Learn Basics
IIT, NIT Alumni Venture

Task 1
Prompt Engineering

ID: PE/2507/IN-1

Dear Candidate,

This task is designed to evaluate your prompt engineering and AI utilization skills in solving a structured data extraction problem from NCERT educational resources. You can use Google Gemini APIs or any equivalent APIs to fetch, analyse, and structure content from Grade 8 NCERT Science textbooks.

Task Description:

1. Create one universal prompt that can handle different chapters of the Class 8 NCERT Science textbook. Assume that a teacher will later use the extracted data and the knowledge graph of the curriculum to teach students, so your prompt must ensure that the output is structured clearly and consistently, with well-defined relationships between chapters, topics, sub-topics, and various content types like activities, examples, figures, and questions.
2. Also create another prompt to create a study/teaching planner for the four chapters, which should include the topic split up, duration, day wise plan. Constraint the number of days should be given by the user of the prompt. This study planner should be created based on the excel what you obtained in step 1 of content extraction.
3. Execute the above prompts through python using Google APIs or any other equivalent APIs of your choice.

Task Input Materials:

Class 8 Science NCERT Book links:

- https://ncert.nic.in/textbook/pdf/hesc106.pdf
- https://ncert.nic.in/textbook/pdf/hesc107.pdf
- https://ncert.nic.in/textbook/pdf/hesc108.pdf
- https://ncert.nic.in/textbook/pdf/hesc113.pdf

Task Details:

1. Create a Detailed /Well structed Prompt:

1.1 Design a standard, reusable prompt that can extract structured content from Science NCERT textbooks.

1.2 The prompt should extract content chapter-wise and topic-wise without adding any extra interpretation or summary. Only extract the following elements wherever they appear:

1.2.1. Topic names

1.2.2. Sub-topic headers

1.2.3. Paragraph, Images, Diagrams, tables, examples, and exercises

1.2.4. Activities, questions, External Sources or boxed facts under their respective sections

1.2.5. Sample output – refer 4.1.1

2. **JSON & Excel Generation**
   3.1 The output from the AI (or extraction logic) must be stored in a structured JSON format as described in sample 4.1.2.
   3.2 Use this JSON file to generate a well-formatted Excel (.xlsx) file with the same structure as sample 4.1.1.
   3.3 The Excel sheet will be used to visually validate your extraction quality, so ensure clarity, accuracy, and consistency in the layout.
   3.4 After generating the Excel file, create a text-based knowledge graph representation of the extracted content.
   This knowledge graph should clearly show the relationships between:
   - Chapters
   - Topics
   - Sub-topics
   - Linked content types (e.g., Paragraphs, Activities, Tables, Images, Questions, etc.)

3. **Use of AI Tools & Methods**

3.1 Use Python along with Google API with free credits available through your Google account or language model APIs to automate the extraction.

3.2 You are free to:

3.3.1. Process the PDF manually for prompt creation.
3.3.2. Use AI systems that first search the textbook for relevant information and then generate accurate answers based on that content (this is called Retrieval-Augmented Generation or RAG)
3.3.3. You can use any tool or method you prefer. What matters most is the accuracy and structure of the extracted output, not the specific approach you take to achieve it.

4. **Output Format**
   This is the expected output format
   4.1 **Science:** Copy of Science-sample-output.xlsx
   4.2 **JSON:** chapter-extract.json

5. Using the extracted data above, create a comprehensive study planner for the listed chapters. The planner should include a clear topic-wise breakdown, estimated duration for each topic, and a structured day-wise schedule. You may design the format and layout of the planner as per your preference to ensure clarity and usability. The minimum duration for the study planner must align with the range specified in the prompt, which is between **5 days a 1 month**, depending on the content and depth of the chapters.

Submission Guidelines:

1. After completing the task, submit the output through the google form link attached with the task message.
2. Submit the GitHub repository link in the google form which contains the following files:
   2.1 Python Code used

2.2 The prompt used for extracting the output
2.3 Json Output File
2.4 Excel Output File
2.5 Study planner document (Format – PDF or Image or any other format of your choice)

Evaluation Criteria:

You will be assessed based on:

1. Accuracy and completeness of the extracted data (Weightage - 35%)
2. Robustness and reusability of your prompt (Weightage - 25%)
3. Clarity and structure of your documentation (Weightage – 20%)
4. Process plan and technical knowledge (Weightage - 20%)

Best wishes!