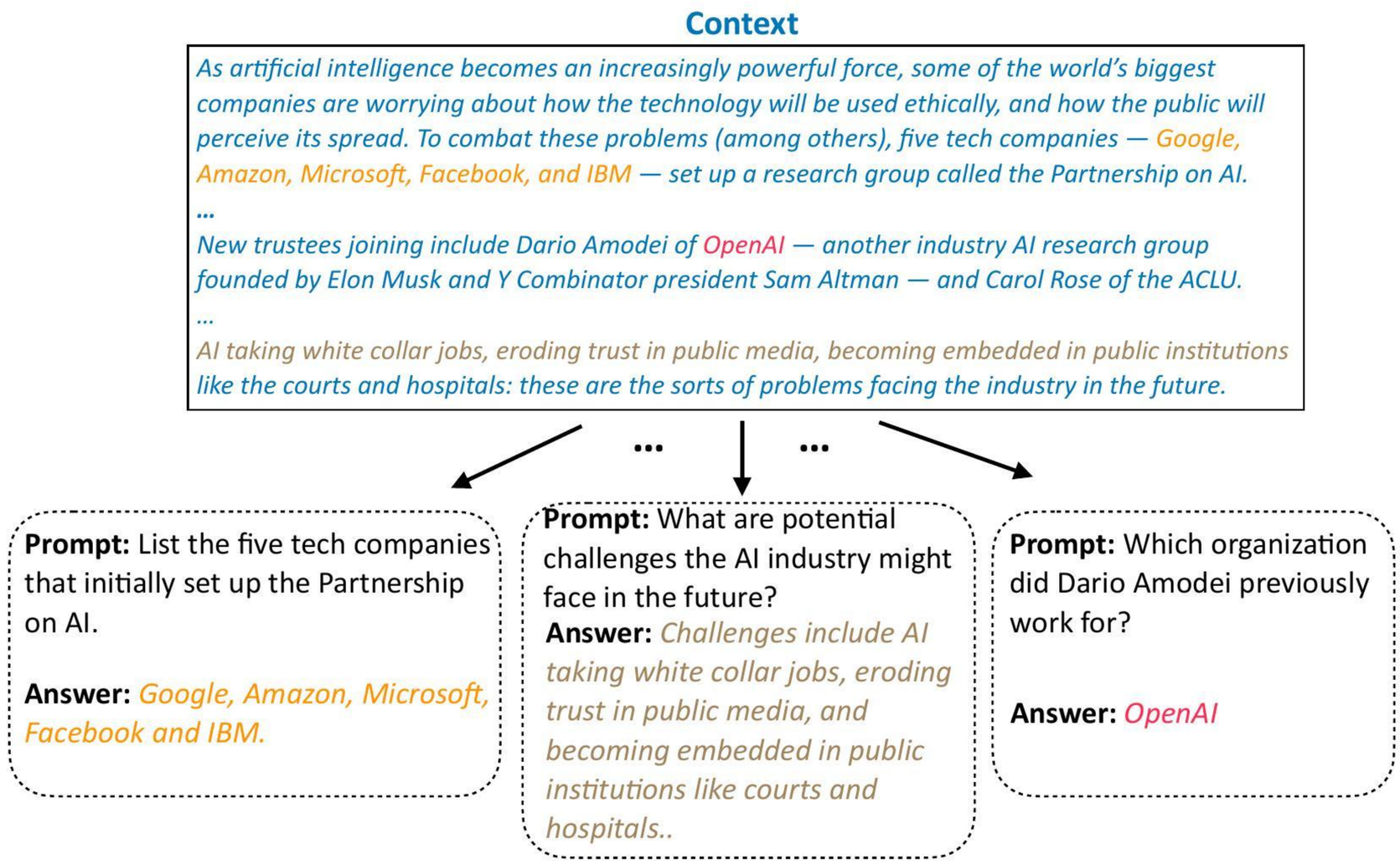

Figure 9: Construction of the PwC dataset: we use the GPT-4 to generate a variety of prompt-answer pairs according to contexts. The resulting dataset is used for instruction fine-tuning (240k for training) and evaluation (18k for testing) in this work.

```
aspects (e.g., topic, genre, structure, style, polarity, key information
and details) of the text as possible. The first half of these prompts
should be like an instruction, the other should be like a question. In
addition to the prompts specified to the above text, please also design
5 general prompts like "rephrase the above text", "summarize the above
text", "write a title for the above text", "extract a few keywords for
the above text" and "write a paragraph (i.e., continuation) that follows
the above text". Each prompt should be outputted in the following
format: [{"prompt": your generated prompt, "answer": the answer to the
prompt}]
```

D GPT-4 Evaluation

According to Mu et al. (2023), we formulate an evaluation prompt to be used with the GPT-4 API. The prompt, as illustrated in Listing 2, consists of a task description along with three specific examples. We supply GPT-4 with a text, a prompt, and two distinct model-generated responses. The task for GPT-4 is to determine the superior answer or recognize a tie. The chosen examples encompass scenarios where Assistant A performs better, Assistant B performs better, and when a tie occurs. This methodology enables us to effectively assess[7] the model's quality. Specially, the orders where the model responses are presented to the GPT-4 are swapped randomly to alleviate bias, as Touvron et al. (2023b) did.

Listing 2: Prompt for the GPT-4 evaluation. This prompt consists of a description of the task and three specific examples.

```
Given a piece of text, an instruction for this text, and two AI
assistant answers, your task is to choose the better answer and provide
reasons. Evaluate the answers holistically, paying special attention to
```

[7]We find the GPT-4 rater tends to prefer longer responses, aligning with observations from recent work such as Zhao et al. (2024). Given that ICAE's responses are generally short (due to instruction fine-tuning with the PwC dataset), its actual performance should be better than the numbers reported in the evaluation.