**Designing an AI based Anti-phishing product based on a Large Language Model (LLM)**

**1. Introduction:** In this study, we discuss the design, implementation, evaluation metrics, constraints and ethical issues related to an AI based product that categorizes emails as either spam or ham (non-spam). By automatically classifying incoming emails using machine learning methods, the software improves user experience by assisting in the removal of undesired content.

**2. Technical Explanation:**

Python's pandas and scikit learn libraries are used to create the AI model. In terms of technical implementation, the following are the main steps:

Data Collection and Preprocessing: A dataset containing email messages and their accompanying labels (spam or ham) is loaded using pandas. For consistency, empty strings are used to replace null values.

Label Encoding: Labels are converted to numeric values. Ham is encoded as 1 while spam is encoded as 0.

Data Splitting: The scikit learn train­\_test\_split function is used to split the dataset into training and test sets.

The features from the email text data are extracted and converted into numerical feature vectors using the Term Frequency Inverse Document Frequency vectorization technique. This technique converts text into a format that machine learning algorithms can comprehend.

Preparing the model: It uses a logistic regression model as the classifier. It is learned using the labels and feature vectors from the training set.

Evaluation: The accuracy of the model is evaluated on both the training and test datasets to gauge how well the trained model performs.

The model is used to forecast whether new email messages will be spam or valuable.

**3. Evaluation Metrics:** Accuracy, a standard statistic for classification tasks is used to judge the efficiency of the AI product. The proportion of incidents that were accurately predicted out of all instances is what is meant by accuracy.

**4. Limitations:**

Data bias: The model's performance is strongly influenced by the calibre and variety of the training set of data. The model might not generalise to real world situations well if the training set of data is biased.

The model's TF IDF vectorization which may not be able to fully capture complicated semantic links in the text is used for feature extraction.

Model choice: While logistic regression is a straightforward classifier, more complex models might perform more effectively.

**5. Ethical Points to Think About:**

Private information may be present in the email content that the product deals with. Data security and privacy must be guaranteed.

False Positives/Negatives: Misclassifying emails can have detrimental effects such as incorrectly labelling critical communications as spam or malignant emails as ham.

Transparency: It is important for users to be aware that AI algorithms are processing their emails and it is also important to make the classification criteria clear.

**6. Conclusion:**

A logistic regression model trained on TF IDF features is used by the AI product to accurately categorize emails as spam or ham. However, there are some performance restrictions on the product and moral concerns are crucial for a responsible deployment.

**7. Future improvements:**

Model selection: Experiment with more sophisticated categorization methods like Random Forest, Naive Bayes or neural networks.

Look at more advanced word embedding approaches while developing features for text representation.

Synonyms, paraphrases and/or more sources can be added to the training data to broaden its diversity.

We can develop a more reliable and powerful AI based solution for email classification by taking these factors into account and continuously improving the model and product.