Experiment is an ablation study designed to assess the individual contribution of the answer template module (Use ‘streamlit run intermediatePrototype.py’ to reproduce the experiment).

The ablated prototype (without the answer template module) is given the same datasets, background knowledge, and questions used in Experiment 2 to generate a new set of FAQ-like 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.