GDM 5 Final Project

Exploring the innovative use of generative AI to extract crucial information from geological documents, enhancing research and data accessibility.

Aicha EL BOU
Juan SIVORI
Samuel NAUTIN
Stephania ZAMBRANO

Tutors:
Henri BLONDELLE
David RHENALS

Evaluating LLM Accuracy in Document Processing

THIS PRESENTATION FOCUSES ON ASSESSING LLM RESPONSES IN EXTRACTING INFORMATION FROM GEOLOGICAL DOCUMENTS, EMPHASIZING ACCURACY AND MODEL COMPARISON.

AgileDD

COMPREHENSIVE PROJECT OVERVIEW

Exploring the Key Aspects of the Project

Our Timeline

Phase 1

GETTING TO KNOW AGILE DD

- Project socialization
- Communicate and gather Geologist insight
- Platform recognition (Agile DD)
- Creation of initial prompt bank
- Benchmark among different LLMs

Phase 2

DATA PREPROCESSING AND ENRICHMENT FOR LLMS USING RAG

- Definition of action plan for second phase, standardization of questions
- Definition of the documents manipulation and final configuration
- Evaluation of 5 documents

Phase 3

PROJECT DEVELOPING

- Data sets standardization
- Evaluation of 15 more documents
- Statistics
- Results and conclusions

Defense

DEFENSE

- PROJECT SUSTENTATION

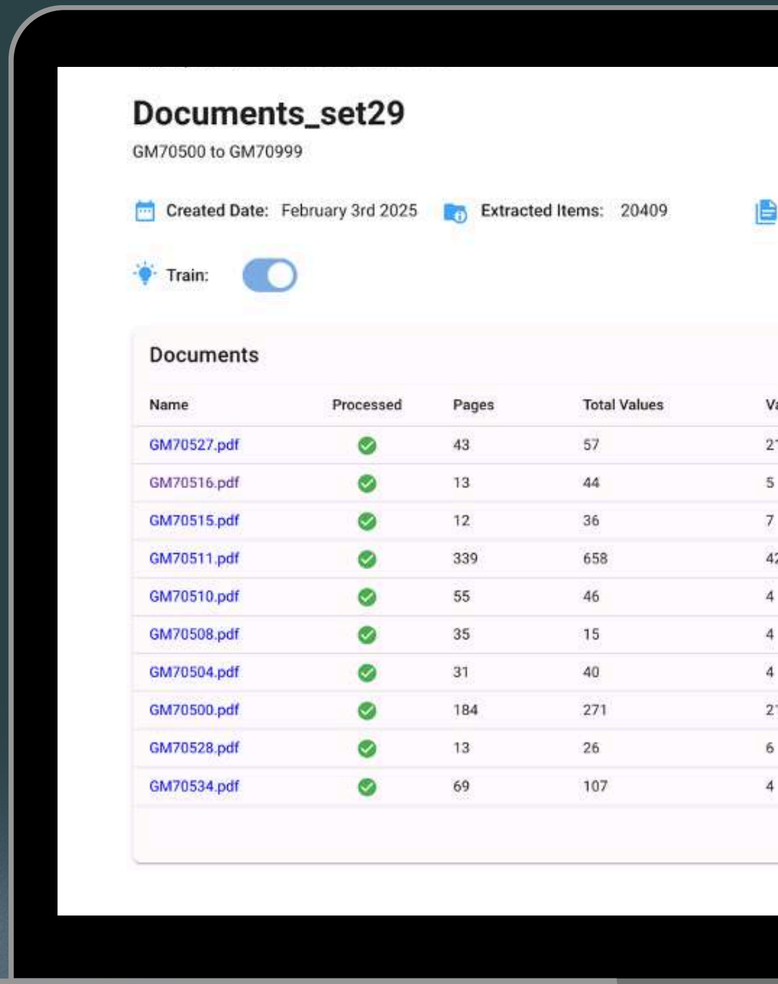
PHASE 1: GETTING TO KNOW AGILE DD

Initial Objective
Generate a benchmark to assess the accuracy of LLM responses in extracting information from technical documents.

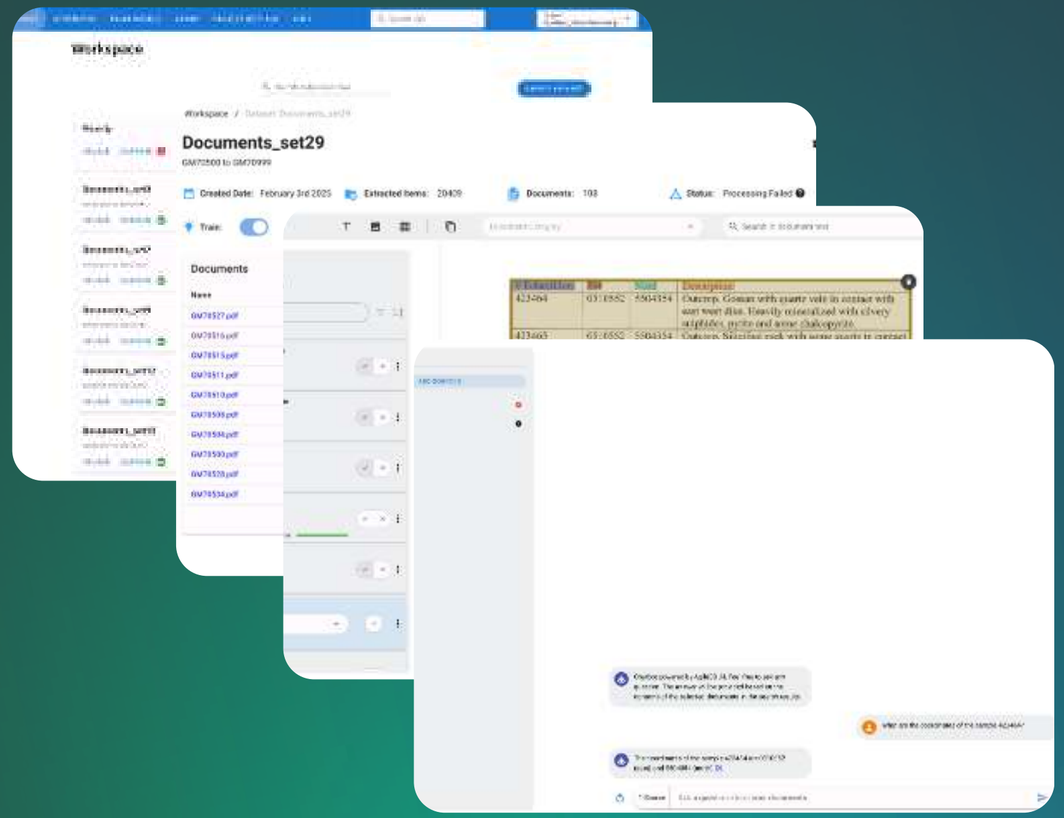
Expanded Scope
Evaluate the accuracy of Agile DD responses and compare them with other LLMs for performance insights.

Importance of Evaluation
Identifying strengths and weaknesses in data comprehension aids in improving model selection for specific use cases.

Relative Performance
Evaluating Agile DD against other LLMs showcases its accuracy and limitations in real-world applications.



Agile DD



Services

- AgileDD is an AI-powered platform that extracts and structures data from complex technical documents. It combines automation with human expertise to enhance data processing for industries like energy, mining, and defense.

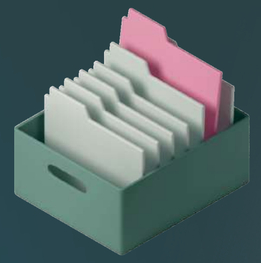
Documents

- Geological Reports
- Assay Certificates (Geochemical Data)
- Sample tables (Mineral Composition)

AgileDD

IFP SCHOOL

Prompt Bank



43
Initial Prompts

36
Refined Prompts

6
Categories

4
Final Prompts

- What are the coordinates of the sample #X
- Focusing on the Json/Xml part, Which sample has the highest concentration of Zinc (Zn) ? excluding #X
- What are the geochemical anomalies, and do they indicate potential mineralization?
- What exploration recommendations were made based on the findings?
- Which sample has the highest concentration of Element (El) ? excluding #X
- Are there historical work results, and what were the assay values?

PHASE 2 : DATA PREPROCESSING AND ENRICHMENT FOR LLMS USING RAG

- **Final Objective**
  Extract text from geological documents using generative AI with a focus on tables.
- **Expanded Scope**
  Incorporate table recognition and structured data extraction to enhance AI responses.
- **Importance of Table Data**
  Tables often contain critical information that AI models find challenging to interpret.
- **Benefits of Structured Formats**
  Utilizing structured formats like JSON/XML can significantly improve accuracy in processing technical documents.

Data extraction and processing workflow

OCR Text Extraction

Utilizing OCR technology to extract text from PDF documents efficiently.

Table Identification

Recognizing and isolating tables from the extracted text for further processing.

Data Conversion

Transforming extracted tables into structured formats like JSON or XML.

Data Combination

Integrating OCR text with structured data formats in various configurations.

DOCUMENT CONFIGURATION

A - original pdf

B - OCR

C - OCR + tables in JSON at the end

D - OCR + tables in XML at the end

E - OCR + tables in JSON at the beggining

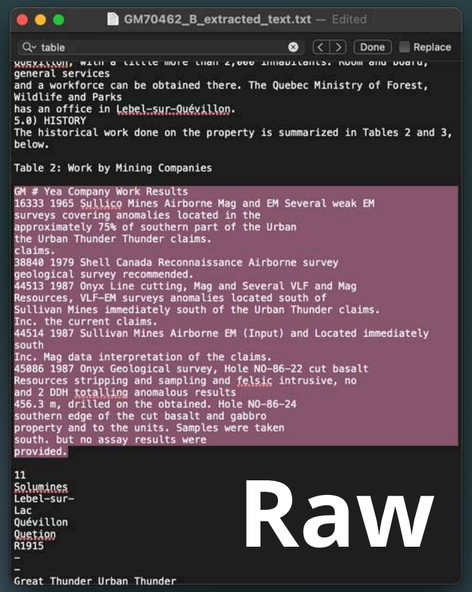
F - OCR + tables in JSON at the end + tables removed

G - OCR + tables in XML at the end + tables removed

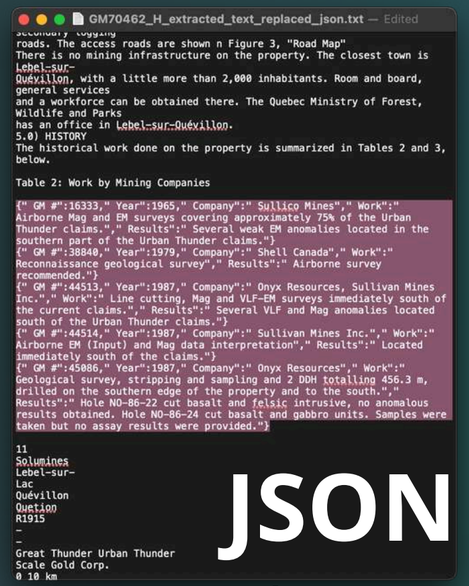
H - OCR + tables in JSON in place of the tables removed

I - OCR + tables in XML in place of the tables removed

Data extraction and processing workflow

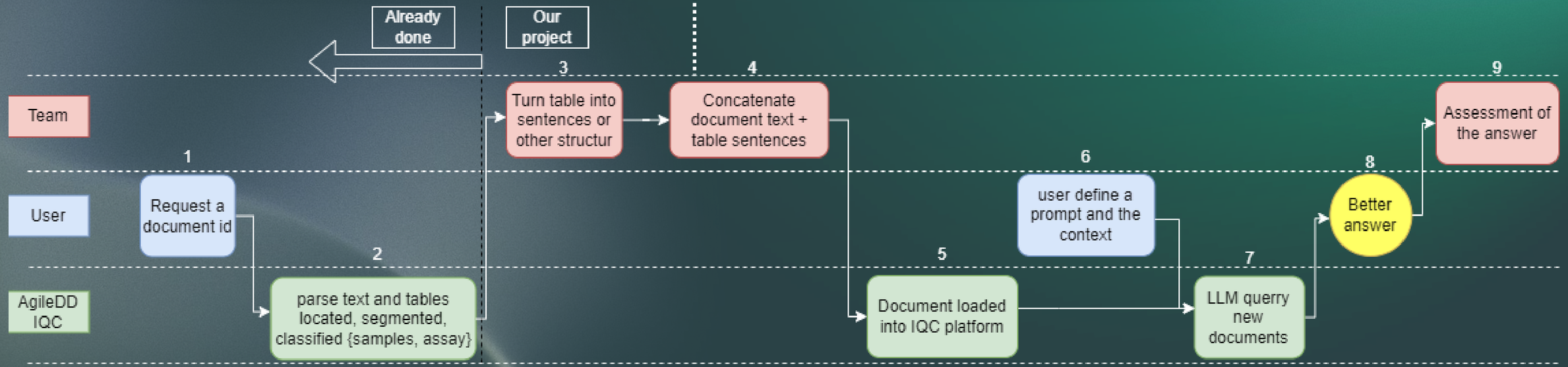


Raw



JSON

XML

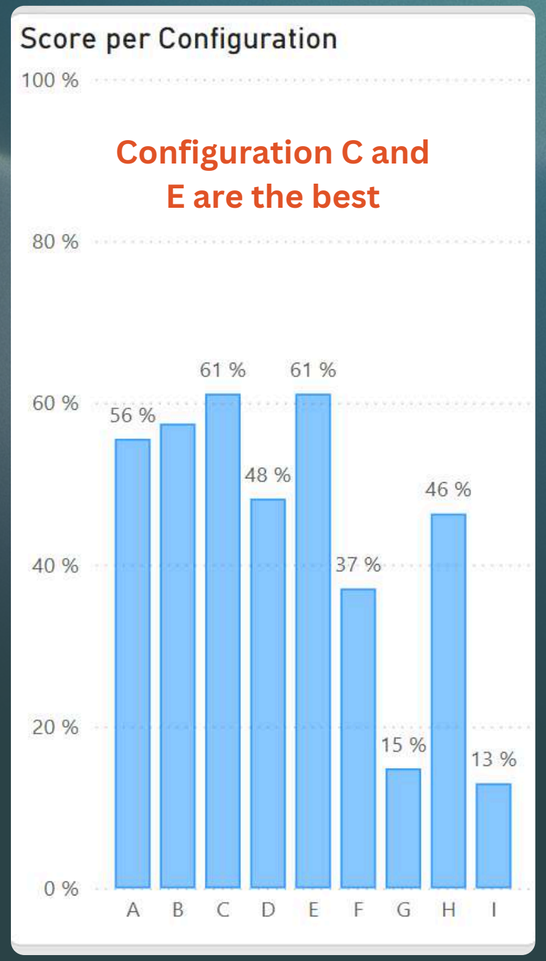


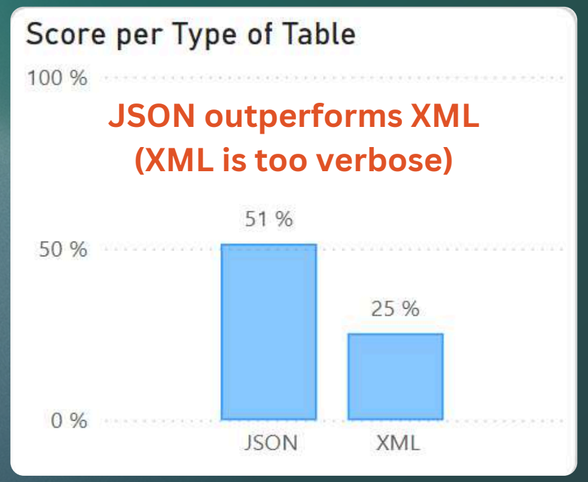
QUERYING & EVALUATION OF AI RESPONSES

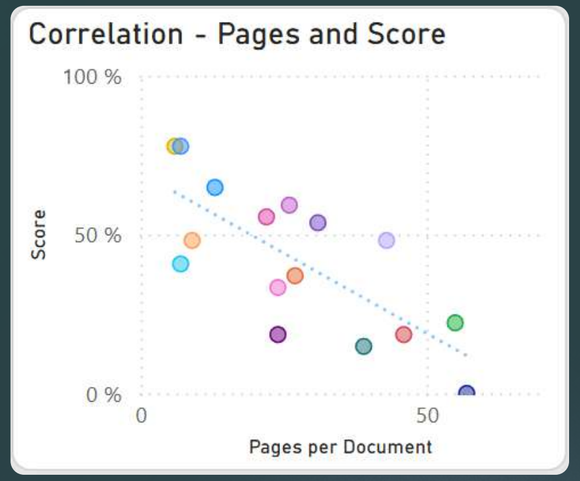
Testing prompt strategies for AI response accuracy

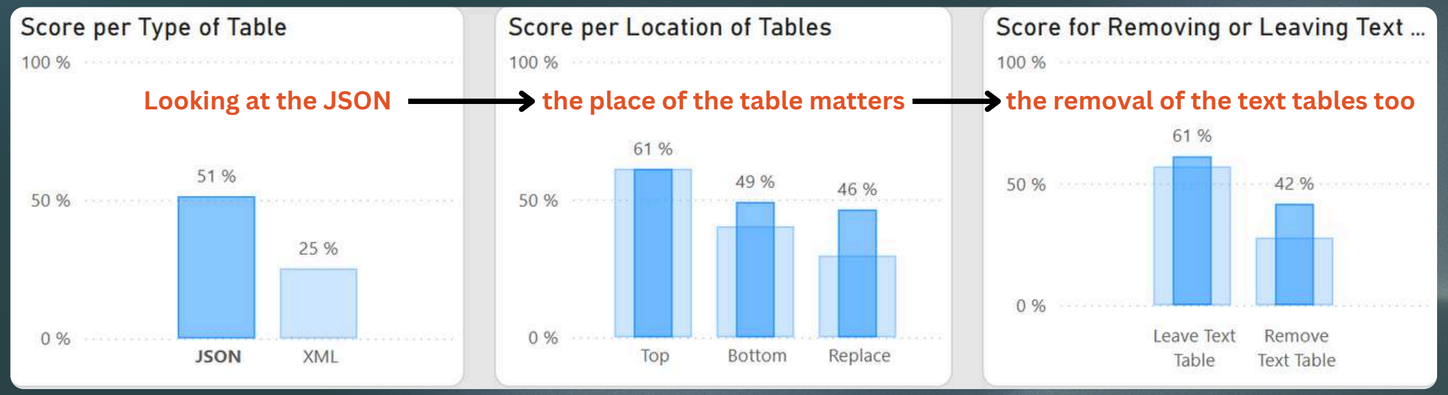
Item	prompt standard	prompt customed
1	What are the **coordinates** of the sample #X	What are the coordinates of the sample CV17-005
2	Which sample has the **highest concentration** of Element (El) ? excluding #X	Which sample has the highest concentration of Zinc (Zn) ? excluding CV17-005
3	Which sample has the **highest concentration** of Element (El), and what is its **description.**	Which sample has the highest concentration of Zinc (Zn), and what is its description.
4	Focusing on the **JSON/XML** part. Which sample has the **highest concentration** of Element (El) ? excluding #X	Focusing on the JSON/XML part. Which sample has the highest concentration of Zinc (Zn) ? excluding CV17-005

RESULTS AND CONCLUSIONS









INSIGHTS

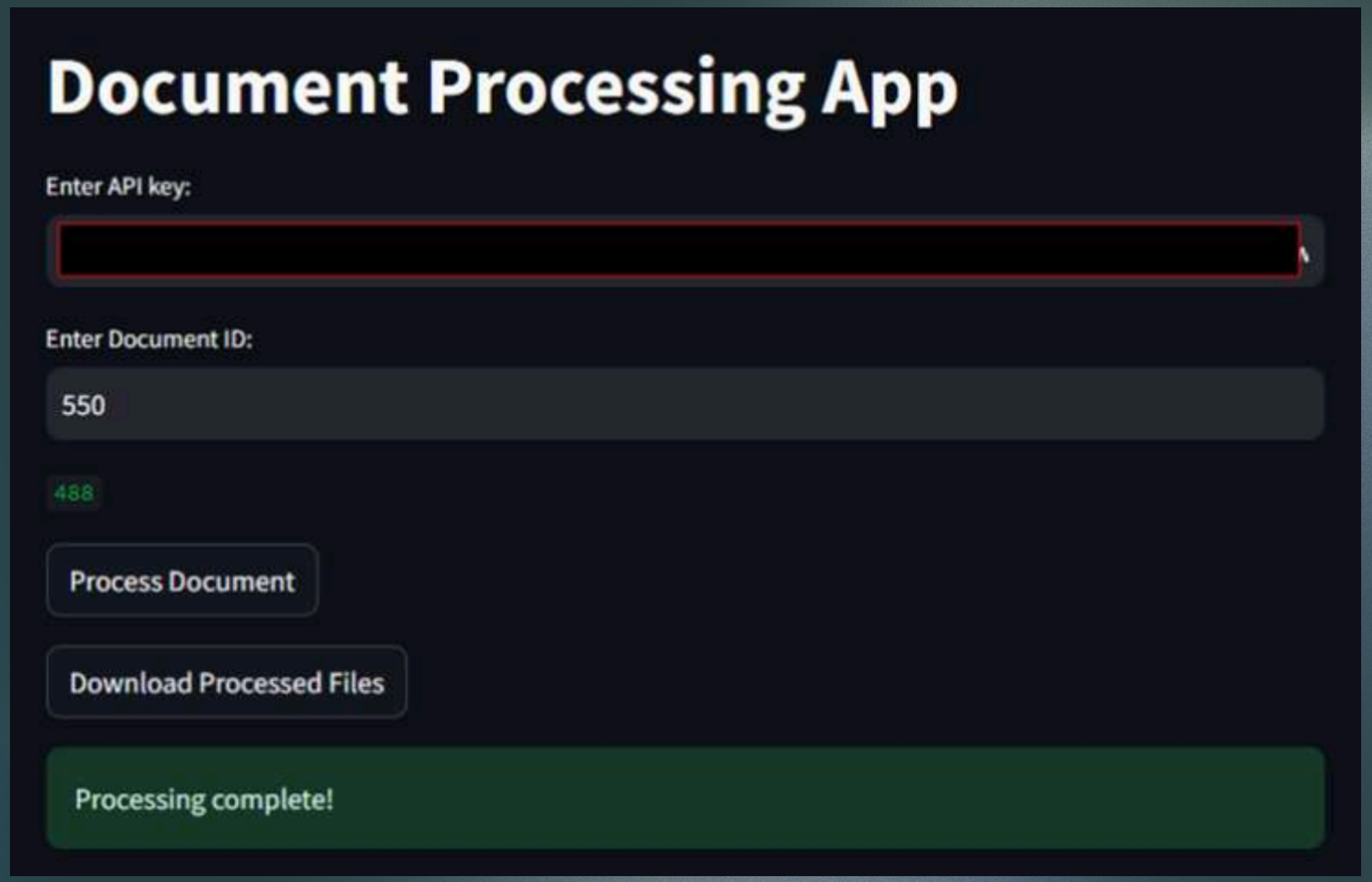
MOST COMMON ERRORS	SOME UNCOMMON ERRORS
“Information not found”	The LLM read the text backward
Error for the application reading inside tables in pdf format and txt ocr format	LLM failed extracting good samples from txt/Json/XML for the higest concentration value
OCR misplacing values inside tables from unstructured data	The LLM described the correct sample without giving its ID
LLM confuses the descrption from one sample to another	
Provides correct answers but incomplete description	
Errors reading the Json/Xml due to the application splitting formats	
If there is a "0" in front of a coordinate number the LLM takes it out	

WORKFLOW AUTOMATION

Custom and time saving web application to increase the speed of the process.
Accessible for everyone in the team.

Characteristics:

- Selection of one GM document.
- Used Agile DD's APIs:
  - Extraction of text from the PDF.
  - Extraction of tables from the PDF.
- Conversion from csv tables to JSON and XML.
- Compression of the text and tables files into a zip file.
- Download of the zip file containing all the inputs for the creation of configuration files.



STANDARDIZATION

Template for data entry:

We created a template in Excel so each member was able to test the GM documents and load the results in a consistent way.

Also, reference data was created to limit the different ways of entering the results from the tests.

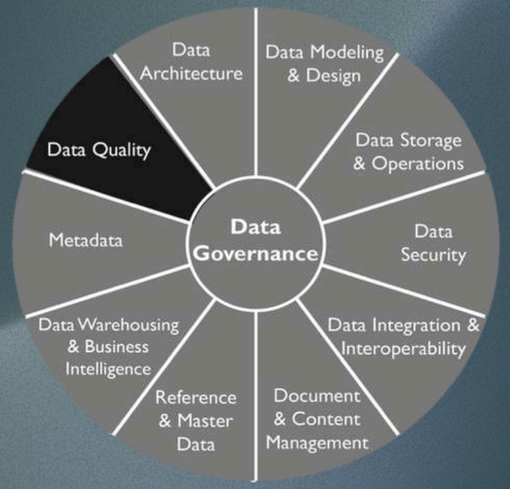
This improved the quality of the gathered data in several dimensions:

- Consistency (same template for all)
- Completeness (the template is like a form, so it is difficult to forget values)
- Validity (reference lists)

Standardized File Naming :

Structured naming conventions applied systematically to track document versions and configurations clearly:

- GMXXXXX_A_ocr_pdf
- GMXXXXX_B_extracted_text
- GMXXXXX_xml or GMXXXXX_json
- GMXXXXX_D_extracted_text_plus_xml
- GMXXXXX_C_extracted_text_plus_json



DATA MODEL

Data Dictionary

Field	Type	Description	Constraints
author_id	int	Unique identifier for the author	Primary Key
author_name	str	Name of the author	Not Null

2. Document

Field	Type	Description	Constraints
document_id	int	Unique identifier for each document	Primary Key
document_name	str	Name or title of the document	Not Null
number_of_pages	int	Total number of pages in the document	Not Null, >= 1
number_of_tables	int	Number of tables present in the document	Not Null, >= 0

3. Prompt

Field	Type	Description	Constraints
prompt_id	int	Unique identifier for the prompt	Primary Key
prompt_std_msg	str	Standardized message used in the prompt	Not Null

4. Configuration

Field	Type	Description	Constraints
configuration_id	str	Unique identifier for a configuration setup	Primary Key
contains_text_tables	bool	Indicates if text-based tables are included	Not Null
contains_json_tables	bool	Indicates if JSON-based tables are included	Not Null
contains_xml_tables	bool	Indicates if XML-based tables are included	Not Null
location_of_added_tables	str	Where additional tables are added to the extracted text	Not Null

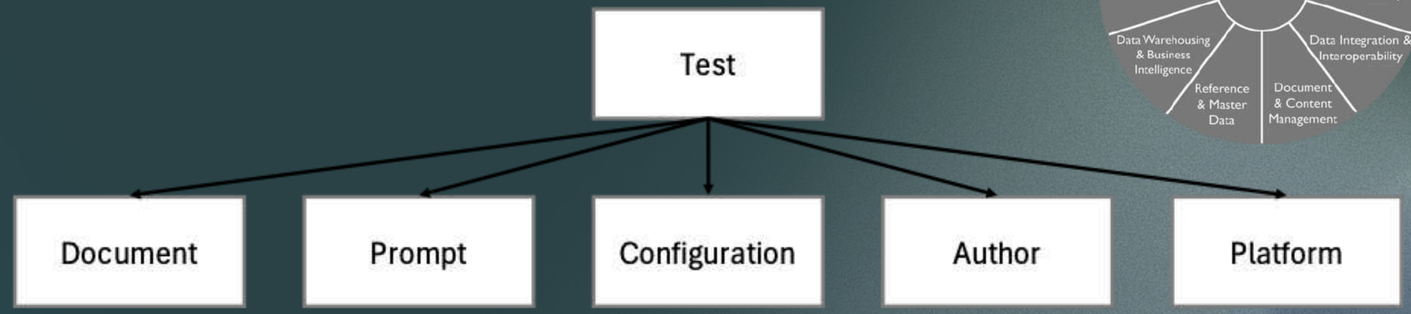
5. Test

Field	Type	Description	Constraints
test_id	int	Unique identifier for a test instance	Primary Key
platform_id	int	Platform used for the test	Foreign Key (Platform)
document_id	int	Document being evaluated	Foreign Key (Document)
author_id	int	Author conducting the test	Foreign Key (Author)
configuration_id	str	Configuration used for the test	Foreign Key (Configuration)
prompt_id	int	Standard prompt used in the test	Foreign Key (Prompt)
prompt_custom_msg	str	Customized version of the prompt	Not Null
answer	str	Extracted value from the document	Not Null
pertinence	bool	Indicates if the answer is relevant	Not Null
correct_coordinates	bool	Whether extracted coordinates were correct	Not Null
correct_sample	bool	Whether the correct sample was identified	Not Null
correct_concentration	bool	Whether the concentration value is correct	Not Null
correct_description	bool	Whether the extracted description is correct	Not Null
remarks	str	Additional comments on test results	Optional

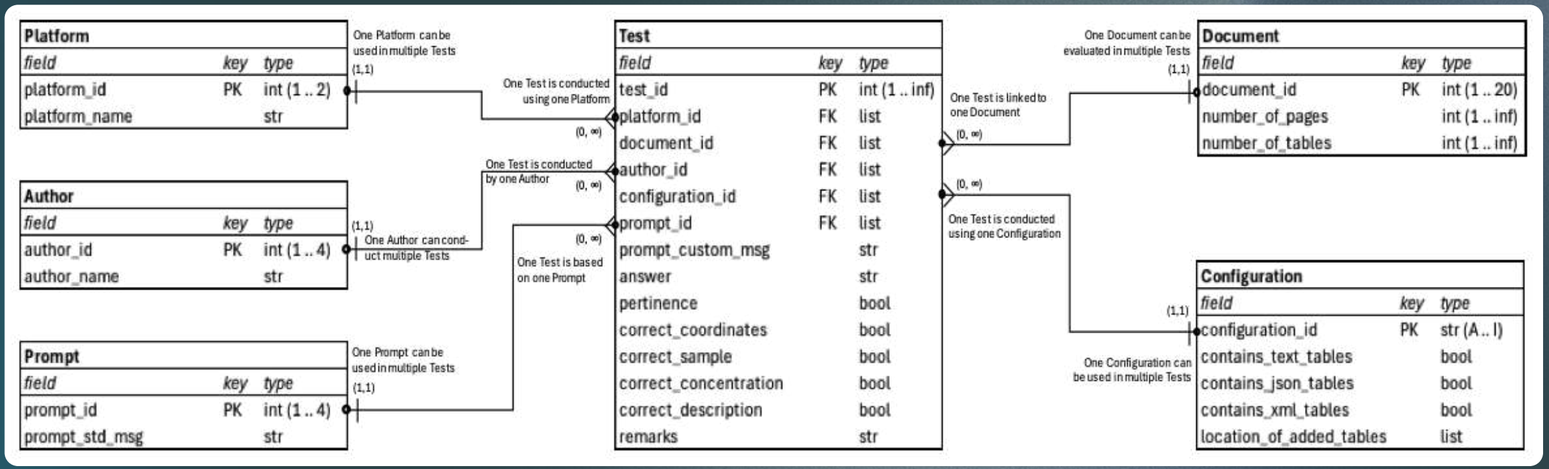
6. Platform

Field	Type	Description	Constraints
platform_id	int	Unique identifier for the platform	Primary Key
platform_name	str	Standardized message used in the prompt	Not Null

Conceptual Data Model

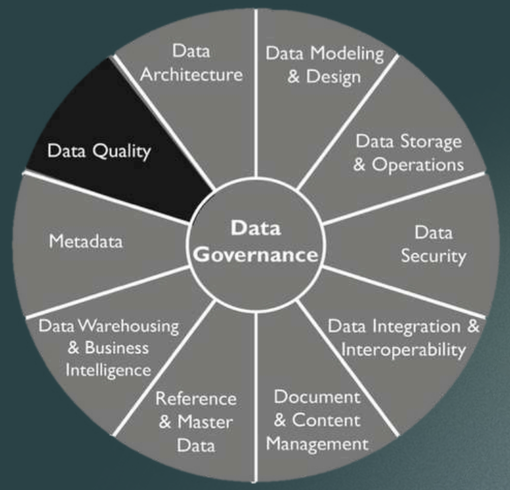


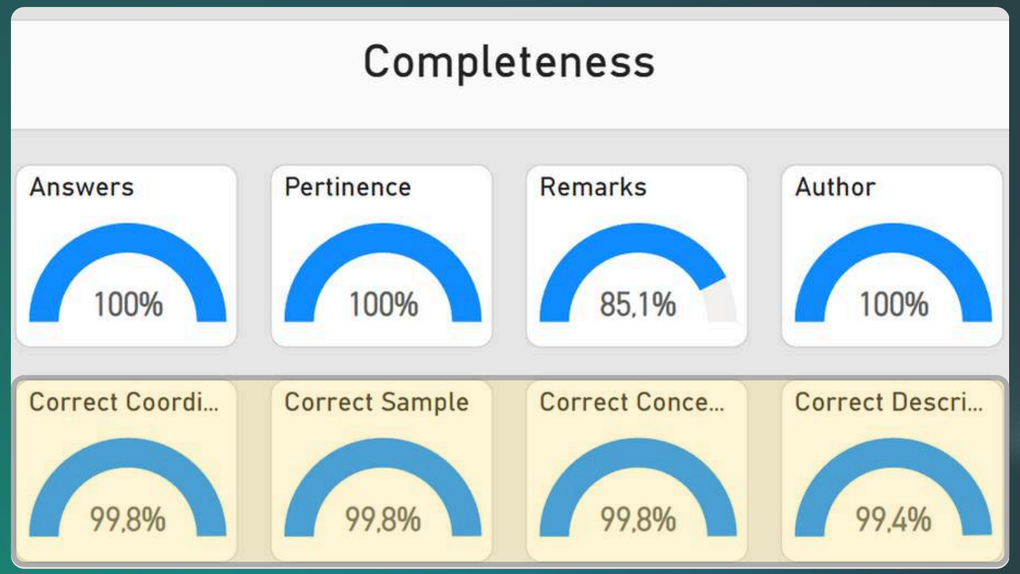
Logical Data Model

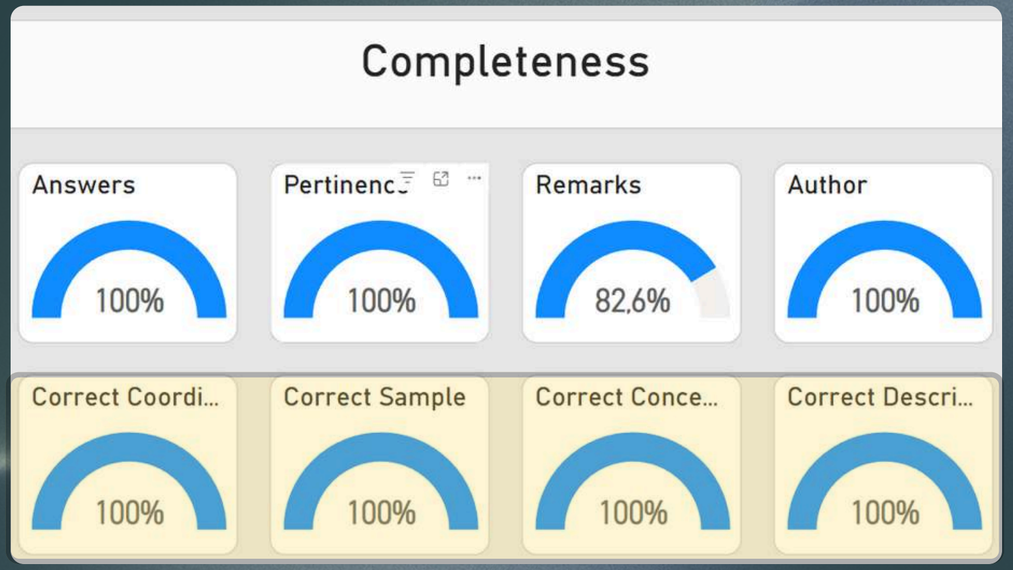


DATA QUALITY

- Data quality dashboard
- Easily detect errors
- Mostly regarding the completeness dimension

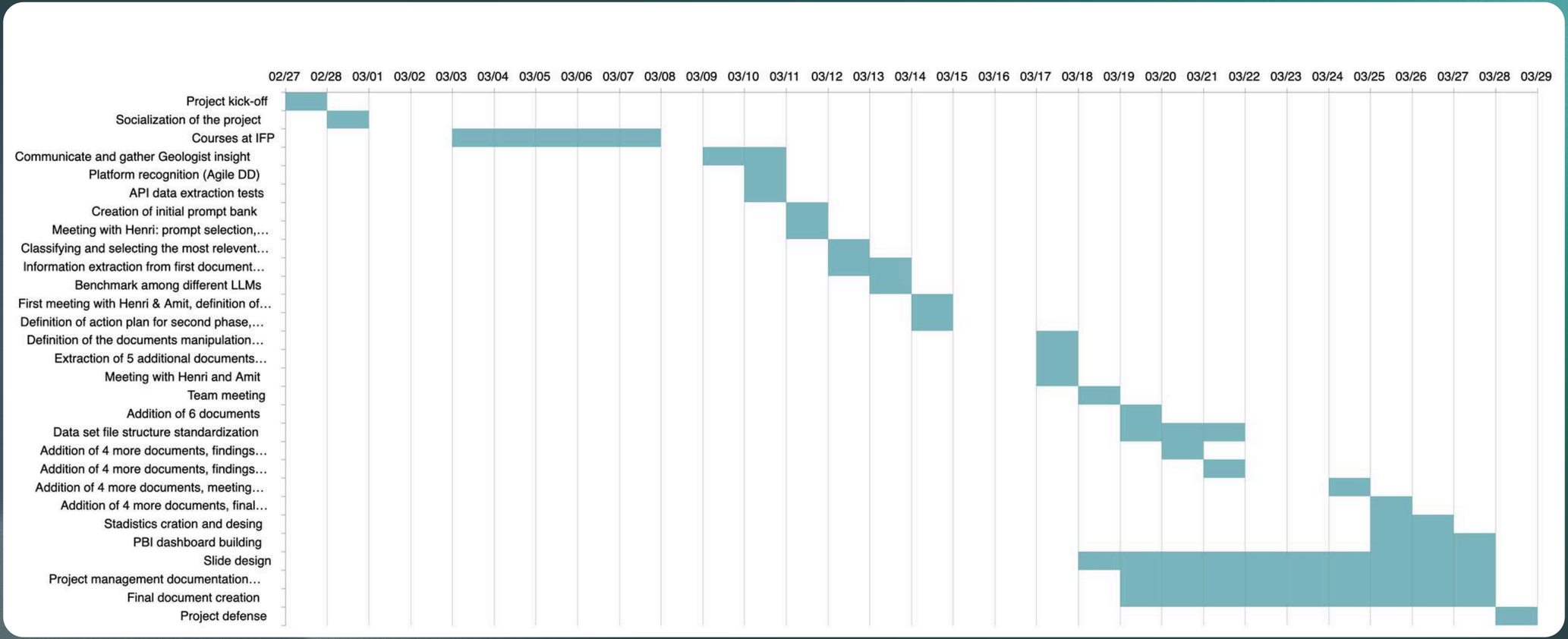




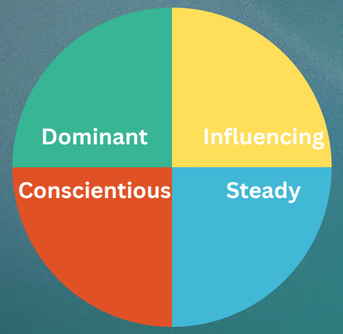


OUR TEAM

WORKING SCHEDULE



WORK BY TASK



	Full Team	Juan SIVORI	Stephania ZAMBRANO	Aicha LE BOU	Samuel NAUTIN
Phase 1	• Recognized the platform. • Tested API data extraction. • Created the initial prompt bank. • Extracted data from the first document (GM70036). • Held daily team meetings. • Held weekly meetings with tutors.	• Automated the workflow for downloading input files through a web-based app. • Defined configuration settings for documents.	• Project Manager • Defined the workflow inspired by agile methodology. • Facilitated communication with tutors and the team. • Documented processes. • Established the document structure for meeting minutes.	• Conducted a benchmark among different LLMs.	• Communicating and gathering Geologist insight • Compilated and Classifying and selecting the most technical questions out of the bank • Presentation on document classification
Phase 2	• Defined new objectives. • Designed an action plan. • Standardized document manipulation processes. • Extracted data from the first five documents. • Held daily team meetings. • Held weekly meetings with tutors.	• Automated and standardized data entry in order to store consistent prompt results. • Calculated statistics to assess configuration performance.	• Developed a new workflow and delegated tasks. • Created the project schedule. • Supervised task completion to ensure progress.	• Organized Excel files daily according to new versions. • Extracted and categorized remarks.	• Configuration definition for documents • Define naming convention for documents • Standardized data entry in order to store consitent prompt results.
Phase 3	• Standardized the dataset file structure. • Added 15 more documents to complete the representative sample. • Discussed and presented the results to Agile DD CEO and tutor • Built visuals and schemas for processes • Created the deliverables	• Created the data model to represent the data gathered. • Developed the final dashboard with performance and data quality metrics.	• Refined the workflow, delegated tasks, and supervised execution. • Ensured the quality of deliverables. • Created the final presentation. • Finalized and adjusted the project document.	• Created the first version of the presentation. • Adjusted the final presentation. • Adjusted the final report	• Creation of final document (structure, visuals, table, redaction)

STRATEGIES

AGILE-INSPIRED WORKFLOW

Adaptive Project Management

- Adopted an Agile-inspired methodology for flexibility and responsiveness.
- Worked iteratively, adjusting tasks and objectives based on evolving insights.

Daily Meetings

- Morning sessions: Defined daily goals and assigned responsibilities.
- Evening sessions: Reviewed progress, resolved issues, and outlined next steps.

Weekly Progress Reviews

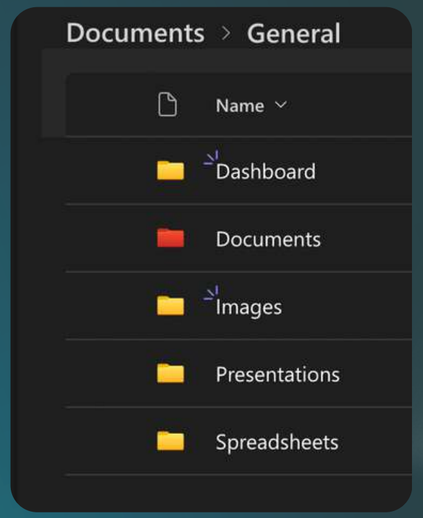
- Regular meetings with our tutor (client role) to validate results.
- Continuous refinement of strategies informed by weekly feedback.

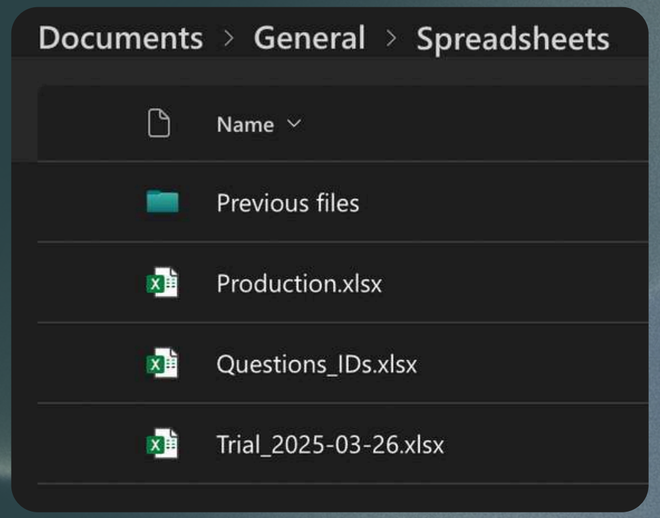
STRATEGIES

FILE ORGANIZATION & DOCUMENT TRACKING

Organized Collaboration with Microsoft Teams
All files were sorted into shared folders (Spreadsheets, Documents, Presentations, Dashboards, Images)

Daily evaluations performed using a standardized Excel template:
New tracking sheets created daily (e.g Trial_2025-03-24.xlsx), with previous versions archived systematically for version control and data traceability.





Thank You