**IT3389 Applied AI Project**

Ethical AI System Implementation Guidelines

Project Name: POSEIDON AI

Group Number: Team 2

|  |  |
| --- | --- |
| Student Admin Number | Student Name |
| 223715Y | GANGULA KARTHIK |
| 220080D | NG JUN MING |
| 221814H | CHOY WEI JUN |
| 220273H | SEAH PIN SHIEN |

# Introduction

This assessment is based on (1) the Model AI Governance Framework (2nd Edition) and (2) Implementation and Self-Assessment Guide for Organizations (“ISAGO”) published by Info-Communications Development Authority (IMDA) & Personal Data Protection Commission (PDPC).

* **Model AI Governance Framework** [[1]](#footnote-1) is a general, ready-to-use tool to enable organizations that are deploying AI solutions at scale to do so in a responsible manner.
* **ISAGO Implementation Guide** [[2]](#footnote-2) helps organizations assess the alignment of their AI governance practices and processes with the Model Framework.

Only 11 of the 64 guiding questions from ISAGO were used in this assessment questionnaire and it is meant for students doing AI-related projects in their studies.

Three additional columns have been added. The Response column allows the project team to specify whether the questions have been responded to or are not applicable to the project. The details of the response are included under the Remarks column. The Stage column suggests when the questions are most applicable throughout the project life cycle.

**Definition of Stages:**

1. Initiation Stage –  Initial Problem Definition
2. Implementation Stage – AI Model Prototyping & Testing and Prototype / Project Development

# System Profile

|  |  |
| --- | --- |
| **Project Name** | POSEIDON AI |
| **Project Description** | **Project Poseidon AI: Enhancing Digital Safety Through AI-Powered Content Classification**   * The project uses multimodal AI models to detect **hateful memes**, preventing the spread of divisive rhetoric linked to real-world harm. * **Mental health classification** models analyze social media content to identify high-risk posts related to anxiety, depression, and suicidal ideation, enabling early intervention. * **Deepfake detection** using CNN-based classifiers combats financial scams, political manipulation, and identity theft by identifying synthetic media inconsistencies. * **Fake news classification** utilizes deep NLP models trained on verified datasets to counter misinformation that erodes public trust.   By integrating cutting-edge AI for content moderation, Project Poseidon AI strengthens digital resilience, aligns with Singapore’s AI governance leadership, and supports sustainable societal impact. |

# Assessment Questionnaire

|  | **Guiding Question** |  | **Response** | **Stage** | **Remarks** |
| --- | --- | --- | --- | --- | --- |
| Section 1: Objectives of deploying AI | | | | | |
| **1.** | Has your project team defined a clear purpose in using the identified AI solution (e.g. operational efficiency and cost reduction)? |  | **YES**  **NO**  **NA** | Initiation | AI models will automate online content classification, significantly reducing reliance on human annotators, who typically take a few minutes per item. In contrast, AI can process thousands of items per second, enabling nonlinear scalability, where content volume increases without a proportional rise in human effort. This allows platforms to handle millions of posts daily, improving efficiency while maintaining accuracy. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Section 3: Determining the level of human involvement in AI-augmented decision-making | | | | | |
| **2.** | Did your project team implement the appropriate level of human involvement in AI-augmented decision-making? |  | **YES**  **NO**  **NA** | Implementation | For the harmful meme classifier, we have established a human-AI oversight framework designed to ensure the model’s accuracy and reliability while allowing for human involvement when necessary. All model responses, including those from the judge model (OpenAI GPT-4), are logged in a PostgreSQL database for continuous monitoring. Human annotators are tasked with reviewing instances where there is a discrepancy between the predictions made by the primary model and the judge model, which provides an explanation as to why the model may have misclassified the content. This review process allows for greater transparency and accountability, as annotators can assess both the model’s decision and the rationale behind the judge model’s classification.  In scenarios where frequent misclassifications occur, a feedback loop is established, where the logged data can be used by developers to refine the model. The developers have the flexibility to either retire the current model or retrain it using the flagged instances, thereby improving the model's performance over time. This ensures that the system remains adaptive to emerging trends in harmful content, such as new meme formats or language use, and helps to reduce the rate of errors in classification. |
| **3.** | [Relevant only in limited scenarios]:  For safety-critical systems, did your project team ensure that:  A. The relevant personnel will be able to assume control where necessary?  B. The AI solution provides sufficient information to assist the personnel to make an informed decision and take actions accordingly? |  | **YES**  **NO**  **NA** | Implementation | 1. For safety critical systems, such as the harmful meme classifier, If the false negative rate exceeds 5% for three consecutive days, the system automatically reverts to a prior version while retraining, giving personnel control to address the issue promptly. 2. Our AI solution also provides sufficient information to assist personnel in making informed decisions. Human annotators are provided with explainability tools, such as saliency maps that highlight specific areas of a meme that influenced the classification (e.g., altered religious symbols or offensive gestures). Additionally, counterfactual explanations are provided, detailing why a specific piece of content was flagged, such as "This meme was flagged because text 'XYZ' combined with imagery 'ABC'. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Section 4: Operations Management | | | | |
| **4.** | Did your project team adopt best practices in data management and protection in dealing with data owned by the customer / project owner? |  | **YES**  **NO**  **NA** | Initiation | For the harmful meme dataset, there were some images which contained user’s twitter/social media usernames. The images were in-painted to remove this information and the meme captions which did contain this text were masked with “[USERNAME]” token.  The other data used does not contain any personally identifiable information for any of the group use-cases or datasets. |
| **5.** | Is the dataset used relevant? |  | **YES**  **NO**  **NA** | Implementation | The dataset for the meme classification task consists of the meme images, the meme caption/description, and the target label describing whether the meme is harmful or harmless. The inputs and the target column are clearly defined for the scope of problem, making the dataset highly relevant. |
| **6.** | Where personal data is involved, is it collected for the intended purposes? |  | **YES**  **NO**  **NA** | Implementation | As mentioned earlier there is no personal data collected, where there is personal data present, like in social media usernames, it has been masked. |
| **7.** | Is the dataset used well-structured and in a machine understandable form? |  | **YES**  **NO**  **NA** | Implementation | The dataset used consists of images, the meme caption and the binary label. The images can be pre-processed using image processing technique along with the meme caption which will undergo the text preprocessing techniques. Both these modalities will be converted into embeddings using various models such as BERT, RESNET, and CLIP, which will be the input to train the deep learning model. |
| **8.** | [Relevant only in limited scenarios:]  If the dataset used has been joined from multiple datasets, were the extraction, transformation and other relevant operations performed correctly? |  | **YES**  **NO**  **NA** | Implementation | The initial dataset which was used is the Facebook harmful meme dataset from Kaggle. This dataset was imbalanced which would cause the model to be more biased to picking the majority class which is the harmless label. To solve this issue, the harmful memes were webscraped from the internet across various harmful topics such as racism, slavery, gender discrimination, etc.  Before merging the datasets, rigorous data preprocessing steps were performed to ensure consistency in structure and attributes. Both datasets were aligned to maintain the same schema, including image file paths, meme captions, and corresponding labels. The data integration process involved schema validation, duplicate removal, and sanity checks to ensure data integrity. Furthermore, exploratory data analysis (EDA) was conducted to verify the effectiveness of the dataset augmentation, ensuring that the combined dataset was more balanced and representative of real-world harmful content. |
| **9.** | Did your project team identify and propose solutions to mitigate unintended biases in the dataset used for the AI model, especially omission bias and stereotype bias? |  | **YES**  **NO**  **NA** | Implementation | During exploratory data analysis (EDA), we observed a significant bias in the dataset, where race- and religion-related terms appeared disproportionately in the harmful meme category compared to the harmless category. This introduced a risk where even harmless memes containing race- or religion-related words were more likely to be misclassified as harmful, leading to stereotype bias in model predictions.  To address this, we implemented a judge LLM as an independent evaluation layer. This model reviewed the AI classifier’s predictions and provided a corrected classification along with a rationale explaining the decision. By systematically logging and analysing discrepancies between the classifier and the judge LLM, we were able to identify and quantify bias patterns. The corrected data was then incorporated into model retraining using an active learning approach, improving fairness and reducing over-sensitivity to specific linguistic patterns. |
| **10.** | Did your organization use different datasets for training, testing and validation of the AI model? |  | **YES**  **NO**  **NA** | Implementation | A training split of 70% for training, 20% for validation and 10% for testing was used. |
| **11.** | Can your project team explain how the deployed AI model functions and arrives at a particular prediction? |  | **YES**  **NO**  **NA** | Implementation | The project team can effectively explain how the deployed AI model identifies fake news and generates predictions. Specifically, SHAP (SHapley Additive Explanations) was used to analyse the model's decisions by highlighting the most influential words contributing to a classification as fake or real. This approach enhances transparency, allowing us to understand which textual patterns or terms the model considers most indicative of misinformation. |

1. Model AI Governance Framework available at <http://go.gov.sg/AI-gov-MF-2> [↑](#footnote-ref-1)
2. ISAGO available at <http://go.gov.sg/ISAGO> [↑](#footnote-ref-2)