Goal:
Design a benchmark to evaluate the performance of a grading system.

The inputs to the grading system would be: Question, ground truth, and a candidate answer.
Output of the system would be: Score for the question and a reasoning for that score.

Designing the benchmark:

- Given a few subject specific pdfs, we need 3 things
    - a prompt to generate QA from that pdf
    - a prompt to rephrase the QA if needed
    - a prompt to evaluate the candidate answers
    - a prompt to generate sample candidate answers

1.) Prompt to generate QA from a pdf: (replace subject matter X with your subject)

You are an expert in [Subject Matter X], creating quiz questions to test understanding of the material.

Given the following text from [Subject Matter X], please generate at-least 50 such questions and answers pairs based on the text such that they capture all the knowledge from the text.

In your output, include the exact phrase from the text that contains the answer to the question as "Context". This phrase must be copied verbatim from the text with no modifications or truncations.

You must obey the following criteria:

- The question must contain all information and context necessary to answer it without the text. Do not reference the text in the question.
- The question must ask about a fact that is newly and uniquely introduced in the text, not about general knowledge outside the information provided.
- You must specify the units of the answer in the question, if applicable. For example, "in millions of US dollars" or "in kilograms".
- Do not ask about dates (e.g., "What date did this happen?") or other non-numerical information, unless it is crucial to the subject matter.
- The answer must be clear and unambiguous.
- If there are no possible questions that meet all of these criteria, return 'None' as the question.

Output should be in JSON format.

Example Response:

[{ "Question": "What is the acceleration due to gravity on Earth's surface in meters per second squared?", "Answer": "9.8", "Context": "The acceleration due to gravity on Earth's surface is approximately 9.8 meters per second squared." }...]

2.) **Prompt to rephrase the QA if needed:**

You are a skilled editor specializing in paraphrasing quiz questions and answers.

Given the following question and answer, rephrase both so that they significantly change the structure and wording without altering the meaning.

Your output should be in the same format as input.

Do not include any additional information or comments.

3.) **Prompt to evaluate the candidate answers:**

Please act as an impartial judge and evaluate the quality of the candidate's responses to the user's questions below, focusing on correctness and helpfulness. For each question, you will be provided with a reference answer and the assistant's answer. Perform the following steps for each question:

1. Compare the candidate's answer with the reference answer, identifying any mistakes or omissions.
2. Provide an objective explanation highlighting the differences and corrections.
3. Rate the candidate's response on a scale from 1 to 10.

Output your evaluation for each question in the following JSON format:

```
[{
  "question": "{question}",
  "explanation": "{your explanation}",
  "rating": {rating}
},...
```

4.) **Prompt to generate sample candidate answers:**

You are an expert in [Subject Matter X], answering quiz questions to test understanding of the material.

Your task is to provide answers to the following questions using your knowledge.

Guidelines:

- If you do not know the answer, reply with "I don't know".