



WAGENINGEN
UNIVERSITY & RESEARCH

Text Extraction

User Manual

Version: 1.0

Prepared by:

Group 4:

Yufei Chen
Johanna Kienzle
Kacper Mozdzen
Valda Rahima
Lone Tsie

Date:

July 1, 2024

Table of Contents

1. Introduction (almost done)	3
2. Getting Started (to complete)	3
2.1. Installation	3
2.2. Setup	3
2.3. Input	4
3. Prompt Engineering and Tracing	4
3.1. How to Write an Effective Prompt (done)	4
3.2. Tracing Prompt Output for Accuracy (not yet)	5
4. Troubleshooting	10
4.1. Common Issues	10
4.2. Frequently Asked Questions	11
5. Glossary of Terms (to do)	11

1. Introduction

This user manual provides step-by-step instructions on efficiently prompting 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 (done: just need hyperlinks)

Setting up your environment:

2.1. Installation

- Create an account of a Large Language Model (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 working environment by installing Python (Version 3.9.18), Jupyter notebook and all required packages (re, PyPDF2, pandas)
- Optional: install list of packages provided for, [here](#)

2.2. Setup

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

2.3. Input

- Inside of 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 (to complete)

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

3.1. How to Write a Good Prompt

ChatGPT might be useful for generating creative input, but sometimes it generates hallucinations when looking 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 this, use the Chain-of-Thought model, encouraging 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:** 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 ensure 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

Steps:

1. Open script and select "Run All" – depending on the hardware, this may take some time
2. Ensure Table of Contents is open and select "Results" to view the results of the script

Understanding Results

With regards 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.

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 however 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 output. To identifying false information from the output.

3.3. Prompt Output

It is important to note that it is quite likely that once you are satisfied with the prompt you've written, the output produced can vary every time it is run.

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 use 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 related to Nutritional Health and Well-being of the 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 for 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 for 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's 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 suggestion for FrieslandCampina for better adapting with 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 aren't yet aligned, and the suggestion for implementation for FrieslandCampina with a bullet points

<p>Points of Alignment</p>	<p>Impact of Food Systems on Health and Well-being:</p> <ul style="list-style-type: none"> • Both reports emphasize the significance of food systems on health. • FrieslandCampina focuses on producing high-quality, nutritious dairy products. <p>Unhealthy Diets as a Key Risk Factor for Global Disease:</p> <ul style="list-style-type: none"> • Addressing lifestyle diseases aligns with FAO's focus on unhealthy diets. • FrieslandCampina promotes dairy as part of a balanced diet to mitigate lifestyle diseases. <p>Importance of Nutritional Quality in Food Production and Supply:</p> <ul style="list-style-type: none"> • Emphasis on improving nutritional quality matches FAO's guidelines. • FrieslandCampina reformulates products to reduce sugar, salt, and fat content.
<p>Points Not Fully Aligned</p>	<p>Broader Systemic Factors:</p> <ul style="list-style-type: none"> • FAO considers social, economic, and environmental determinants of health. • FrieslandCampina's report focuses primarily on product-specific improvements without addressing broader systemic factors. <p>Comprehensive Approach to Food Systems:</p> <ul style="list-style-type: none"> • FAO calls for a holistic understanding of food systems. • FrieslandCampina's report could better reflect the interconnected nature of food systems beyond product reformulation.

<p>Suggestion for FrieslandCampina</p>	<p>Integrate Broader Systemic Factors:</p> <ul style="list-style-type: none"> • Address social, economic, and environmental determinants of health. • Develop initiatives that support sustainable farming, equitable food access, and consider socio-economic impacts. <p>Adopt a Holistic Approach to Food Systems:</p> <ul style="list-style-type: none"> • Reflect the comprehensive and interconnected nature of food systems. • Launch programs that improve health at multiple levels (e.g., community health initiatives, education on balanced diets) and report on these efforts. <p>Expand Reporting on Health Impacts:</p> <ul style="list-style-type: none"> • Provide detailed reporting on the health impacts of products and initiatives. • Conduct and publish long-term health impact studies of FrieslandCampina’s products, guiding further product development and health strategies. <p>Engage in Collaborative Efforts:</p> <ul style="list-style-type: none"> • Collaborate with public health organizations, governments, and industry peers. • Participate in multi-stakeholder initiatives aimed at improving global health through better food systems and share insights with the industry.
---	--

5. Troubleshooting

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

5.1. Common Issues

- Error Identifying column name from LLM output
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 adjust the prompt to follow the specified expectations

- Python code not running

Use 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've 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.

6. Glossary of Terms (to do)

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