6,227
	Category 4 GHG emissions, upstream transportation & distribution [tonnes CO2e]	66
	Category 6 GHG emissions, business travel [tonnes CO2e]	19,704
	Category 7 GHG emissions, employee commuting [tonnes CO2e]	5,711
GRI 102-56	Verification status of reported Scope 3 emissions	Third party verified
GRI 305-5	Emissions reductions from energy efficiency projects [tonnes CO2e]	95
GRI 305-4	Normalized carbon intensity [tonnes CO2e (Scope 1+2 market-based)/FTE]	1.0
GRI 305-7	Nitrogen oxides (NOX), sulfur oxides (SOX), and other significant air emissions	0

KPI in CSR Report

Motivation

- Advancements in **large language models (LLMs)** offer great potential
  - Transform human-generated unformatted information into machine-readable standardized databases (Gu et al., 2023)

- Develop an **LLM-enabled framework** that can extract financial data from unstructured sources
  - Provide valuable insights for market participants, policymakers, and researchers.

Introduction

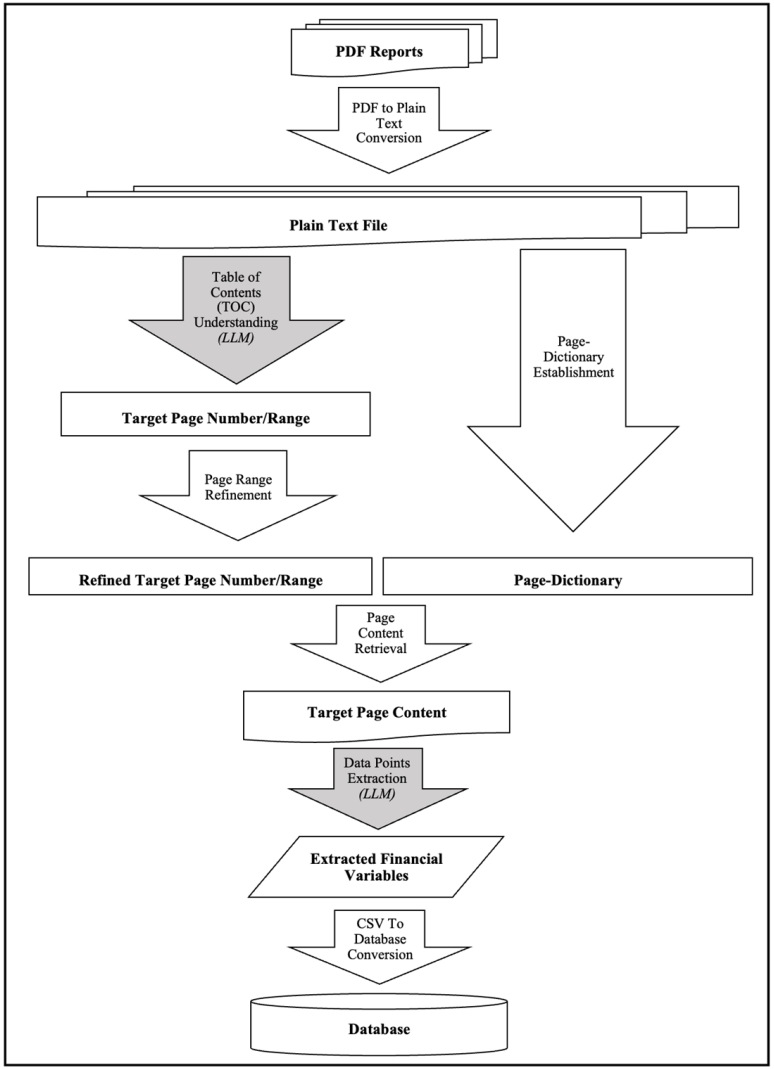
- Develop an LLM-enabled framework that can process PDF-formatted data source and extract predefined financial data from it
- Following the six-step **design science** research methodology
- Incorporate **data preparation**, **prompt engineering**, **batch querying**, and **database construction**
- **Illustration:** Extract financial data from local government's annual financial reports (ACFR)
  - There is no centralized, electronic, and publicly accessible database of governmental financial data (W. J. Kim, Plumlee, and Stubben 2022)
  - ACFR is the primary source of a comprehensive set of financial information for U.S. local governments.

Objectives of the Framework

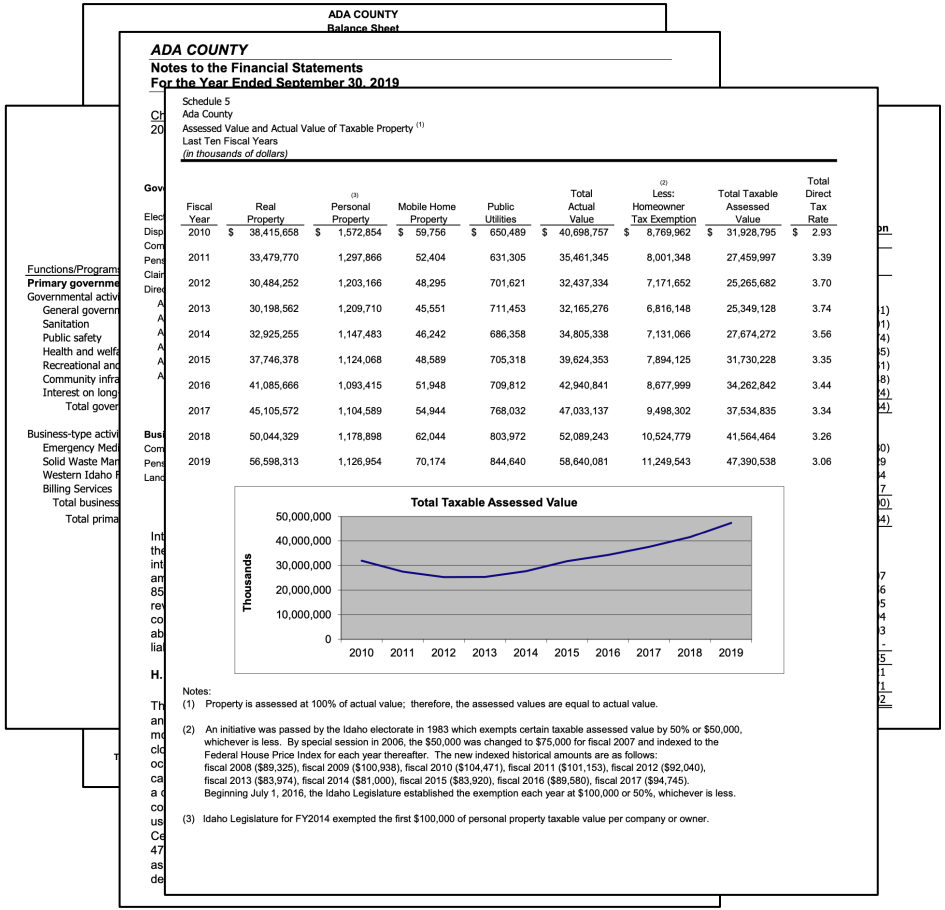
- **Effectiveness**
  - Extracted data can be matched with the manual extracting results by human experts

- **Efficiency**
  - Maximize the efficiency of extraction and make sure the extraction can be automatic in batches

Design the Artifact



Design the Artifact



ADA COUNTY
Balance Sheet

ADA COUNTY
Notes to the Financial Statements
For the Year Ended September 30, 2019

Schedule 5
Ada County
Assessed Value and Actual Value of Taxable Property [1]
Last Ten Fiscal Years
(in thousands of dollars)

Fiscal Year	Real Property	[3] Personal Property	Mobile Home Property	Public Utilities	Total Actual Value	[2] Less: Homeowner Tax Exemption	Total Taxable Assessed Value	Total Direct Tax Rate
2010	$ 38,415,658	$ 1,572,854	$ 59,756	$ 650,489	$ 40,698,757	$ 8,769,962	$ 31,928,795	$ 2.93
2011	33,479,770	1,297,866	52,404	631,305	35,461,345	8,001,348	27,459,997	3.39
2012	30,484,252	1,203,166	48,295	701,621	32,437,334	7,171,652	25,265,682	3.70
2013	30,198,562	1,209,710	45,551	711,453	32,165,276	6,816,148	25,349,128	3.74
2014	32,925,255	1,147,483	46,242	686,358	34,805,338	7,131,066	27,674,272	3.56
2015	37,746,378	1,124,068	48,589	705,318	39,624,353	7,894,125	31,730,228	3.35
2016	41,085,666	1,093,415	51,948	709,812	42,940,841	8,677,999	34,262,842	3.44
2017	45,105,572	1,104,589	54,944	768,032	47,033,137	9,498,302	37,534,835	3.34
2018	50,044,329	1,178,898	62,044	803,972	52,089,243	10,524,779	41,564,464	3.26
2019	56,598,313	1,126,954	70,174	844,640	58,640,081	11,249,543	47,390,538	3.06

Notes:
(1) Property is assessed at 100% of actual value; therefore, the assessed values are equal to actual value.

(2) An initiative was passed by the Idaho electorate in 1983 which exempts certain taxable assessed value by 50% or $50,000, whichever is less. By special session in 2006, the $50,000 was changed to $75,000 for fiscal 2007 and indexed to the Federal House Price Index for each year thereafter. The new indexed historical amounts are as follows: fiscal 2008 ($89,325), fiscal 2009 ($100,938), fiscal 2010 ($104,471), fiscal 2011 ($101,153), fiscal 2012 ($92,040), fiscal 2013 ($83,974), fiscal 2014 ($81,000), fiscal 2015 ($83,920), fiscal 2016 ($89,580), fiscal 2017 ($94,745). Beginning July 1, 2016, the Idaho Legislature established the exemption each year at $100,000 or 50%, whichever is less.

(3) Idaho Legislature for FY2014 exempted the first $100,000 of personal property taxable value per company or owner.

FileName	City	Year	Long-Term Liabilities	Net Pension Liability	OPEB Liabilities	Unrestricted Net Position	SNP_Thousand	SNP_Million
Gwinnett County - Public Financial Report_2021	Gwinnett Co	2021	[111477, 10	[]	[]	[708826]	[1000]	[]
Hamilton County - Public Financial Report (2)_2019	Hamilton Co	2019	[87262, 145	[687808]	[243310]	[-224968]	[1000]	[None]
Hamilton County - Public Financial Report (2)_2020	Hamilton Co	2020	[88749, 741	[741809]	[218552]	[-282257]	[1000]	[None]
Hamilton County - Public Financial Report_2021	Hamilton Co	2021	[92913, 356	[356008]	[59749]	[54526]	[1000]	[None]
Harris County - Public Financial Report_2019	Harris Coun	2019	[30222335	[]	[]	[-72476538	[None]	[None]
Harris County - Public Financial Report_2020	Harris Coun	2020	[31301246	[None]	[None]	[-69067717	[None]	[None]
Harris County - Public Financial Report_2021	Harris Coun	2021	[36348347	[None]	[None]	[-10187858	[None]	[None]
Hennepin County - Public Financial Report_2019	Hennepin C	2019	[19473500,	[76448131	[]	[-81363439	[None]	[None]
Hennepin County - Public Financial Report_2021	Hennepin C	2021	[8109000, 1	[59488658	[15076314	[-67198992	[None]	[None]
Hillsborough County - Public Financial Report_2019	Hillsboroug	2019	[22427, 667	[1031454]	[118871]	[55831]	[1000]	[None]
Hillsborough County - Public Financial Report_2021	Hillsboroug	2021	[15580, 475	[398407]	[140126]	[744443]	[1000]	[None]
Horry County - Public Financial Report_2019	Horry Coun	2019	[41549, 523	[207419]	[41972]	[-23530]	[1000]	[None]
Horry County - Public Financial Report_2020	Horry Coun	2020	[51287, 508	[217400]	[43650]	[-25779]	[1000]	[None]
Horry County - Public Financial Report_2021	Horry Coun	2021	[39667, 435	[253730]	[49316]	[18107]	[1000]	[None]
Howard County - Public Financial Report_2019	Howard Co	2019	[10522793	[]	[]	[-85093012	[None]	[None]
Howard County - Public Financial Report_2020	Howard Co	2020	[10916021	[]	[]	[-67480768	[None]	[None]
Howard County - Public Financial Report_2021	Howard Co	2021	[12187053	[None]	[None]	[-54467420	[None]	[None]
Jefferson County - Public Financial Report_2019	Jefferson C	2019	[17339340,	[]	[]	[30538041]	[None]	[None]
Jefferson County - Public Financial Report_2020	Jefferson C	2020	[18211350,	[None]	[None]	[97889534]	[None]	[None]
Jefferson County - Public Financial Report_2021	Jefferson C	2021	[17982409,	[None]	[None]	[94590037]	[None]	[None]
Jefferson Parish - Public Financial Report_2019	Jefferson Pa	2019	[126291, 41	[167988]	[164213]	[-114882]	[1000]	[None]
Jefferson Parish - Public Financial Report_2020	Jefferson Pa	2020	[116728, 16	[116728]	[167473]	[-79092]	[1000]	[None]
Jefferson Parish - Public Financial Report_2021	Jefferson Pa	2021	[81936, 185	[100906]	[184547]	[-142849]	[1000]	[None]
Johnson County - Public Financial Report_2019	Johnson Co	2019	[57339498,	[]	[]	[18672874	[None]	[None]
Johnson County - Public Financial Report_2020	Johnson Co	2020	[3962583, 7	[None]	[None]	[25243078	[None]	[None]
Johnson County - Public Financial Report_2021	Johnson Co	2021	[82644307,	[]	[]	[31525554	[None]	[None]
Kane County - Public Financial Report_2019	Kane Count	2019	[6200000, 2	[2886546	[9659855]	[10825840	[None]	[None]
Kane County - Public Financial Report_2020	Kane Count	2020	[6555000, 2	[3097725, 1	[10618922]	[12667492	[None]	[None]
Kane County - Public Financial Report_2021	Kane Count	2021	[2690000, 2	[168088, 10	[10645485]	[13582685	[None]	[None]
King County - Public Financial Report_2019	King County	2019	[298072, 58	[None]	[None]	[981633]	[1000]	[None]
King County - Public Financial Report_2020	King County	2020	[97094826,	[97094826]	[9957109]	[-28855319	[None]	[None]
King County - Public Financial Report_2021	King County	2021	[366764, 58	[]	[]	[2420230]	[1000]	[None]
Knox County - Public Financial Report_2019	Knox Count	2019	[97094826,	[97094826]	[9957109]	[-28855319	[None]	[None]
Knox County - Public Financial Report_2020	Knox Count	2020	[12129216	[12129216	[6211127]	[-27922578	[None]	[None]
Knox County - Public Financial Report_2021	Knox Count	2021	[58770238,	[87185359]	[44205205]	[10427297	[None]	[None]
Lake County - Public Financial Report_2019	Lake Count	2019	[10941525	[17761432	[34685096]	[-15263682	[None]	[None]
Lake County - Public Financial Report_2020	Lake Count	2020	[38730967	[23105701	[37048311]	[-17997800	[None]	[None]
Lake County - Public Financial Report_2021	Lake Count	2021	[22780769	[68108707]	[47833626]	[-15658687	[None]	[None]
Larimer County - Public Financial Report_2019	Larimer Co	2019	[13926888	[None]	[None]	[16826455	[None]	[None]
Larimer County - Public Financial Report_2020	Larimer Co	2020	[1086538, 6	[]	[]	[20698016	[None]	[None]
Larimer County - Public Financial Report_2021	Larimer Co	2021	[11501194	[None]	[None]	[21592077	[None]	[None]
Lexington-Fayette Urban County - Public Financial Report_	Lexington-F	2019	[15269978	[47963853	[31845531	[-71116025	[None]	[None]
Lexington-Fayette Urban County - Public Financial Report_	Lexington-F	2021	[18154309	[65541317	[36917724	[-77470918	[None]	[None]

Design the Artifact - Data Preparation

- PDF to Plain text conversion
  - Facilitate machine readability
- Table of Contents (TOC) understanding
  - Extract the targets page number/range
  - Increase accuracy
  - Save computational cost
- Page Range Refinement
  - Shorten page range (e.g., *Notes*)
- Page Dictionary Establishment
  - Transform long document (e.g., annual report)

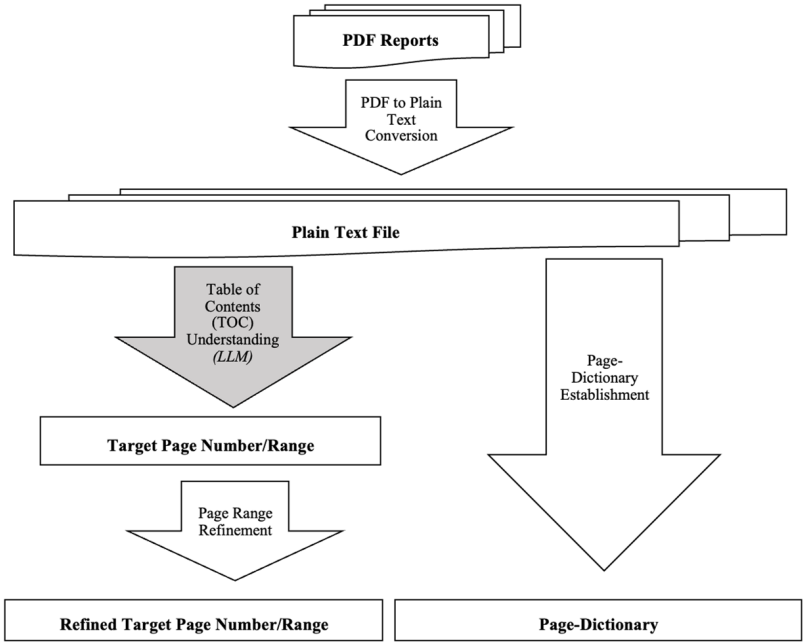
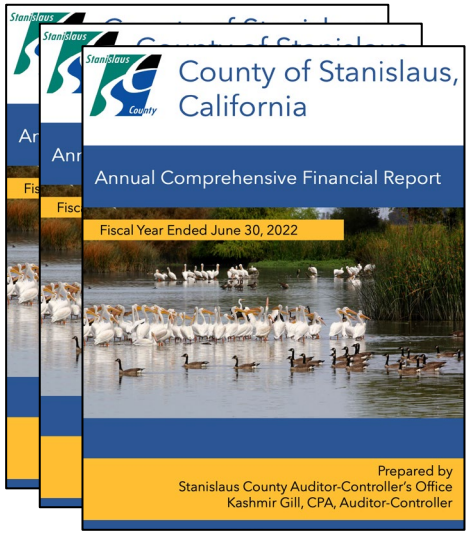


Illustration - Data Preparation - Page Dictionary Establishment



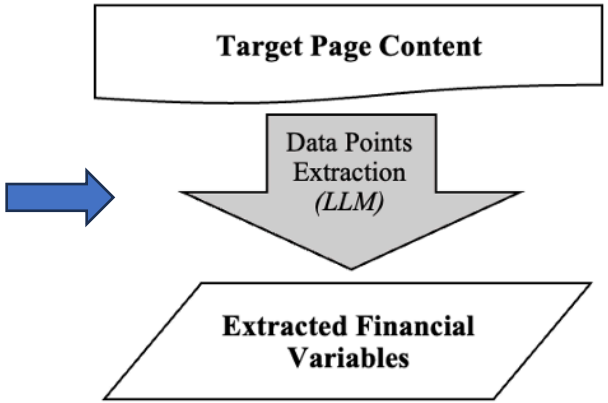
Page Dictionary

page	content
30	County of Stanislaus Notes to the Basic Financial Statements June 30, 2022 NOTE 1 SUMMARY OF SIGNIFICANT ACCOUNTING POLICIE...
31	County of Stanislaus Notes to the Basic Financial Statements June 30, 2022 Supervisors, which is the same governing body as the Coun...
32	County of Stanislaus Notes to the Basic Financial Statements June 30, 2022 In the government-wide statements, eliminations have been...
33	County of Stanislaus Notes to the Basic Financial Statements June 30, 2022 services, outpatient treatment services, and an array of edu...
34	County of Stanislaus Notes to the Basic Financial Statements June 30, 2022 The Custodial Funds account for assets held by the Coun...

Design the Artifact - Prompt Engineering

- Systematic development and optimization of prompts to enhance interactions in alignment with specific objectives or requirements

Prompt Engineering



Design the Artifact - Prompt Engineering

- ***Instruction Learning (Chung et al. 2022; Gu et al. 2023)***
  - Tasks described through explicit instructions

Example

[Role and Context]: "You are an assistant who is good at extracting financial information from unstructured textual data."

[Rule]: "Strictly obey the following rules when extracting:

Rule 1. Find each value by recognizing the relevant row and column names.

Rule 2. Output in the JSON schema: {"Total Asset": [], "Total Expenditure":[]}

[Task]: "The page content is a financial statement. Extract the following values from the statement:

1. Row "Total primary government" for column "Expenses"

Illustration - Prompt Engineering - Zero/Few Shot Learning

- ***Zero/Few-shot Learning (Brown et al. 2020; Kojima et al. 2022; Zhao et al. 2021)***
  - *With or without examples in the prompt*

Example

[Task]: "Row account "Long-Term Liabilities" for column "Total":

a. Some example names for the line items: 'Lease liability', 'Compensated absences payable', 'Post-closure care costs'."

Illustration - Prompt Engineering - CoT Prompting

- ***Chain-of-Thought Prompting (Gao et al. 2023; Gu et al. 2023)***
  - a series of short, interrelated statements or sentences, serving to direct the reasoning process of the LLM in a manner similar to how a human might approach a task

Example

[Task]:

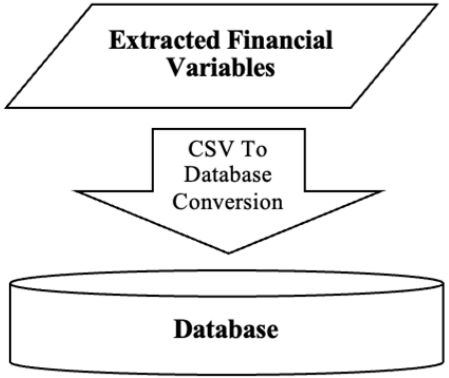
What is the first page containing the Statement of Net Position? Assign it to A.

What is the page number/range of the immediate next statement/item following A? Assign it to B.

Form list_1 with A and B in [A, B] list format.

Design the Artifact - Batch Querying with LLM & Database Construction

- Existing LLM research in accounting
  - Rely on user interface (UI) for interaction
- We formalize prompts as a Python function
  - Easy of deployment, maintenance, adaptation
- Preprocess the data extracted from LLM
  - Unify data unit, format..
- Database Management System (DBMS)
  - PostgreSQL
  - Relational database



PostgreSQL

Evaluation

- Evaluation of effectiveness and efficiency
- Regular meetings with GFOA to obtain expert review

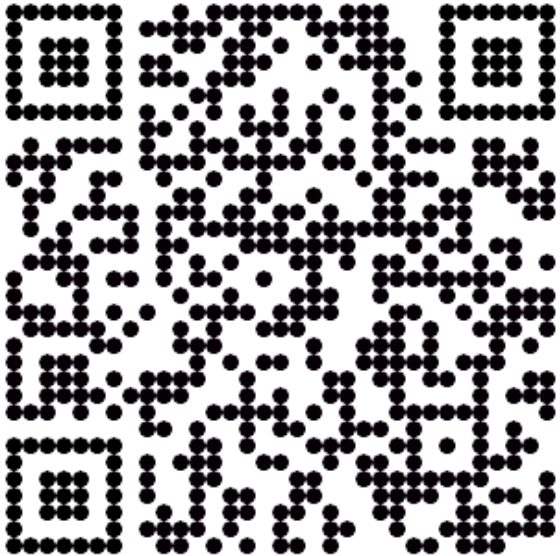
	GPT-4 - initial test	GPT-4 - refined prompts	Experts
Total Count of Data Points	152	152	152
Actual Count of Correct Data Points	146	152	150
% Correct Data Extraction	96.1%	100%	98.7%
% Average Absolute Variance	0.03%	0%	5.2%
Total Time to Extract Data (in minutes)	8	4	200
PDF Conversion Time	4	NA	NA
Code Running Time	4	4	NA

Conclusion

- Introduces a framework for the extraction of financial data from unstructured PDF format, employing state-of-the-art LLM technology.

- Contributions:
  - Devise and validate a framework to extract financial data from unstructured sources
  - Shed light on the potential of LLMs as an alternative approach to traditional costly data standardization methods:
    - Data labeling (XBRL) → Data post-processing (LLM extraction)

Scan to check our paper on SSRN →



Thank You

✉ huaxia.li@rutgers.edu