Greetings!

Dear candidate, we are pleased to inform you that your application has been shortlisted for our job posting "**50-2024 Research Software Engineer**" (m/f/d). Before we proceed with the interview round, as part of the selection process, we ask you to please complete the task described below.

Task Details

Screening Task: Semantic NLP Filtering for Deep Learning Papers in Virology/Epidemiology

General Task Information: The aim of this task is to filter and classify academic papers from a dataset created through a keyword-based search on PubMed. The dataset is provided in CSV format and contains 11,450 records. The specific goal is to identify papers that implement deep learning neural network-based solutions in the fields of virology and epidemiology.

Dataset: The dataset can be accessed at Virology AI Papers Repository. It includes a header row and multiple data rows generated from keyword-based searches. The list of keywords used for the searches is available here.

Task Requirements:

1. Implement semantic natural language processing techniques to filter out papers that do not meet the criteria of utilizing deep learning approaches in virology/epidemiology.
2. For the papers deemed relevant, classify them according to the type of method used: ["text mining", "computer vision", "both", "other"].
3. Extract and report the name of the method used for each relevant paper.

Submission Instructions:

- Please upload your code to a Github repository.
- Your README file should clearly document your system explaining to the end user what your solution components are. Consider answering the following questions: which NLP technique for

filtering the papers have you used? Why do you think your approach will be more effective than keywords-based filtering? What are the resulting dataset statistics?

Evaluation Criteria

All solutions will be evaluated on two criteria:

I. Clarity of readme

The README is your initial introduction to potential users of your solution. It should be clear and succinct, effectively presenting your solution's purpose and usage.

II. Simplicity and code cleanness

For this task, avoid complex LLMs and opt for lightweight solutions using smaller language models suitable for personal computers or free platforms like Google Colab. Heuristics-based approaches are also encouraged. Your solution should streamline the early stages of article collection for review, effectively minimizing manual scanning and filtering of numerous articles.