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

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Real-time government reporting with expanded tags and real-time encoding and posting

Miklos A. Vasarhelyi
Yu Gu
Qingman Wu
Huaxia Li
Continuous Audit and Reporting Lab (CarLab)

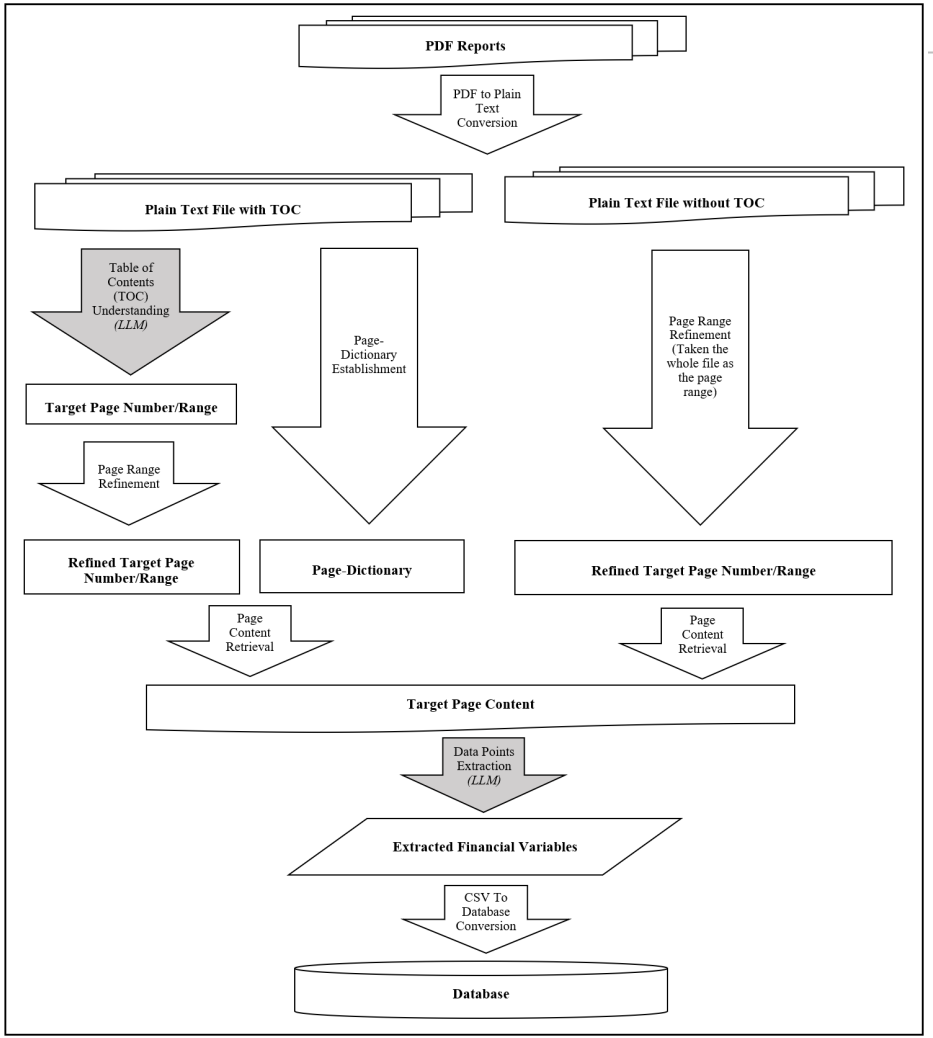
GovFin 2024, July 30, 2024 New York

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.

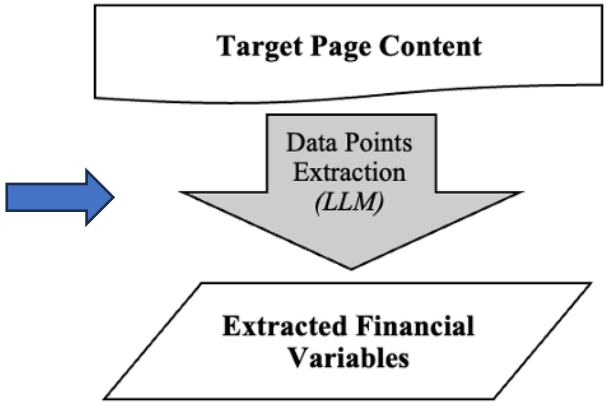
Design the Artifact



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

- ***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 & Evaluation

- ACFRs of US local governments

	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%
Total Time to Extract Data (in minutes)	8	4	200
PDF Conversion Time	4	NA	NA
Code Running Time	4	4	NA

Illustration & Evaluation

- ESG reports of HKEX-listed companies

	GPT-4 - initial test	GPT-4 - refined prompts	Authors
Total Count of Data Points	90	90	90
Actual Count of Correct Data Points	84	89	90
% Correct Data Extraction	93.3%	98.9%	100%
Total Time to Extract Data (in minutes)	5	2.5	45
PDF Conversion Time	2.5	NA	NA
Code Running Time	2.5	2.5	NA

Discussion

- Additional tests on large-scale extraction tasks
  - 4,000 county-year ACFRs resulting in over 80,000 data point
  - Average accuracy of 96%

- Comparison between various LLMs
  - GPT-4: 96.8%
  - Claude 2: 93.7%
  - Bard: exceed input limit

- Comparison with whole PDF Q&A tools
  - ChatGPT-4: inaccurate results
  - Claude 2: exceed input limit

Some Additional Thoughts

XBRL and AI

AI in Enhancing the XBRL Taxonomy

- Identifies and corrects inconsistencies in taxonomy definitions.
- Analyzes the semantics of extension tags across various filings to identify opportunities for integration and to streamline new taxonomy developments.

AI in Optimizing the XBRL Process

- Automates data extraction and validation, reducing manual effort.
- Facilitates real-time preparation, error detection and correction.
- Automates updates to ensure compliance with evolving standards and stakeholders' interest.
- Assists professionals and enhances accuracy and efficiency in financial reporting.

AI in Improving the XBRL File Useability

- Converts complex financial data into more accessible formats.
- Enhances data visualization for better stakeholder understanding.
- Uses natural language processing to simplify narrative sections.
- Incorporates Natural Language Processing (NLP) to extract and process data from XBRL reports, providing advanced analyses and facilitating real-time responses to user' customized queries.

Speculations

- Exogenous variables will be a very important portion of the future reporting schema
  - They can provide futurity to a largely obsolete financial reporting schema
  - E.g. google queries, sentiment, IoT captures, weather, etc
- They are oblique (not orthogonal) to some financial measurements, have different rhythms, measurements, etc.
  - If tagged and integrated will present great value to the business measurement and assurance (audit) ecosystem
- Tailored reporting apps that eventually will replace the one-for-all current reporting schema will be able to integrate these pre-prepared tagged schemata
  - Most of this probably can be done independently with generative technology but their heavy computing content will preclude its widespread utilization
- The inclusion of thousands of categories from ESG reporting and the constant flux of needs, concerns, and standards will make “automatic taxonomy creation,” and “automatic tag embedding” an inevitable need