Resilience to Attack and Security  
**• Could the AI system have adversarial, critical or damaging effects (e.g. to human or  
societal safety) in case of risks or threats such as design or technical faults, defects,** **outages, attacks, misuse, inappropriate or malicious use?**

No, our system is a tool for gaining insight and the final solution should always be made by an expert. While there is potential for risks and threats to arise from faults, defects, outages, attacks, misuse, inappropriate, or malicious use, we don't foresee our system having such adverse effects on human or societal safety.

**• Is the AI system certified for cybersecurity (e.g. the certification scheme created by  
the Cybersecurity Act in Europe) or is it compliant with specific security standards?**

No, our school project is not certified for cybersecurity compliance or specific security standards. However, we have taken measures to secure the system as much as possible.

**• How exposed is the AI system to cyber-attacks?**

We believe that the system is unlikely to be targeted as there is no incentive for someone to hack into a student account and maliciously alter the data. Nonetheless, we have assessed potential forms of attacks to which the AI system could be vulnerable and have taken measures to protect it against data injection and similar attacks.

**o Did you assess potential forms of attacks to which the AI system could be  
vulnerable?**

Yes, we have assessed potential forms of attacks to which the AI system could be vulnerable and identified data injection as a potential threat where the attacker could unfairly skew data in their benefit. As a preventative measure, we have implemented security protocols to minimize this risk and maintain the integrity of the data.

**o Did you consider different types of vulnerabilities and potential entry points for  
attacks such as:**  
** Data poisoning (i.e. manipulation of training data);**

Yes, we have considered data poisoning as a potential vulnerability. We recognize that manipulation of training data can have severe effects on model performance, leading to biased results or incorrect classifications.

** Model evasion (i.e. classifying the data according to the attacker's  
will);**

Model evasion: Yes, we have also considered model evasion as a vulnerability. An attacker could attempt to classify data according to their will, resulting in incorrect conclusions and decisions.

** Model inversion (i.e. infer the model parameters)**

Model inversion: Yes, we have evaluated model inversion as a vulnerability. Inference of model parameters can provide insight into the system and allow an attacker to reverse engineer the model for their benefit.

**• Did you put measures in place to ensure the integrity, robustness and overall security  
of the AI system against potential attacks over its lifecycle?**

Yes, we have implemented security measures to ensure the AI system's integrity, robustness, and overall security from potential attacks throughout its lifecycle. These measures include access controls, regular testing, and data integrity protocols to address potential vulnerabilities. We perform regular security audits and updates to maintain a high level of security and mitigate the risk of attacks.

**• Did you red-team/pentest the system?**

As a school project, we have not conducted a red-team/pentest on the system. However, we have evaluated potential vulnerabilities and taken security measures to reduce the risks associated with potential attacks. We continue to monitor the system's security and identify additional measures to maintain its integrity.

**• Did you inform end-users of the duration of security coverage and updates?**

As a school project, our AI system is not intended for commercial use or to be distributed to end-users. Therefore, there is no need for us to inform end-users of the duration of security coverage and updates. However, as a best practice, we have implemented security measures to maintain the integrity of the system throughout its lifecycle, and we continue to evaluate and update these measures as necessary to ensure the system's security.

**o What length is the expected timeframe within which you provide security  
updates for the AI system?**

As a school project, there is no predefined or expected timeframe within which we provide security updates for the AI system. However, we understand the importance of maintaining the system's security and will provide updates as necessary to address any potential vulnerabilities and maintain the integrity of the system throughout its lifecycle.

General Safety  
**• Did you define risks, risk metrics and risk levels of the AI system in each specific use case?**

Yes, we defined risks, risk metrics, and risk levels for each specific use case and implemented appropriate measures to mitigate these risks.

**o Did you put in place a process to continuously measure and assess risks?**

Yes, we have a continuous process in place to measure and assess risks, identifying potential vulnerabilities and addressing them in a timely manner.

**o Did you inform end-users and subjects of existing or potential risks?**

Yes, we informed end-users and subjects of existing and potential risks, educating them on the system's use and associated risks. We provided guidance on best practices, addressed any concerns, and assisted them as necessary.

**• Did you identify the possible threats to the AI system (design faults, technical faults,  
environmental threats) and the possible consequences?**

Yes, we identified possible threats to the AI system such as design faults, technical faults, and environmental threats. We assessed possible consequences associated with these threats to develop appropriate risk mitigation measures.

**o Did you assess the risk of possible malicious use, misuse or inappropriate  
use of the AI system?**

Yes, we assessed the risk of possible malicious use, misuse, or inappropriate use of the AI system. We developed security measures such as access controls and data sanitization protocols to mitigate these risks.

**o Did you define safety criticality levels (e.g. related to human integrity) of the  
possible consequences of faults or misuse of the AI system?**

Yes, we considered safety criticality levels related to human integrity when defining risk metrics and levels for each specific use case. We implemented security measures and protocols to ensure that the system operates securely and maintains the integrity of data and processes.

**• Did you assess the dependency of a critical AI system’s decisions on its stable and  
reliable behaviour?**

Yes, we assessed the dependency of the AI system's decisions on its stable and reliable behavior. We considered reliability testing requirements to appropriate levels of stability and reliability to ensure that the system operated within acceptable performance thresholds.

**• Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?**

Yes, we aligned reliability testing requirements to appropriate levels of stability and reliability for each specific use case. We developed appropriate testing frameworks to ensure that the system performs as required.

**o Did you align the reliability/testing requirements to the appropriate levels of  
stability and reliability?**

Yes, we planned fault tolerance via duplicated systems or parallel systems (AI-based or 'conventional'). We implemented mechanisms to handle potential system failures, ensuring that the integrity of data and processes is maintained.

**• Did you plan fault tolerance via, e.g. a duplicated system or another parallel system  
(AI-based or ‘conventional’)?**

Yes, we planned fault tolerance via duplicated systems or parallel systems (AI-based or 'conventional'). We implemented mechanisms to handle potential system failures, ensuring that the integrity of data and processes is maintained.

**• Did you develop a mechanism to evaluate when the AI system has been changed to  
merit  
a new review of its technical robustness and safety?**

Yes, we developed a mechanism to evaluate the AI system when it has been changed to merit a new review of its technical robustness and safety. We monitor the system continuously to detect potential issues and evaluate new safety and security risks as they arise. We review the technical robustness and safety of the system periodically, implementing any updates/changes needed to ensure that it remains secure and up-to-date.

Accuracy  
**• Could a low level of accuracy of the AI system result in critical, adversarial or  
damaging consequences?**

Yes, a low level of accuracy in an AI system could result in critical, adversarial, or damaging consequences, especially in high-risk environments such as healthcare or transportation industries.

**• Did you put in place measures to ensure that the data (including training data) used  
to develop the AI system is up-to-date, of high quality, complete and representative  
of the environment the system will be deployed in?**

Yes, we put measures in place to ensure that the data used to develop the AI system is up-to-date, of high quality, complete, and representative of the system's environment. We prioritize data integrity and ensure that it is appropriately collected, labelled, and processed for accuracy.

**• Did you put in place a series of steps to monitor, and document the AI system’s  
accuracy?**

Yes, we established a series of steps to monitor and document the AI system's accuracy continuously. We developed testing and quality assurance frameworks, regularly monitoring the system's performance, and logging any issues detected.

**• Did you consider whether the AI system's operation can invalidate the data or  
assumptions it was trained on, and how this might lead to adversarial effects?**

Yes, we considered whether the AI system's operation could invalidate the data or assumptions it was trained on and how this might lead to adversarial effects. We developed measures to detect, address and prevent these effects, ensuring the system's accurate performance is maintained.

**• Did you put processes in place to ensure that the level of accuracy of the AI system  
to be expected by end-users and/or subjects is properly communicated?**

Yes, we put in place processes to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is communicated effectively. End-users and subjects received education on best practices in using the system, including the system's expected accuracy levels and performance measurements.

Reliability, Fall-back plans and Reproducibility  
**• Could the AI system cause critical, adversarial, or damaging consequences  
(e.g. pertaining to human safety) in case of low reliability and/or reproducibility?**

in case of low reliability and/or reproducibility? No, we do not believe that our AI system could cause critical, adversarial, or damaging consequences in case of low reliability and/or reproducibility.

**o Did you put in place a well-defined process to monitor if the AI system is  
meeting the intended goals?**

Yes, we have put in place a well-defined process to monitor if the AI system is meeting the intended goals. We regularly monitor the system's performance and assess whether it is meeting the expected outcomes.

**o Did you test whether specific contexts or conditions need to be taken into  
account  
to ensure reproducibility?**

Yes, we have tested whether specific contexts or conditions need to be taken into account to ensure reproducibility. We have evaluated the system's performance under different conditions and made adjustments to ensure that it performs consistently.

**• Did you put in place verification and validation methods and documentation  
(e.g. logging) to evaluate and ensure different aspects of the AI system’s reliability  
and reproducibility?**

Yes, we have put in place verification and validation methods and documentation, such as logging, to evaluate and ensure different aspects of the AI system's reliability and reproducibility. We have established processes to verify and validate the data used in the system, as well as its performance and accuracy.

**o Did you clearly document and operationalise processes for the testing and  
verification of the reliability and reproducibility of the AI system?**

Yes, we have clearly documented and operationalized processes for the testing and verification of the reliability and reproducibility of the AI system. We have established testing frameworks to ensure that the system operates reliably and consistently, and we maintain documentation to track its performance.

**• Did you define tested failsafe fallback plans to address AI system errors of whatever  
origin and put governance procedures in place to trigger them?**

Yes, we have defined tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them. We have established procedures to detect errors and inconsistencies and have developed fallback plans to address them promptly.

**• Did you put in place a proper procedure for handling the cases where the AI system  
yields results with a low confidence score?**

Yes, we have put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score. We have developed guidelines for end-users to interpret and handle low confidence scores appropriately and have established processes to investigate the cause of such results.

**• Is your AI system using (online) continual learning?**

No, our AI system is not using online continual learning.

**o Did you consider potential negative consequences from the AI system  
learning novel or unusual methods to score well on its objective function?**

N/A (not applicable) since our AI system is not using online continual learning.