Experiment evaluates the overall effectiveness of the prototype (Use ‘streamlit run stUIforAPI.py’

or ‘python defaultUI.py’ to reproduce the experiment.) compared to a baseline that directly uses GPT-4.

Both methods are provided with the same datasets, background knowledge, and questions to generate FAQ-like textual reports.

Participants are recruited, they were given an assessment guide and random assignment reports. They assess whether each FAQ-like textual report as an answer is correct and, if incorrect, classify the error type.

The collected annotations are aggregated and analyzed to compare the accuracy, hallucination rate, and error distribution between the two methods.