
# Text Extraction

## User Manual

Version: 1.0

**Prepared by:**

Group 4:

Yufei Chen

Johanna Kienzle

Kacper Mozdzen

Valda Rahima

Lone Tsie

**Date:**

June 27, 2024

## Table of Contents

## 1. Introduction

This user manual provides step-by-step instructions on efficiently a Large Language Model (LLM) to compare two reports. To use this manual, access to both an LLM and a Jupyter Notebook is needed. For detailed information on the development of these prompt recommendations, please refer to the accompanying report. The user manual has been developed to not only produce usable output but also trace the output for accuracy and reliability. LLMs may generate unreliable information, so it is important to handle their outputs cautiously.

The use case example mentioned below has been generated from publicly available data. No copyright infringement is intended.

## 2. Getting Started

Setting up your environment:

### 2.1. Installation

- Create an account of a LLM of your choosing that can handle the upload of PDF documents in your required language. For our example we used ChatGPT4o.
- Download a package manager (we used Conda 22.9.0).
- Set up a working environment by installing Python (Version 3.9.18), Jupyter Notebook, and all required packages (re, PyPDF2, pandas). The official website can be found here.
- Optional: install the list of packages provided, here.

### 2.2. Setup

- Configuration of basic preferences (e.g., language, time zone).
- Connecting external devices or peripherals, if relevant.

### 2.3. Input

- Inside the project folder “PDF_Scanner_Script”,
    - Move the **two** report documents (PDF) that will be compared to the specified folder called “*Reports*”.

- Move the LLM output (csv format) to the specified folder called "*ComparisonTables*".

## 3. Prompt Engineering and Tracing

This section provides a comprehensive guide to generating effective prompts to achieve the desired output.

### 3.1. How to Write a Good Prompt

To achieve a desired output, a good prompt is essential. While ChatGPT can be useful for generating creative output, it can sometimes produce hallucinations, especially when seeking for factual consistent information. Hallucinations can occur due to:

1. Lack of context: Insufficient details of the prompt.
2. Unclear prompt: Ambiguous or unclear specification of format.
3. Training data: The LLM may lack specific knowledge.

To mitigate these issues, use the Chain-of-Thought model, which encourages the model to break down its reasoning process step by step in a more detailed and structured manner. Additionally, uploading internal documents as input helps restrict ChatGPT to domain-specific data.

These are the key points of the prompt-generating methods:

- **Be clear on the objective**: Clearly state what you want the model to do. For example, "compare A and B".
- **Specify the desired format**: Indicate how the results should be presented (e.g., tables).
- **Ensure accuracy**: Direct quotes should be accurately extracted, and page numbers should be correctly referenced.
- **Avoid unnecessary elements**: Exclude hyperlinks, citations, or references in the table.
- **Include handling for missing information**: Instruct the model to leave a column blank if it does not find relevant context.
- **Encourage detailed reasoning**: Mention "*Think Step by Step*."

### 3.2. Tracing Prompt Output for Accuracy

This section outlines the use of the provided Python script to determine whether the LLM-generated quotes are correct.

Before running the script, please ensure the following:

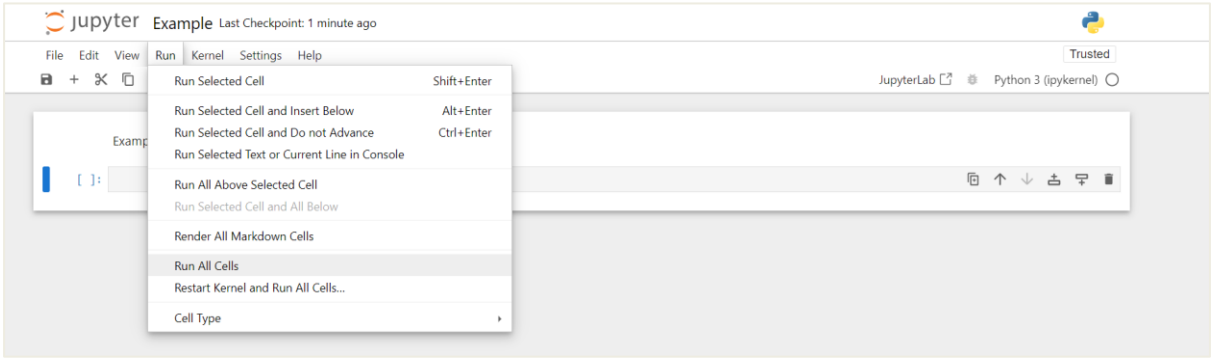
- The two documents needed for comparison are in the designated folder.
- The LLM output is in the designated folder.
- The necessary packages are installed in the Python environment.

Note

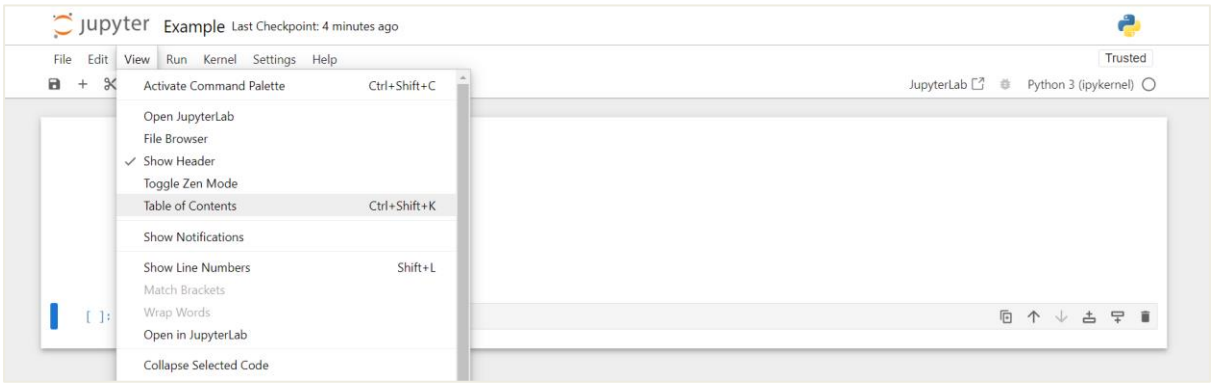
The Python script will read the two documents in the designated folder in alphabetical order. Therefore, the first document will be called ***Report#1,*** and the second, ***Report#2.***

Steps:

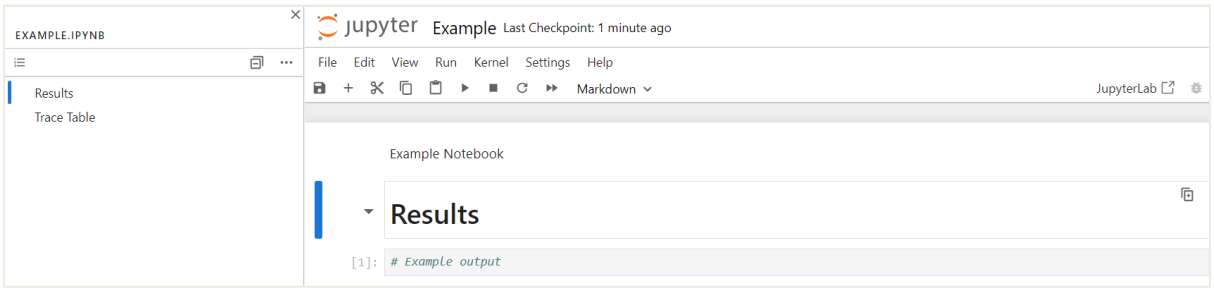
1. Open the script and select the **Run** tab followed by "*Run All Cells*" – depending on the hardware, this may take some time.




2. Ensure the Table of Contents is viewable by selecting the **View** tab followed by "*Table of Contents"*.



3. Click "*Results*" from the list on the left-hand side and scroll past the code to view the results of the script.




Understanding Results

With regard to the results, there are four sets of values that describe the validation. The value counts are also grouped into three categories, Same Page, Different Page and Not Found.

Categories

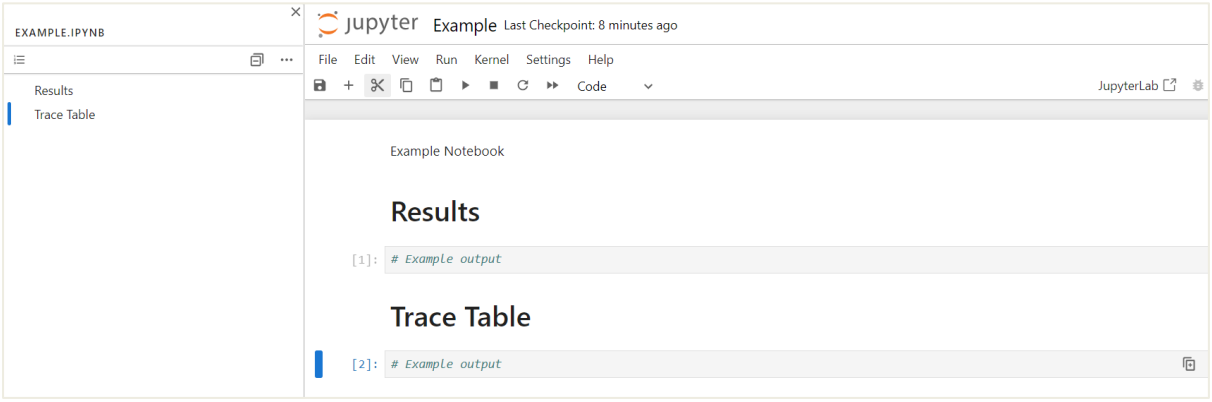
- **Same Page**: This means that the output from the LLM produced supporting quotes that are accurately identified to which page they come from.
- **Different Page**: This means that although the supporting quote generated from the LLM exists in the specified document, it gave the wrong page number.
- **Not Found**: This means that the supporting quote generated from the LLM does not exist in the specified document. This is a hallucination and the purpose of this script - to identifying false information from the output.

Values

1. The first set of values shows the result of the search of quotes that are said to be found in Report#1. Ideally, quotes are expected to be identified as *Same Page*.
2. The second set of values shows the result of the search of quotes that are said to be found in Report#1 but rather were extracted from Report#2. Ideally, all quotes will be identified as *Not Found* to signify that the LLM produced a supporting quote from the specified PDF document.
3. The third set of values shows the result of the search of quotes that are said to be found in Report#2 but rather were extracted from Report#1. Ideally, all quotes will be identified as Not Found to signify that the LLM produced a supporting quote from the specified PDF document.

4. The last set of values shows the result of the search of quotes that are said to be found in Report#2. Ideally, quotes are expected to be identified as *Same Page*.

To view the output of the validation along with the specific insight and accompanying quote, select "*Trace Table*" from the table of contents. Here you will see a table with various columns for each report that was in the designated folder.




### 3.3. Prompt Output

It is important to note that it is possible that even though you are satisfied with the prompt you have written, the output produced can vary every time it is run. It is recommended to produce at least 3 sessions of output with the LLM of your choosing. The Python script is able to handle multiple output of csv tables and will run them all together.

## 4. Use Case Example

For this example, FrieslandCampina would like to do a content comparison for their annual report with a policy recommendation report from FAO (The Food and Agriculture Organization of the United Nations). The documents used for this particular example are:

1. *150 Years, A Nourishing Company (FC - 2021).*
2. *The State of Food Security & Nutrition In The World (FAO -2021).*

**Goal**

The company wants to compare the key themes of Nutritional Health and Well-being of FrieslandCampina's 2021 annual report and the 2021 FAO report.

**Good Prompt**

***Compare the key themes*** of Nutritional Health and Well-being between the two provided annual reports: "FAO 2021 Report" and "FrieslandCampina 2021 Report." ***Create a detailed table*** with the following columns:

- FAO 2021 Report: Description of all instances of Nutritional Health and Well-being as presented in the FAO 2021 Report. Create a new row for each instance.
- Supporting Quote from FAO Report: ***Direct quotes*** from the FAO report to show evidence for each description. ***Include a page number***.
- FrieslandCampina 2021 Report: Description of all instances of Nutritional Health and Well-being as presented in the FrieslandCampina 2021 Report. Create a new row for each instance.
- Supporting Quote from FrieslandCampina Report***:*** ***Direct quotes*** from the FrieslandCampina report to show evidence of the description. ***Include a page number***.

Use the quotes and information ***directly from the provided PDF documents*** to complete the table accurately. Provide the information ***without including any hyperlinks, citations, or references***. ***Let us think step by step***.

**Output**

## Comparison of Nutritional Health and Well-being Themes in FAO 2021 Report and FrieslandCampina 2021 Report

| FAO 2021 Report: Description | Supporting Quote from FAO Report | FrieslandCampina 2021 Report: Description | Supporting Quote from FrieslandCampina Report |
|---|---|---|---|
| Impact of Food Systems on Health and Well-being | "Food systems may exert both positive and negative impacts on human health through multiple interrelated pathways, which are influenced by factors arising from within and outside food systems, including social, economic and environmental determinants of health." (p. 112) | Focus on Nutritional Value of Products | "We aim to produce nutritious and high-quality dairy products that contribute to health and well-being." (p. 24) |
| Unhealthy Diets as a Key Risk Factor for Global Disease | "Unhealthy diets are among the key risk factors driving the global burden of disease; moreover, the negative health impacts associated with poor quality diets are significant." (p. 112) | Addressing Lifestyle Diseases | "We are actively engaged in addressing lifestyle diseases by promoting the consumption of dairy as part of a balanced diet." (p. 26) |
| Importance of Nutritional Quality in Food Production and Supply | "The nutritional quality of foods produced and supplied affects dietary goals and diet-related health goals." (p. 113) | Product Reformulation for Nutritional Improvement | "We have reformulated several of our products to reduce sugar, salt, and fat content to improve their nutritional profile." (p. 29) |

**Future Continuation with the Output**

There are many things we can do with the comparison output. For example, from the comparison of 2021 FrieslandCampina's Annual Report and FAO 2021, we can analyze the points of alignment, points not fully aligned, and suggestions for FrieslandCampina for better adapting to the current nutritional health and well-being trends of the world.

**Another Good Prompt**

Provide me a critical analysis of the comparison output, which points are aligned, and which points are not yet aligned, and the suggestion for implementation for FrieslandCampina with bullet points.

| | |
|---|---|
| **Points of Alignment** | **Impact of Food Systems on Health and Well-being:**<br>• Both reports emphasize the significance of food systems on health.<br>• FrieslandCampina focuses on producing high-quality, nutritious dairy products.<br><br>**Unhealthy Diets as a Key Risk Factor for Global Disease:**<br>• Addressing lifestyle diseases aligns with FAO's focus on unhealthy diets.<br>• FrieslandCampina promotes dairy as part of a balanced diet to mitigate lifestyle diseases.<br><br>**Importance of Nutritional Quality in Food Production and Supply:**<br>• Emphasis on improving nutritional quality matches FAO's guidelines.<br><br>FrieslandCampina reformulates products to reduce sugar, salt, and fat content. |
| **Points Not Fully Aligned** | **Broader Systemic Factors:**<br>• FAO considers social, economic, and environmental determinants of health.<br>• FrieslandCampina's report focuses primarily on product-specific improvements without addressing broader systemic factors.<br><br>**Comprehensive Approach to Food Systems:**<br>• FAO calls for a holistic understanding of food systems.<br><br>FrieslandCampina's report could better reflect the interconnected nature of food systems beyond product reformulation. |

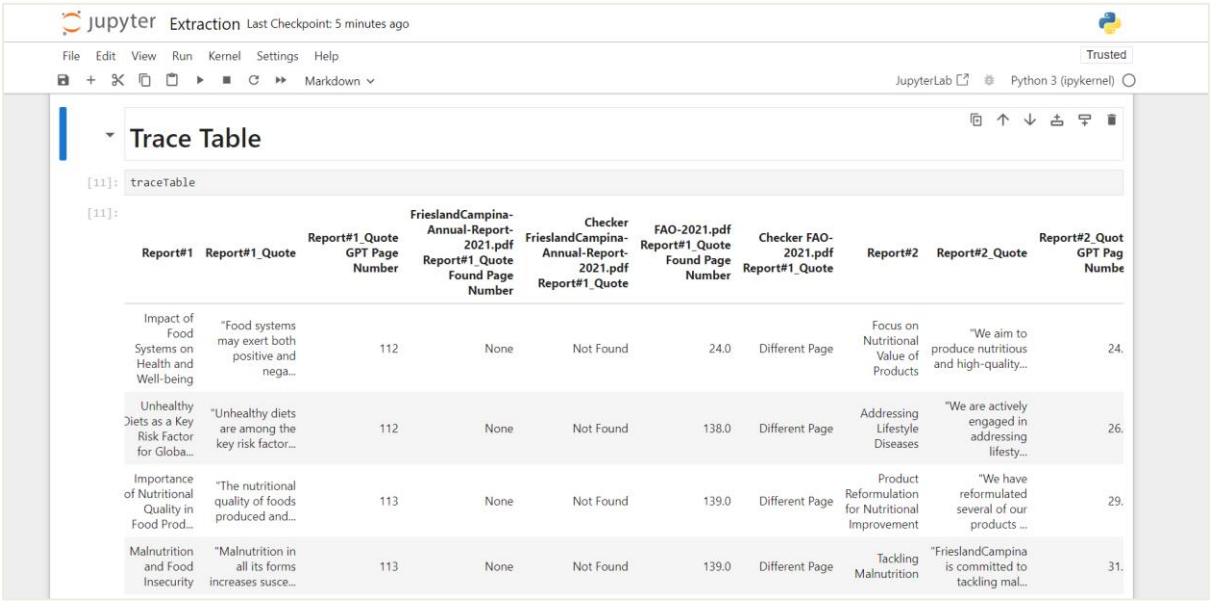
| | |
|---|---|
| **Suggestion for FrieslandCampina** | **Integrate Broader Systemic Factors:**<br>• Address social, economic, and environmental determinants of health.<br>• Develop initiatives that support sustainable farming, equitable food access, and consider socio-economic impacts.<br><br>**Adopt a Holistic Approach to Food Systems:**<br>• Reflect the comprehensive and interconnected nature of food systems.<br>• Launch programs that improve health at multiple levels (e.g., community health initiatives, education on balanced diets) and report on these efforts.<br><br>**Expand Reporting on Health Impacts:**<br>• Provide detailed reporting on the health impacts of products and initiatives.<br>• Conduct and publish long-term health impact studies of FrieslandCampina's products, guiding further product development and health strategies.<br><br>**Engage in Collaborative Efforts:**<br>• Collaborate with public health organizations, governments, and industry peers.<br><br>Participate in multi-stakeholder initiatives aimed at improving global health through better food systems and share insights with the industry. |

Tracing Output

For this use case, the tracing results were as follows:


As we can see, Report#1 – FAO-2021, had 7 supporting quotes for the produced key insights that were not found and thus hallucinations. Both checks for the quotes appearing in the wrong document came out accurately as all were identified as not found. This means the LLM did not assign a quote from the wrong document. For Report#2 – FrieslandCampina-Annual-Report-2021, the LLM performed not as well as the first report analysis. Here 14 quotes were not found and there are only 4 that exist and were on the page that was mentioned.

For this use case, the trace table is as follows:




Here we can see that from Report#1 – FAO-2021, the key insight generated by ChatGPT as the *Impact of food systems on health and wellbeing*, was given a supporting quote that has been marked as None for the page number and Different Page number for its existence. This means that the quote generated does exist within the specified document on page 24, however, the number given was incorrectly stated as page 112. For instances like this, it is

therefore advised to cautiously use this information for any possible report writing. Additional confirmation can be made by finding the quote within the document at the specified correct page and reading it in context to ensure its usability.

## 5. Troubleshooting

This section should assist users in identifying and resolving common issues.

### 5.1. Common Issues

- *Error identifying column name from LLM output*

  The current code renames columns so this is unlikely to occur.
- *The phrasing of the LLM output is not what is expected*

  Rerun prompt till the LLM produces an output that follows the requirements or adjusts the prompt to follow the specified expectations
- *Python code not running*

  Use the provided packages script to run the required packages. If the issue persists, use ***pip install*** instead

### 5.2. Frequently Asked Questions

- *What type of report document can I use?*

  This user manual is designed for reports in PDF format that are majority textual in form.
- *How many documents can I compare at a time?*

  This will depend upon the LLM that you have chosen. For example, ChatGPT4o can handle up to 10 documents. If a comparison of more than two documents is needed, technical skills in Python are required to customize the code to the new output for a 2+ comparison. Otherwise, only two documents can be compared.
- *What language can the documents be? Is English required?*

  As long as both documents are in the same language and the LLM used produces output in that language, the manual is still usable. The Python code search function can work on Latin alphabet-based languages. Special care and customization may be needed for character-based languages such as Mandarin.

- *What additional processing can be done?*

  For those more familiar with Python and Jupyter notebook, customizable processing can be done to the code and LLM output. For example, if the user wants to only filter the key insights whereby the supporting quote exists and whose page was correctly identified, a filter can be added to the corresponding columns.

## 6. Glossary of Terms

| | |
|---|---|
| LLM | Large language model |
| Prompt | The input request for the LLM |
| Prompt engineering | The process of designing a prompt to elicit desired output |
| CoT | Chain of Thought |