Assessing fairness of AI systems for education – Attachments

Summary

Appendices

Appendix 1 – AI-Driven Assistant UI

The following images show the user interface of the AI-Driven Documentation assistant tool used to assess the semantic completeness of the given documentation.

The user can filter questions from TAI-SDF database and display the reports:

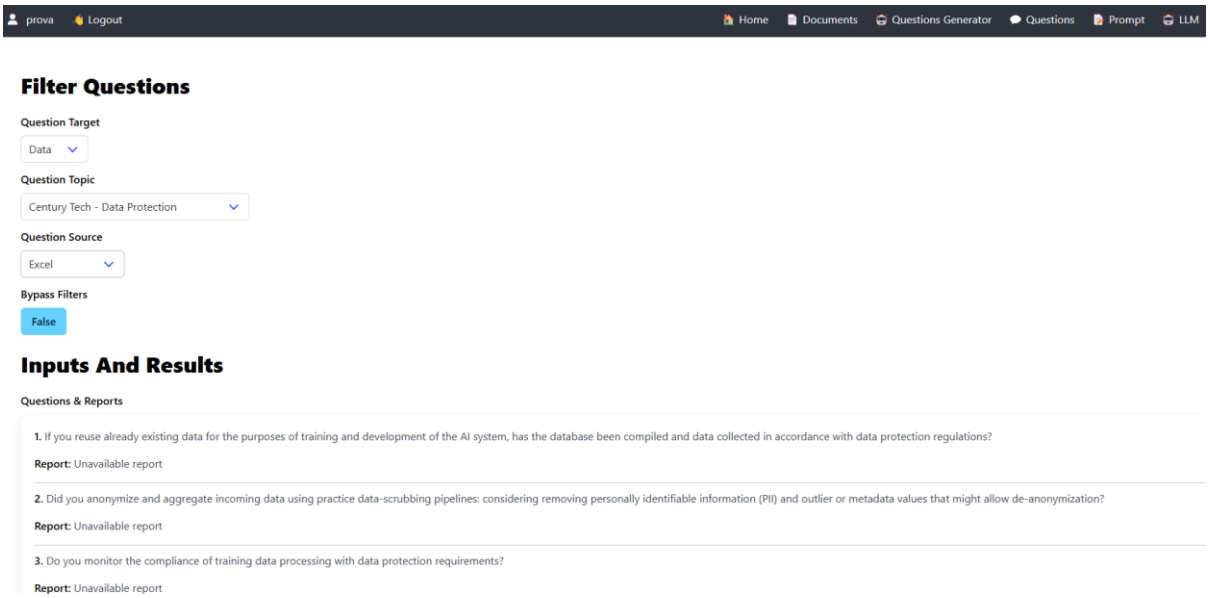


Figure 1: AI Tool UI - Questions

The user can load the documentation from PDF files and edit it, as well as filling the different prompt sections (i.e. task, constraints, examples and additional information):

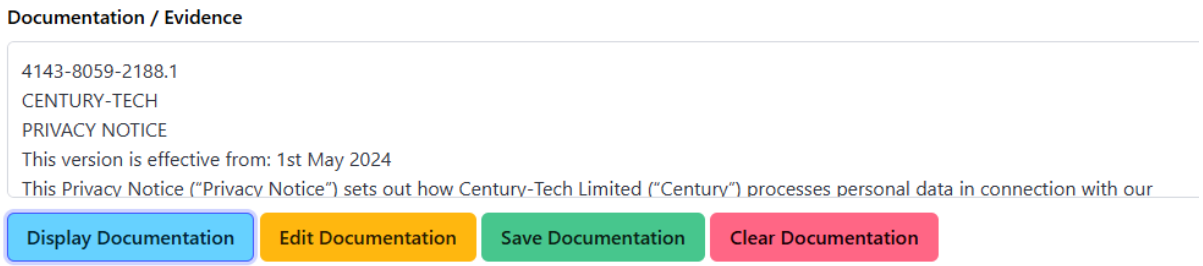


Figure 2: AI Tool UI - Documentation

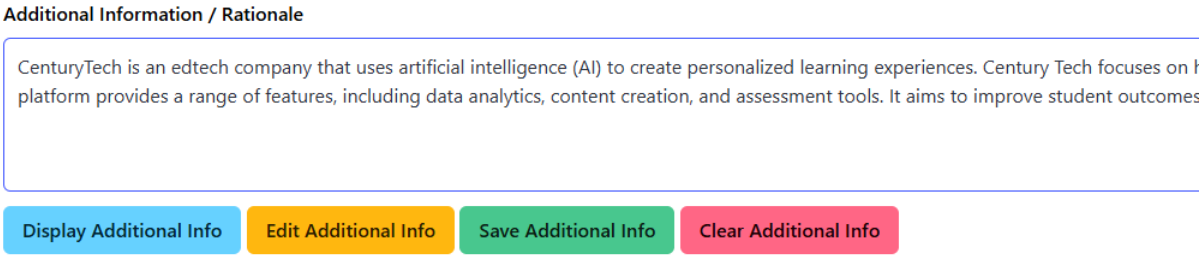


Figure 3: AI Tool UI - Additional Information

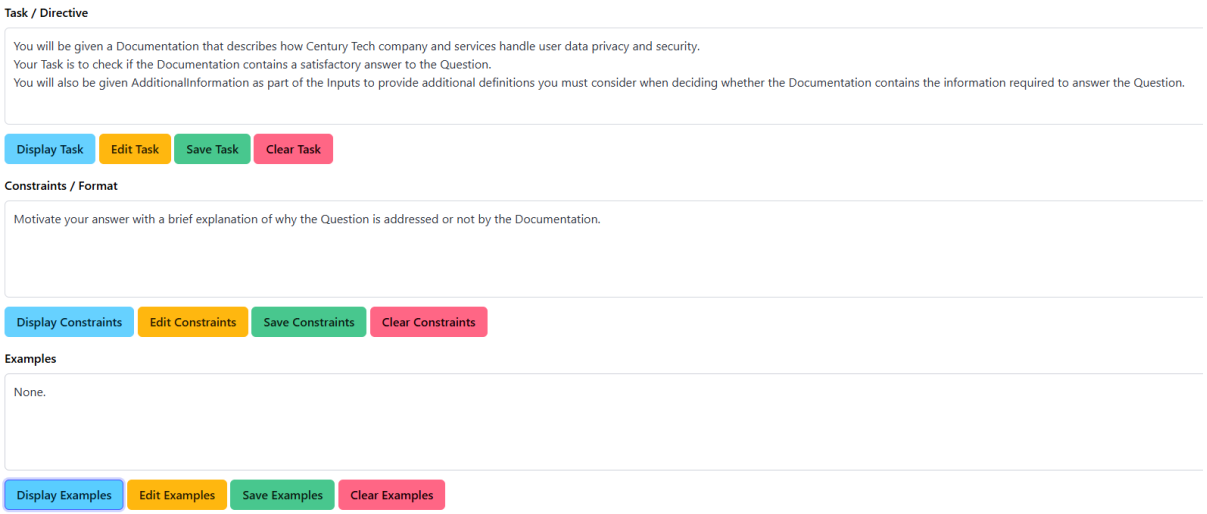


Figure 4 : AI Tool UI - Task, Constraints and Examples

This facilitates the user in assembling a model prompt tailored for the given use case.
Finally, the user can set up the system prompt, select the LLM backend and generate an assessment report for each question:

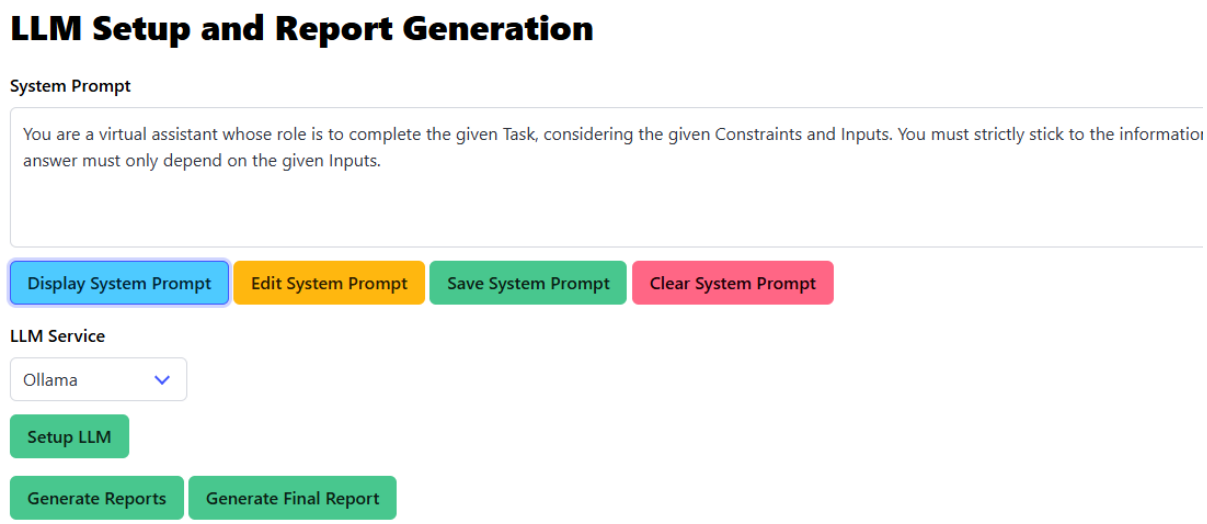


Figure 5: System Prompt and AI model configuration

The configuration of the tool for use case #1 and #2 is shown in the following sections.

Use Case #1: Fairness Assessment of Thrively

This use case analysis has been performed configuring the application with DeepSeek-R1:8B model locally hosted on Ollama as NLM backend.
The natural language model is queried with the following prompt setup:

- *System prompt*: "You are an assistant whose role is to complete the given Task, considering the given Constraints and Inputs. You must strictly stick to the information in the Inputs in order to complete the Task. You are prohibited to infer or make assumptions: your answer must only depend on the given Inputs."

- *Task*: "You will be given in the Inputs a Question related to Thrively software, an AI-based learning platform targeted to discover and develop personal strengths, interests, and aspirations of every student to help them reach their full potential. Thrively can identify and understand the ways in which a student's best learns and his/her habits, as well as helping with identification of their career pathways, connecting their interests and personality types to the world of work. In the Inputs you'll also be given Documentation about Thrively that contains the information you must consider answering the Question. Your Task is to check if the Documentation contains a satisfactory answer to the Question. You will also be given Additional Information as part of the Inputs to provide additional definitions you must consider when deciding whether the Documentation contains the information required to answer the Question."

- *Constraints*: "Motivate your answer with a brief explanation of why the Question is addressed or not by the Documentation."

- *Additional Information*: none.

- Examples: none.

The considered TAI-SDF questions are the following:

1. Do you compute fairness metrics in evaluating the AI solution?
2. Does the AI solution explain why it made a particular decision or return a specific result?
3. Is the AI solution accessible for persons with disabilities?
4. Did you consider diversity and representativeness of end-users and/or subjects in the data?
5. Did you test for specific target groups or problematic use cases?
6. Did the AI solution ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?
7. Did you assess the risk of the possible unfairness of the system onto the end user's communities?
8. Did the AI solution write clear privacy notices for children so that they are able to understand what will happen to their personal data and what rights they have?
9. In which way the result of the profiling made by the AI solution affect the children?
10. Can the AI solution stop profiling of children if they ask you for?

Use Case #2: Data Security of Century Tech

This use case analysis has been performed configuring the application with DeepSeek-R1:8B model locally hosted on Ollama as NLM backend.
The natural language model is queried with the following prompt setup:

- *System prompt*: "You are an assistant whose role is to complete the given Task, considering the given Constraints and Inputs. You must strictly stick to the information in the Inputs in order to complete the Task. You are prohibited to infer or make assumptions: your answer must only depend on the given Inputs."

- *Task*: "You will be given a Documentation that describes how Century Tech company and services handle user data privacy and security. Your Task is to check if the Documentation contains a satisfactory answer to the Question. You will also be given AdditionalInformation as part of the Inputs to provide additional definitions you must consider when deciding whether the Documentation contains the information required to answer the Question."

- *Constraints*: "Motivate your answer with a brief explanation of why the Question is addressed or not by the Documentation."

- *Additional Information*: "CenturyTech is an edtech company that uses artificial intelligence (AI) to create personalized learning experiences. Century Tech focuses on helping educators and students by offering tools for adapting lessons to individual learning styles and needs. The platform provides a range of features, including data analytics, content creation, and assessment tools. It aims to improve student outcomes by making learning more tailored and efficient."

- Examples: none.

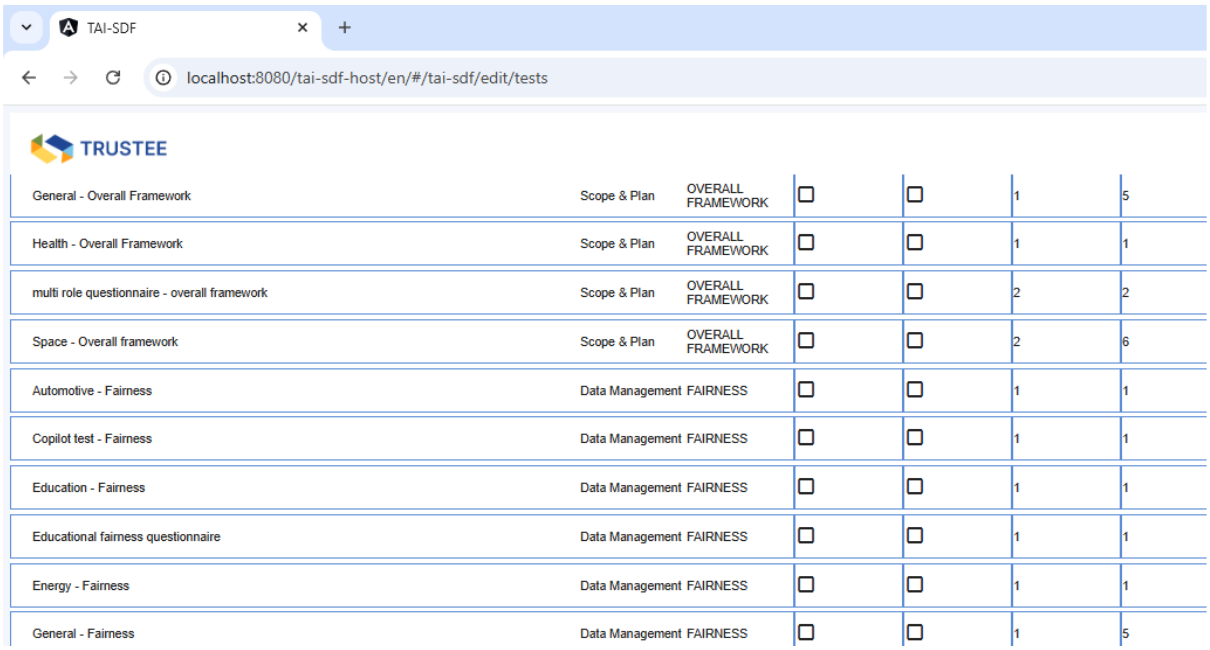
The considered TAI-SDF questions are the following:

1. If you reuse already existing data for the purposes of training and development of the AI system, has the database been compiled and data collected in accordance with data protection regulations?
2. Did you anonymize and aggregate incoming data using practice data-scrubbing pipelines: considering removing personally identifiable information (PII) and outlier or metadata values that might allow de-anonymization?
3. Do you monitor the compliance of training data processing with data protection requirements?
4. Did you take measures to enhance privacy-by-design, such as via encryption, pseudonymization, anonymization and aggregation?
5. Did you handle any sensitive data with care: complying with required laws and standards, providing users with clear notice and give them any necessary controls over data use, following best practices such as encryption in transit and rest?
6. In case of data processing for vulnerable individuals (children, patients, employees, etc), did you define and if so, do you implement a storage policy and further privacy strategies (e.g. minimization, hiding, separation or abstraction) for the personal data?
7. Did you assess the degree of anonymization and possible risk of re-identification?
8. Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle?

9. Did you define the requirements concerning data protection and security at the origin, while taking into account the available standards and best practices?
10. Did you define the time limits for erasure of stored personal data?
11. Did you put in place procedures to verify the implementation of the storage periods?
12. Did you define and put in place the necessary and appropriate technical measures for the storage of personal data?
13. Did you define and put in place the necessary and appropriate organizational measures for the storage of personal data?

Appendix 2 – TAI-SDF Questionnaire UI

Figure 6 and Figure 7 show the user interface of the Questionnaire based assessment tool. Figure 6**Errore. L'origine riferimento non è stata trovata.** shows the questionnaires list:



General - Overall Framework	Scope & Plan	OVERALL FRAMEWORK	☐	☐	1	5
Health - Overall Framework	Scope & Plan	OVERALL FRAMEWORK	☐	☐	1	1
multi role questionnaire - overall framework	Scope & Plan	OVERALL FRAMEWORK	☐	☐	2	2
Space - Overall framework	Scope & Plan	OVERALL FRAMEWORK	☐	☐	2	6
Automotive - Fairness	Data Management	FAIRNESS	☐	☐	1	1
Copilot test - Fairness	Data Management	FAIRNESS	☐	☐	1	1
Education - Fairness	Data Management	FAIRNESS	☐	☐	1	1
Educational fairness questionnaire	Data Management	FAIRNESS	☐	☐	1	1
Energy - Fairness	Data Management	FAIRNESS	☐	☐	1	1
General - Fairness	Data Management	FAIRNESS	☐	☐	1	5

Figure 6: Available Questionnaires

Figure 7 briefly shows the UI to create a new questionnaire:

Figure 7: Creation of a New Survey