Tab 1

**Social Sustainability in the Professional Journey**

**IT3389 Applied AI Project | Nanyang Polytechnic | Year 3**



| ***Team member’s name*** | Siew Wei Heng | Ethan Lee | Lim Hon Lin Gavin |
| --- | --- | --- | --- |
| ***Admin Number*** | 220342H | 220664H | 224729C |
| ***Roles played*** | Team Leader | Team Member | Team Member |

# Table of Contents

[**Table of Contents 2**](#_rxokkcuyofqh)

[**Problem statement 3**](#_vw82pmlogcws)

[General Problem statement 3](#_m6cr1brny3jd)

[**Research into problem 3**](#_vw82pmlogcws)

[1. Income Inequality across Geography (Wei Heng) 4](#_1pt5r1gfnuc8)

[“Income inequality hinders demand” [1] 4](#_7devfyodi42l)

[“Income inequality increases social tensions” [1] 4](#_mbpx8v1t6uen)

[“Income inequality is the 10th goal of the SDGs” [2] 5](#_miho6yvgxr1c)

[Specific Problem statement 5](#_33zo9tovxtu4)

[Problem Description 5](#_dsawnwiwj9in)

[Linking back to the Problem Statement 5](#_yqjze7xs7w1e)

[2. Inequality in Educational Opportunities Geographically (Gavin) 6](#_3ooo8ojp941u)

[Problem Description: 6](#_p0bnpul35mbq)

[Insight/Analysis: 6](#_1k9ry1ca87qw)

[Linking back to the Problem Statement: 6](#_sizbeo6kszrj)

[3. Lack of Access to Personalised Guidance (Ethan) 7](#_l0e04s7lp9sf)

[Problem Description: 7](#_682b6u677r1)

[Research: Enhancing Engagement for Personalised Guidance 7](#_xcaj3zebc3qo)

[Link to General Problem Statement: Promoting Equal Opportunities through Optimised Engagement 7](#_gur6nmc4vsl7)

[**Proposed Solution 8**](#_vw82pmlogcws)

[Overall Web App 8](#_9y1gzc4iljg9)

[1. Income Inequality across Geography (Wei Heng) 8](#_5l4st46suu8c)

[Solution description 8](#_dglxavfrr90v)

[Benefits of solution 9](#_rmv5i1todr4s)

[How it fulfils Social Sustainability 9](#_x02ba2qtn6le)

[Expected Input & Output Table for Web App 9](#_hexjix6glvff)

[2. Inequality in Educational Opportunities Geographically (Gavin) 10](#_e3q2g8azmggs)

[Solution description 10](#_1qaz3btrduod)

[Benefits of Solution 10](#_l7r6iub3ek3g)

[How it fulfils Social Sustainability 11](#_w1hcm56wu2ho)

[Expected Input & Output Table for Web App 11](#_3fxonv4njhl1)

[3. Lack of Access to Personalised Guidance (Ethan) 12](#_ljytaiit79tt)

[Solution description 12](#_k1p1zjmjkbav)

[Benefits of solution 13](#_7u36po8g8no5)

[How it fulfils Social Sustainability 13](#_q48cvr5qydil)

[Expected Input & Output Table for Web App 13](#_dv7x2zn9vdq0)

[**Scope of Work (Components) 13**](#_vw82pmlogcws)

[1. Income Inequality across Geography (Wei Heng) 13](#_v3j3rplxr2g2)

[Scraping 13](#_h80ipeozah3u)

[EDA 14](#_4131vx3ejnak)

[Model 16](#_mxub32z3j4ow)

[Expected Input & Output Table for ML 16](#_cxbpmjp86xr)

[Web App 17](#_f9myjogxow2h)

[2. Inequality in Educational Opportunities Geographically (Gavin) 17](#_f39fn1rnf7x9)

[Data Source 🌐 17](#_gd8b8n8qpsro)

[EDA 18](#_4qno21f1x15f)

[Model 25](#_h814ufw2afzc)

[Web App 28](#_cie1a626n0yv)

[3. Lack of Access to Personalised Guidance (Ethan) 29](#_tu2ophz0ck1d)

[Scraping 29](#_ehhmxin4vf7m)

[EDA 29](#_pbim32iex4on)

[Model 31](#_1sgebuttc90q)

[Web App 32](#_mbx4ufv5ho9z)

[**Milestones & Deliverables 33**](#_vw82pmlogcws)

[Overall Gantt chart 33](#_jfmxsf4558z8)

[Conclusion 34](#_o7olb9z4cxgr)

[**Reference links (for any research done) 34**](#_vw82pmlogcws)

[Income Inequality across geography (Wei Heng) 34](#_5iibnugtxk45)

[**Global Income Inequality (Wei Heng) 35**](#_2715es3vaxf5)

# Problem statement

## General Problem statement

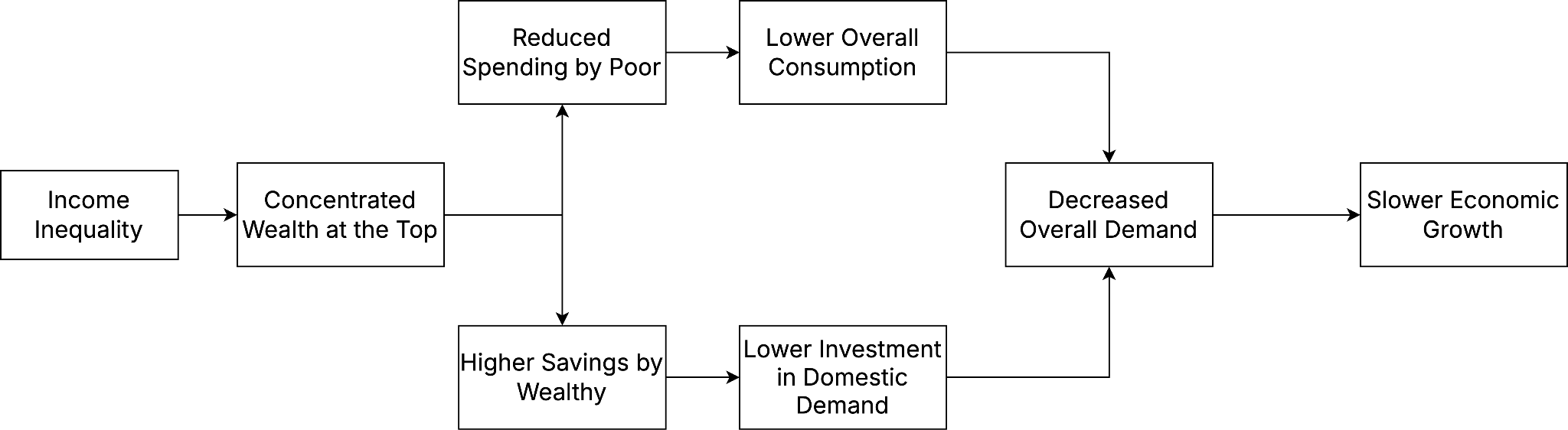
"Many working professionals and students face unequal opportunities in their educational and professional journeys due to factors beyond their control, hindering their ability to thrive and achieve their full potential."

# Research into problem

To support the problem statement, we have identified and researched the 3 following problems with social sustainability that have high relevance, and the potential to have insights mined by AI.

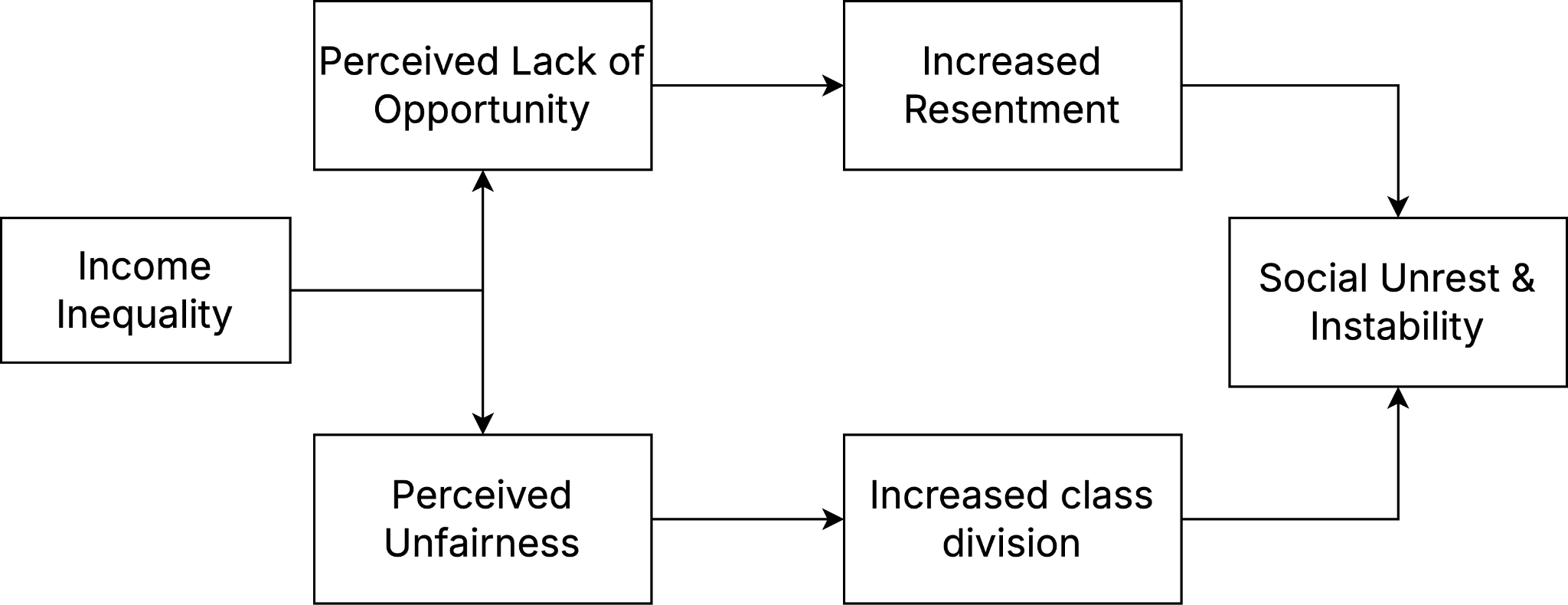
## 1. Income Inequality across Geography (Wei Heng)

### “Income inequality hinders demand” [1]

****

The unequal income distribution inside and across nations has become a policy conundrum for national and international bodies. Increasing income inequality in an economy leads to a low proportion of labor income to national income, generating insufficient overall demand.

### “Income inequality increases social tensions” [1]



One significant policy-level problem stemming from income inequality is the aggravation of social disturbance and instability. Income disparity at high levels frequently causes discontent within underprivileged communities, hence exacerbating social tensions and creating political instability.

Inequality also reduces the effectiveness of poverty-reduction initiatives, limiting progress toward a more fair and sustainable society.

### “Income inequality is the 10th goal of the SDGs” [2]



Reducing Income Inequality across and among countries is the 10th goal of the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership. Adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future.

### Specific Problem statement

“While people everywhere aspire to better careers and fair compensation, especially when considering opportunities abroad, getting reliable and personalized information to make those decisions is incredibly difficult. This can lead to missed opportunities or misinformed moves.”

### Problem Description

Talented individuals, especially those from less developed countries, may not have the same access to information and opportunities as those in more developed nations. This lack of access can create a misinformed idea that moving abroad is always the best option, without fully understanding the realities of the job market and fair compensation in their destination country. This can lead to individuals making mistakes that didn’t benefit their careers or financial well-being as expected.

### Linking back to the Problem Statement

Our problem statement states that many individual’s potentials are hindered, and part of this is due to the fact that working professionals from less developed countries may not have equal opportunities to their career growth compared to developed countries. They may also develop a misinformed bias that moving to other countries is the right move without in-depth research. This will result in mistakes and can set people back in achieving their highest potential.

## 2. Inequality in Educational Opportunities Geographically (Gavin)

### Problem Description:

Educational disparities hinder individuals from reaching their full potential, perpetuating cycles of inequality and limiting societal progress. Factors like economic constraints, inadequate government spending, and unequal access to quality resources create barriers, particularly for underprivileged communities. Addressing these disparities is crucial to fostering social sustainability, inclusivity, and equitable development globally.

### Insight/Analysis:

**Access to Education**:

* Around 258 million children, adolescents, and youth globally were out of school in Year 2018 (UNESCO). | [Link](https://shorturl.at/GJxHf)
* Key barriers include economic constraints, gender inequality, lack of infrastructure, and socio-political instability.
* Rural areas, particularly in low-income countries, experience significant gaps in primary and secondary school enrollment rates compared to urban regions.

**Quality of Education**:

* Even in regions with access to schools, the quality of education varies widely.
* High pupil-teacher ratios, underqualified teachers, and outdated curricula diminish the effectiveness of learning.
* For instance, in Sub-Saharan Africa, the average pupil-teacher ratio in primary schools exceeds 40:1. | [Link](https://shorturl.at/oUemb)

**Government Expenditure**:

* Government investment in education significantly impacts access and quality.
* Countries with higher GDP allocation to education often see better literacy rates, enrollment numbers, and professional outcomes.
* For example, Nordic countries consistently allocate over 6% of GDP to education, resulting in equitable opportunities and high societal well-being. | [Link](https://shorturl.at/JMAKV)

### Linking back to the Problem Statement:

Unequal access to education limits individuals' potential, as stated in the problem: "**Many face unequal opportunities in their educational and professional journeys, hindering their ability to thrive**." Our approach addresses disparities in enrollment, resources, and quality, promoting inclusivity and contributing to social sustainability. By improving educational equity, we support societal well-being and global progress.

## 3. Lack of Access to Personalised Guidance (Ethan)

### Problem Description:

Many individuals face challenges in accessing advice that fits their specific context. For students choosing their educational paths and professionals considering career moves, the inability to filter through vast amounts of content can lead to suboptimal decisions. This is especially critical in situations where reliable information can dramatically influence their future. This leads to creating posts on these forums, which usually go unnoticed or gather too little attention to get the best advice.

### Research: Enhancing Engagement for Personalised Guidance

The challenge of accessing context-specific guidance is a global issue, particularly for students and professionals who are navigating important decisions about their education and careers. Many of these individuals turn to online forums for advice, yet the posts they create often go unnoticed or fail to attract enough attention to yield meaningful feedback. The post predictor addresses this issue by forecasting the engagement level of posts, helping users ensure that their inquiries are more likely to receive the attention and personalized advice they need. This approach helps promote social sustainability by optimising the visibility of valuable, relevant posts, ensuring a more equitable exchange of knowledge.

### Link to General Problem Statement: Promoting Equal Opportunities through Optimised Engagement

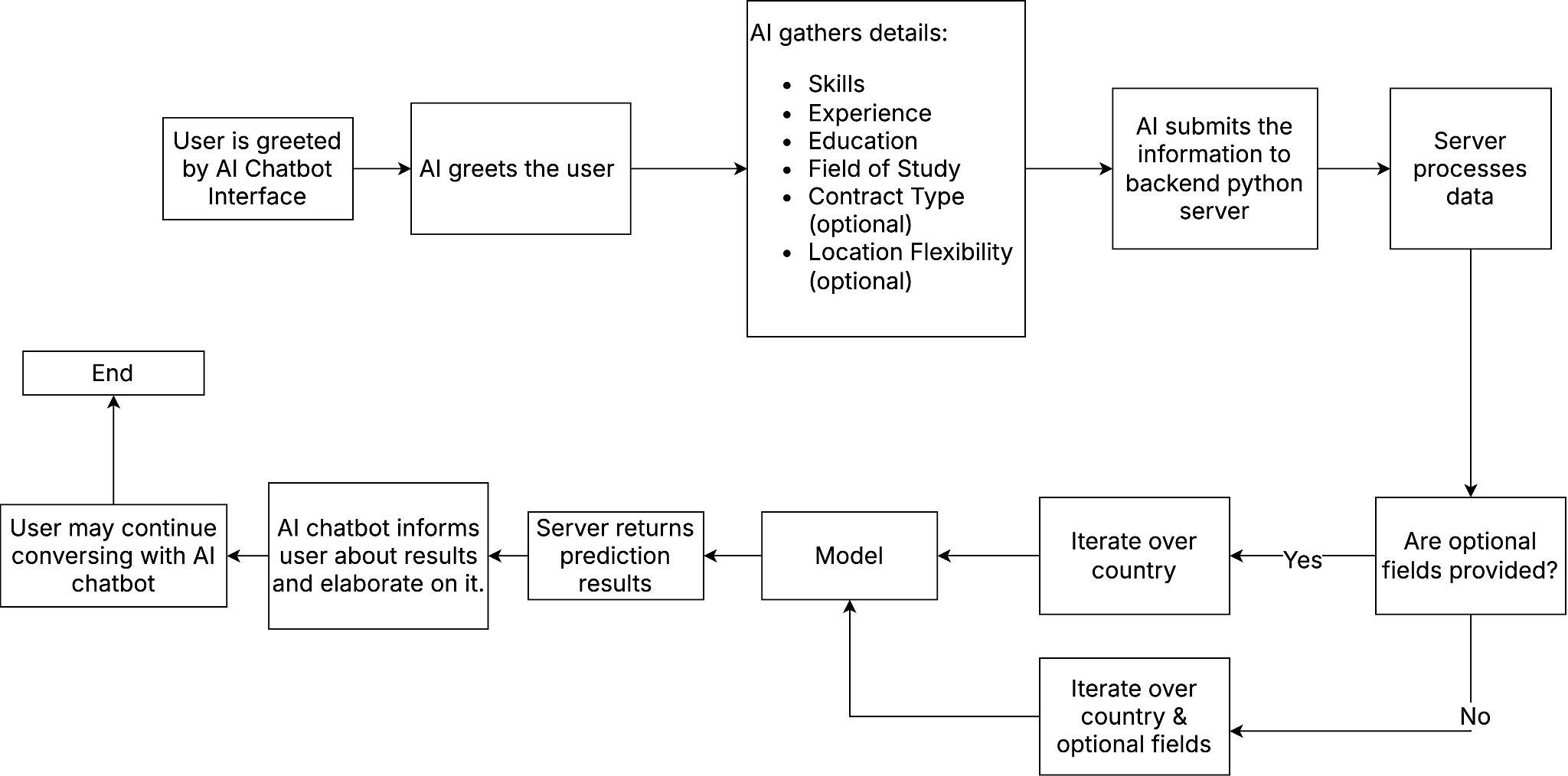
The post predictor directly addresses the general problem of unequal opportunities by ensuring that users, regardless of their background, can gain equal access to relevant and personalised guidance. By predicting which posts will garner the most engagement, it increases the likelihood that users receive the feedback they need, empowering them to make informed decisions in their educational and professional paths. This enhances social sustainability by reducing barriers to useful information and promoting fairer access to career and educational opportunities.

# Proposed Solution

## Overall Web App

## 1. Income Inequality across Geography (**Wei Heng**)

### Solution description



We are developing an AI-powered chatbot that provides users with personalized salary predictions and insights into the cost of living in the US, Singapore, and India. The chatbot will guide users through a conversational interface, gathering information about their skills, experience, and career aspirations. It will then leverage machine learning models, trained on a rich dataset of job postings and economic data, to deliver:

* Predicted Salary (USD): Estimated salary for the user's profile in each of the three target countries.
* After-Tax Salary: Estimated salary after accounting for local income tax rates.
* Cost of Living Considerations: Contextual information about the cost of living in each location, helping users understand the real value of their potential earnings.
* General Inquiries: General information regarding this project, and what the project aims to solve.

### Benefits of solution

**Empowers Informed Decision-Making**

Provides individuals with the data they need to make informed decisions about international career moves.

**Promotes Fair Compensation**

Increases salary transparency and helps individuals understand their worth in different markets.

**Mitigates Risk of Misinformed Choices**

Reduces the likelihood of individuals making career moves based on inaccurate or incomplete information.

**Addresses Information Asymmetry**

Helps bridge the information gap between individuals in less developed countries and opportunities in more developed economies.

### How it fulfils Social Sustainability

Fair compensation globally is an aspect of social sustainability, because working professionals and students can now, with this web application, consider the relative value of their skills in different countries. This can contribute to global economic development, education of fair salary ranges, and improved salary negotiation.

The main goal is to allow working professionals to understand the relative value of their skills in different countries, which can be useful for those who either are considering international career moves, or need empowerment to negotiate for better wages and working conditions in their own countries. The goal is NOT to convince working professionals to move to a different country. The goal is to provide working professionals a global view of the value of their skills, talents and experience, whilst taking into account cost of living, taxes etc.

### Expected Input & Output Table for Web App

| Input | Output |
| --- | --- |
| **AI Inferred from User conversation:**  Soft skills (Multi-value Categorical)  Hard skills (Multi-value Categorical)  Location flexibility (Categorical) (Optional)  Contract Type (Categorical) (Optional)  Educational Level (Ordinal)  Field of Study (Categorical)  Min Years of Work experience (Numerical) | Predicted Salary (USD) from Singapore  Predicted Salary (USD) from Singapore (After tax)  Predicted Salary (USD) from USA  Predicted Salary (USD) from USA (After tax)  Predicted Salary (USD) from India  Predicted Salary (USD) from India (After tax) |

## 2. Inequality in Educational Opportunities Geographically (**Gavin**)

### Solution description

Our solution provides a **web-based platform** that uses **machine learning** to predict which **countries need intervention in their education systems**. By analyzing key metrics such as government expenditure, primary completion rates, and school enrollment figures, the platform flags countries where educational disparities are most evident. Through an easy-to-use interface, users can input country data, select relevant education indicators, and receive predictions along with actionable recommendations.

Our platform also includes **visualizations**—such as scatter plots, trend lines, and heatmaps—that help users understand relationships between different education metrics and identify areas that require urgent attention.

### Benefits of Solution

1. Educational Research:Individuals can learn how different factors, such as education spending, impact literacy and enrollment, leading to better understanding and access to educational opportunities.

2. Predictive Modeling:By predicting educational outcomes, individuals can identify potential challenges early, allowing targeted interventions to improve completion rates and overall success.

3. Global Education Analysis:Individuals can gain insights into how global education systems operate, helping them understand resource distribution and how it impacts access to quality education in different regions.

4. Policy Development:This solution helps policymakers create data-driven reforms, ensuring educational policies promote equal opportunities, which ultimately benefits individuals by improving their access to education.

### How it fulfils Social Sustainability

1. Equal Education Access:  
Identifies and addresses educational gaps, ensuring all individuals have equal opportunities to thrive.

2. Economic and Social Growth:  
Empowers individuals through education, contributing to both personal and societal advancement.

3. Informs Policy for Equity:  
Data-driven insights guide policies that reduce educational inequalities and promote fairness.

4. Global Collaboration:  
Analyzes global trends, fostering international cooperation for a more sustainable, equitable future.

### Expected Input & Output Table for Web App

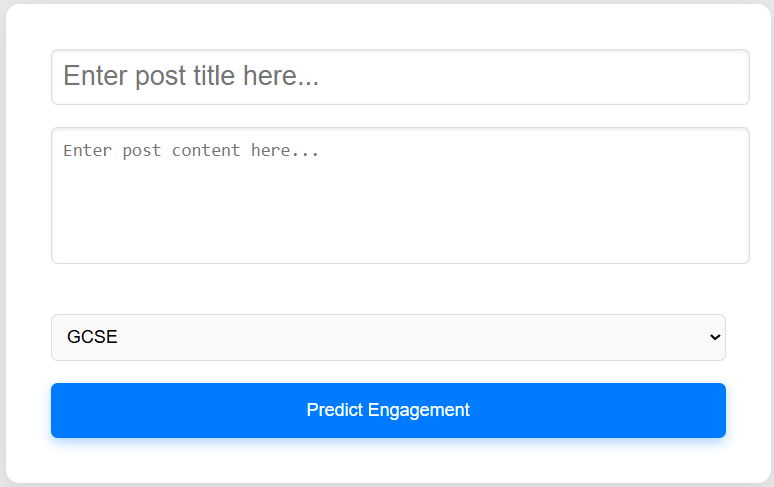
| Input | Output |
| --- | --- |
| Country  Country Code  Year  Government Expenditure (% of GDP)  Primary Completion Rate (%) | School Enrollment Rates:   * Primary (%). * Secondary (%). * Tertiary (%). |

## 3. Lack of Access to Personalised Guidance (**Ethan**)

### Solution description



The goal is to do Post Engagement Prediction on Career and Education related forums. Users will be able to create a post with a similar interface to Figure 3.1, and receive a prediction of the engagement of their post. Moreover, Gen AI will be connected to the post content and title, acting as countermeasures to PII leaks and nonsensical content.



*Figure 3.1*

### Benefits of solution

* Improves Post Quality
* Increase odds of attracting forum users to provide personalised guidance
* Enhances the overall quality of forum discussions.
* Uses Gen AI to filter out PII and nonsensical content, maintaining safety and trust.

### How it fulfils Social Sustainability

High-quality posts receive more engagement, which in turn leads to increasingly tailored advice for individual needs, while also providing quality content for other users of the forum, generating a cycle of sustainability for students and professionals alike.

### Expected Input & Output Table for Web App

| **Input** | **Output** |
| --- | --- |
| Post Title  Post Content Post Category (Either A Levels, GCSE, Study support or Job Experience) | Post Engagement (Categorical)  GenAI Feedback (If countermeasures triggered) |

**Why this works**: Users will be able to get feedback on their post engagement before submission, wasting precious time compared to if they were to blindly post their problems on the internet and pray for good engagement. Moreover, the proposed countermeasures not only protect us from unnecessary calls to the predictive model, but also protect the user from accidentally leaking private details onto the internet.

# **Scope of Work (Components)**

## 1. Income Inequality across Geography (**Wei Heng**)

### Scraping

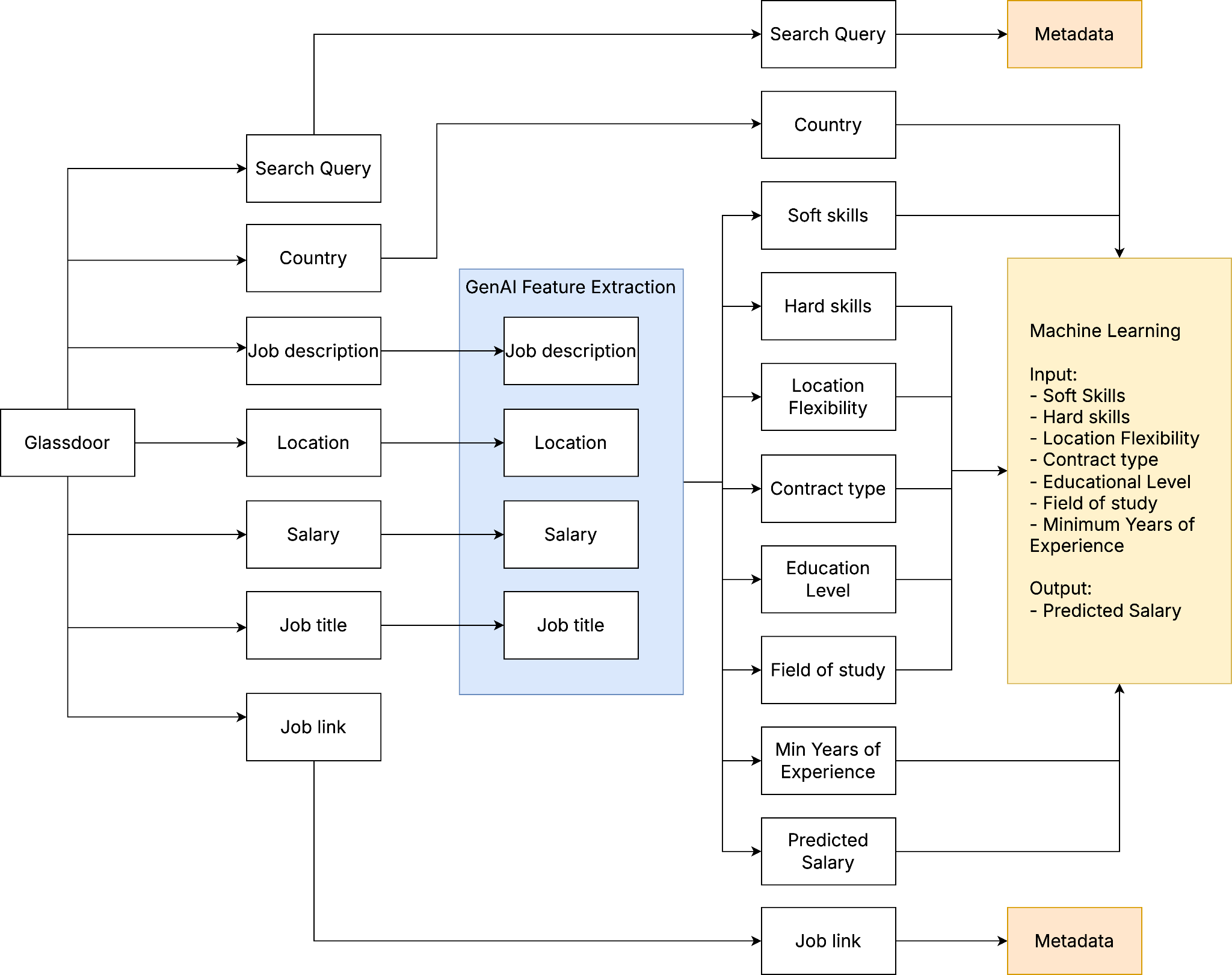
**Expected Data structure of scraping**

Search term (Categorical)

Job Title (Categorical)

Job Description

Salary Range



We will be mostly scraping Glassdoor, since it provides plenty of data from the US, SG and India.

The three countries we will focus on are: Singapore, US and India, as this provides a good contrast over developed and developing economies. There are likely interesting insights here.

### EDA

**Feature Extraction with GenAI**

Not only will we remove missing fields and do the necessary data preparations, we will enrich the data by converting the job descriptions into categorical fields like years of experience, soft skills and hard skills using Generative AI.

Some considerations between model providers are as such: cost > speed > intelligence

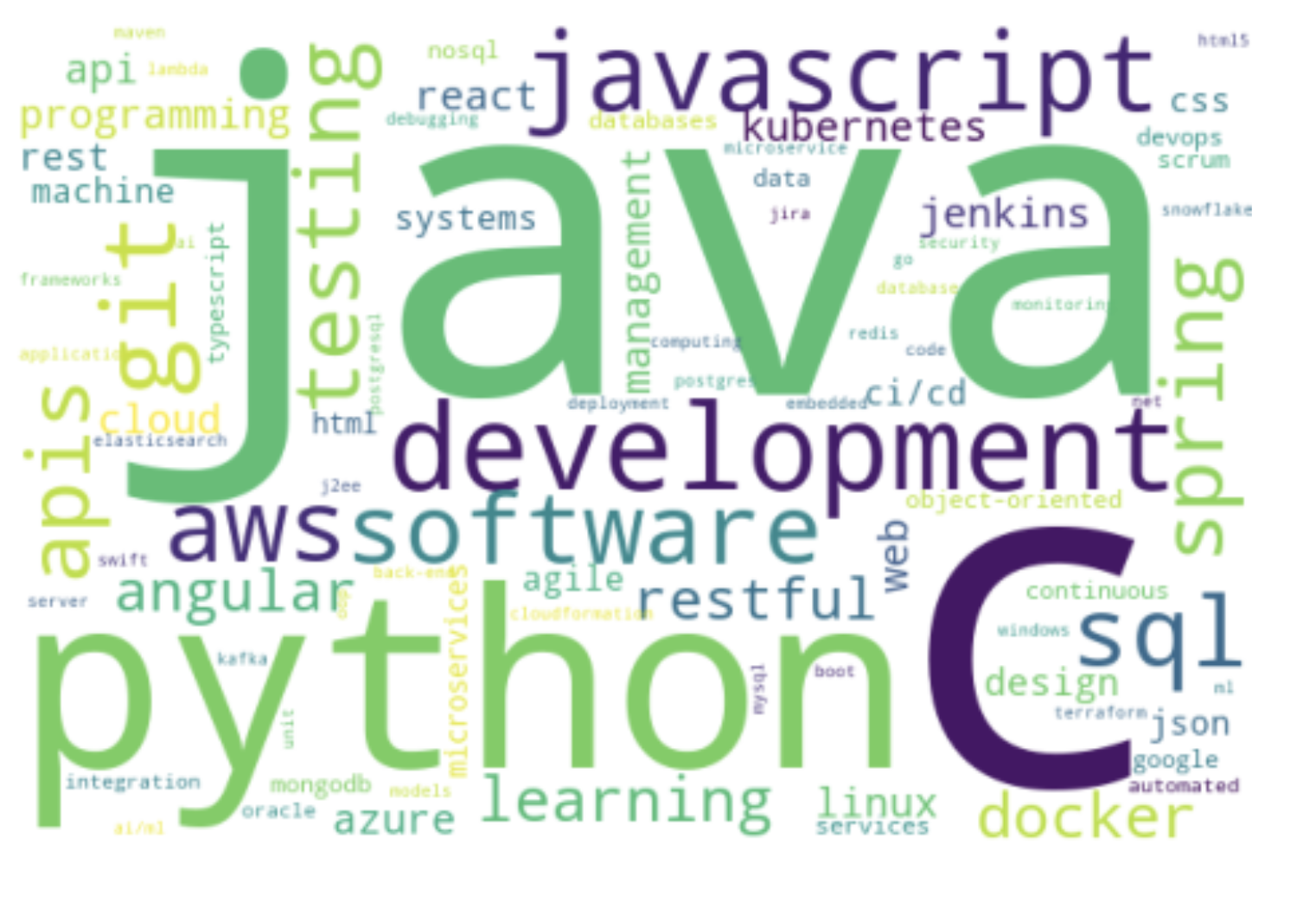
As such, I have settled with the Gemini models as they are free albeit with rate limit, and have decent intelligence to extract features.

**Word clouds**

Soft\_skills:



Hard\_skills:



The GenAI model is able to accurately identify correct soft and hard skills.

### Model

Machine Learning:

1. Try to train a classical model on a 1k job listing and see if any changes are needed for the data.
2. Apply changes and train a deep learning model on 50k job listings.

### Expected Input & Output Table for ML

| Input | Output |
| --- | --- |
| Search term (Categorical)  Job Title (Categorical)  Country (Categorical)  AI Inferred from Job Description:  Soft skills (Multi-value Categorical)  Hard skills (Multi-value Categorical)  Location flexibility (Categorical)  Contract Type (Categorical)  Educational Level (Ordinal)  Field of Study (Categorical)  Min Years of Work experience (Numerical) | Predicted Salary (USD) |

### Web App

1. Landing page
2. Global personalized salary tab

## 2. Inequality in Educational Opportunities Geographically (**Gavin**)

### Data Source **🌐**

**World Bank**: This dataset is compiled from the World Bank's educational database, providing reliable, updated statistics on educational progress worldwide.

**Source Link:** <https://shorturl.at/v4L5g>



The above depicts the data-source file in .csv format, showcasing the initial **different rows** and **columns** on education across the world.

### EDA

Dataset Inspection:

* Loaded the dataset and examined its structure, including columns and missing values.

Data Cleaning:

* Identified and dropped the top three columns with the most missing values.
* Removed rows with any remaining missing data, resulting in a cleaned dataset.

Normalization:

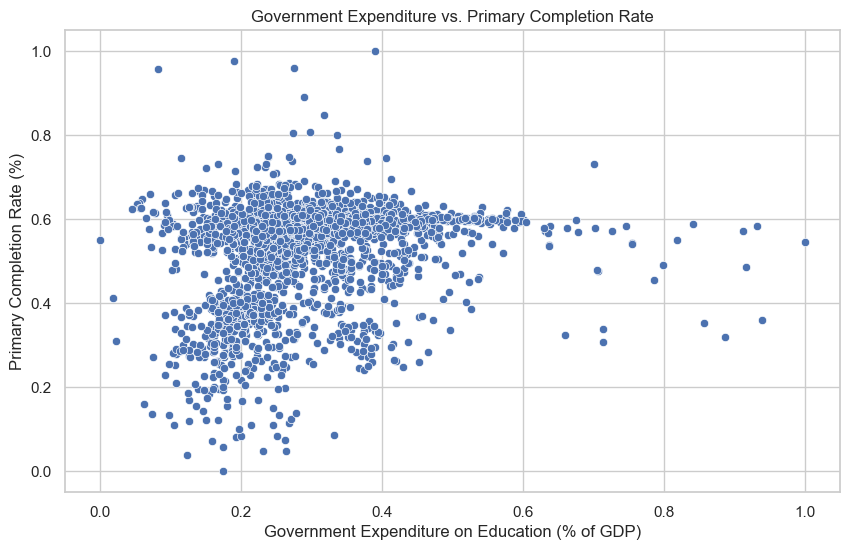
* Normalized numeric columns (excluding the *year* column) using Min-Max Scaling to bring values into a 0–1 range. (For Simplicity.)

Feature Encoding:

* Applied Label Encoding to categorical variables, while **retaining the original *country* names** for interpretability.

Data Visualization:

* A **scatter plot** was used to analyze the relationship between government spending on education (% of GDP) and primary completion rates (%).



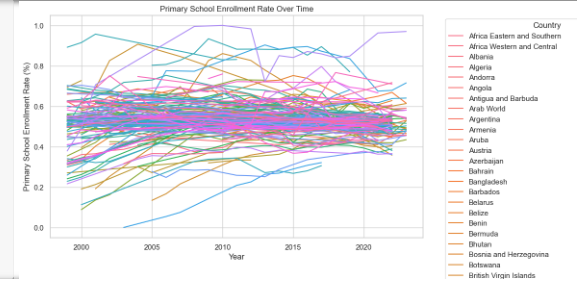
**Insights (Distribution Analysis)**

1. Positive Correlation: There appears to be a positive correlation between government's expenditure on education as a percentage of GDP and the primary completion rate. This suggests that as government's spending on education increases, the proportion of students completing primary education tends to increase as well.

2. Scatter: The data points are scattered, indicating that the relationship is not perfectly linear. This means that while there's a general trend of higher primary completion rates with increased spending, there are many instances where countries with similar spending levels have vastly different completion rates.

\* The scatter plot **highlights** how unequal government's spending on education can create unequal educational outcomes, particularly in terms of primary completion rates. This disparity in access to resources and quality education contributes to the problem of unequal opportunities faced by many individuals.

**Solution**: Addressing this issue requires ensuring equitable distribution of resources and adequate funding for education in all countries.

* A **line plot** visualized changes in primary school enrollment rates across countries over the years.

**Insights (Trend Analysis):**

The line graph depicting primary school enrollment rates across different countries is seen to have a general upward trend which is a positive sign, indicating progress in providing access to primary education around the world. Additionally, it also directly connects to the problem statement by visualizing one key aspect of educational inequality: unequal access to education at the foundational level.

1. **Differing Progress:**

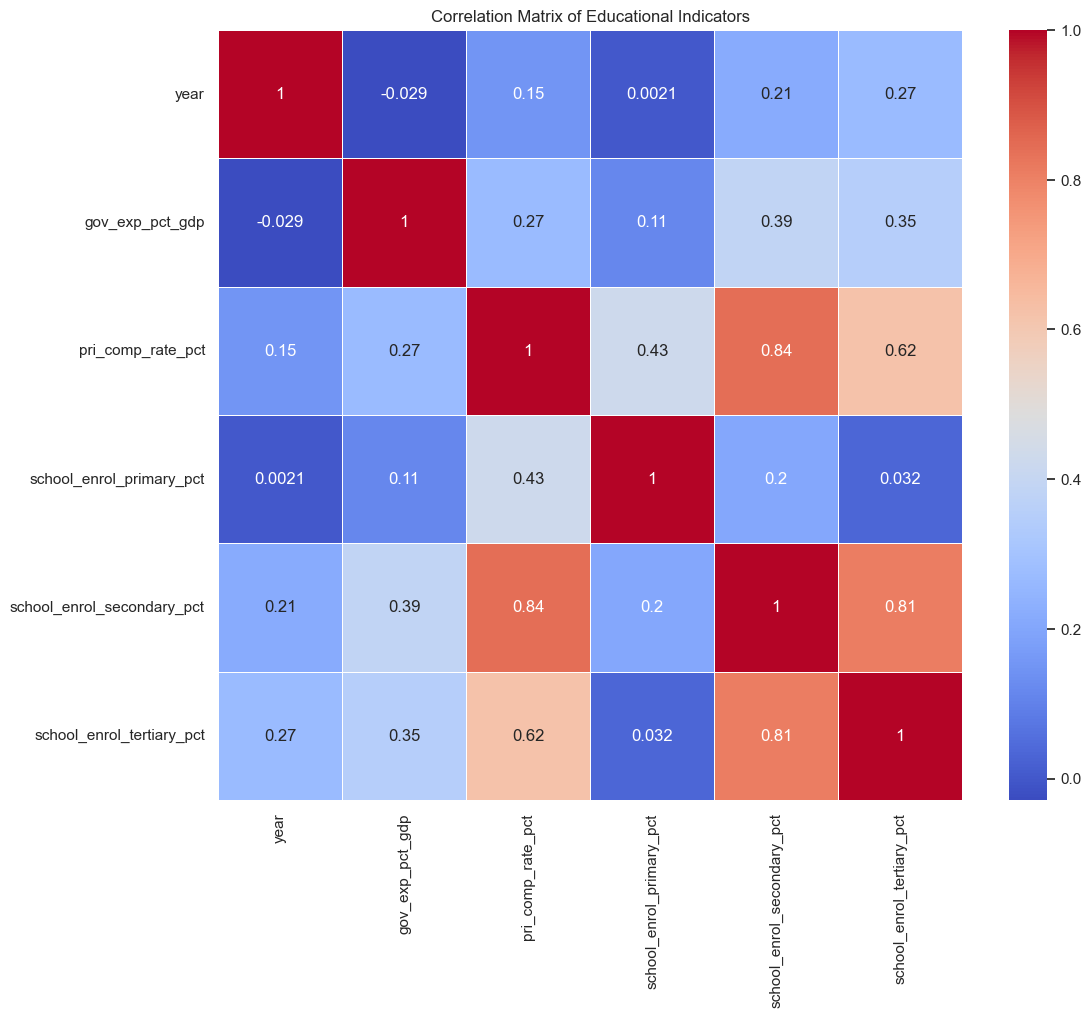
The diverse trajectories of enrollment rates across countries highlight the unequal progress being made in expanding access to primary education. Some countries are making significant strides, while others are struggling to improve. This disparity in progress further contributes to educational inequalities.

1. **Differing Start:**

The above graph illustrates how countries start with varying enrollment rates. This emphasizes that children in some countries have better access to primary education from the very beginning, while others face significant barriers. This disparity in access creates an uneven playing field, impacting future educational and professional opportunities.

\* The line graph demonstrates how unequal access to primary education, as evident in the varying enrollment rates across countries, is a **significant contributor to the problem of unequal educational opportunities.**

* A **heatmap** illustrated the relationships among numeric educational indicators, highlighting strong and weak correlations.



**Insights (Heatmap of Correlations):**

**1. Strong Relationship between Primary Completion and Enrollment:** 😊

As depicted above, the strong positive correlation between primary completion rates and enrollment at all levels (primary, secondary, and tertiary) is a significant finding. This highlights the importance of ensuring high completion rates at the primary level and that countries that successfully educate children at the primary level are more likely to have higher enrollment rates at subsequent educational stages. Low completion rates can create a barrier to further education, exacerbating existing inequalities.

**2. Potential Impact of Government Expenditure:** 😊

While the correlation between government expenditure and primary completion rates is moderate, it indicates a potential positive influence. This suggests that increased government investment in education could contribute to higher completion rates and, consequently, higher enrollment rates.

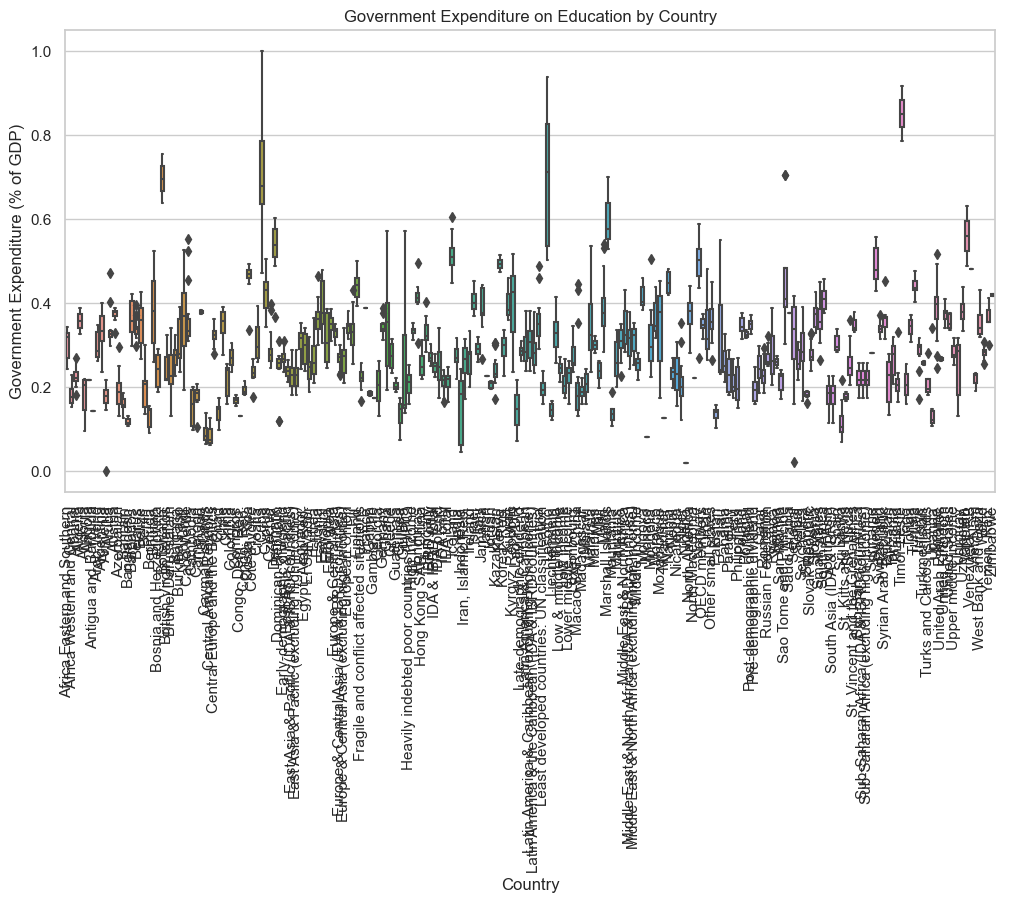
**3. Limited Influence of Country-Level Factors:** 🙁

The weak correlations between "*country*" and "*country\_code*" with most other variables suggest that country-specific characteristics (e.g., culture, historical context) may not have a strong direct impact on these educational indicators.

**4. Potential Time-Series Effects:** 😐

The mixed correlations with "year" suggest that the relationships between these variables might change over time. For example, the impact of government expenditure on education might vary depending on the economic context or policy changes over time.

* A **box plot** showed the variation in government spending on education (% of GDP) across different countries.



**Insights (Box-Plot)**:

The above **box-plot** visualizes the distribution of government's spending on education across various countries. This insight presents to us how much each country allocates to Education as a percentage of its Gross Domestic Product (GDP).

**Evidently**, the chart reveals a wide range of spending with some countries investing significantly higher/lower in education than others, suggesting diverse priorities and economic contexts.

The above visualization **may not** be **clear**,

however, some insights gathered were:

**Highest Government Expenditure**:

- Sweden (47.32%)

- Denmark (49.19%)

- Finland (54.05%)

- France (58.47%)

**Lowest Government Expenditure**:

- Haiti (10.17%)

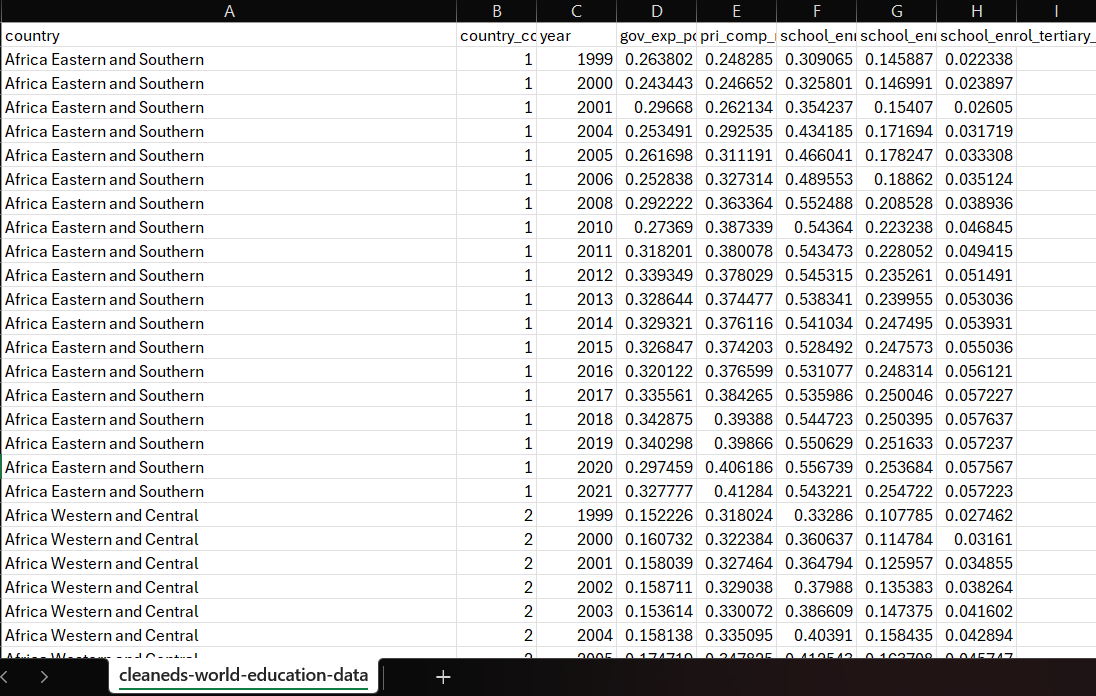
- Indonesia (17.54%)

- Guatemala (14.57%)

- Macao (25.00%)

**Summary Review**:

Countries with the highest government expenditure on education as a percentage of GDP, such as Sweden, Denmark, Finland, and France, typically allocate a significantly larger portion of their economic resources to education compared to countries with lower spending like Haiti, Indonesia, Guatemala, and Macao. This suggests that these high-spending countries prioritize education as a key driver of social and economic development so as to be better positioned to provide quality education for all, while those with lower spending may struggle to ensure equitable access and quality education for all.



The above depicts the **finalised** data-source file in .csv format, after **Data Preparation and EDA Amendments** have been appropriately made.

### Model

**1. What Will My Model Learning (ML) Be Like?**

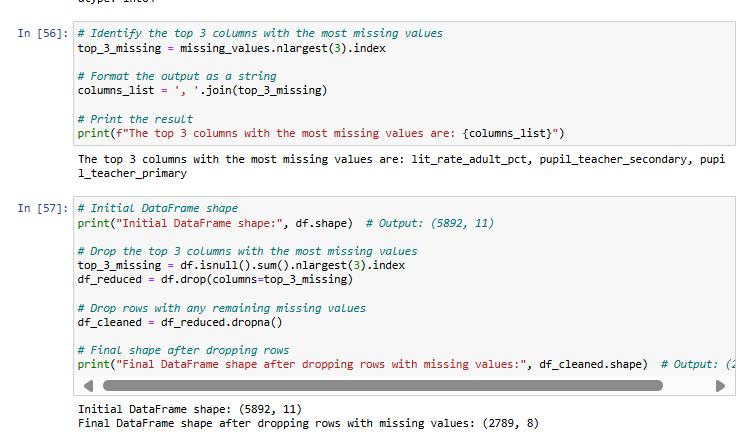
**Model Type:**

* **Random Forest Classifier**:
  + Predicts whether a country requires intervention based on education metrics. (*gov.exp, pri comp rate etc*)
  + Chosen for its ability to handle nonlinear relationships and rank feature importance effectively.

**Input Features:**

After removing the **three columns** with the **most missing values** (*lit\_rate\_adult\_pct, pupil\_teacher\_secondary*, and *pupil\_teacher\_primary*), the remaining features are:

* **Country**:
* **Year**:
* **Government expenditure on education (% of GDP)** (*gov\_exp\_pct\_gdp*):
* **Primary completion rate (%)** (*pri\_comp\_rate\_pct*):
* **School enrollment rates**:
  + **Primary** (*school\_enrol\_primary\_pct*):
  + **Secondary** (*school\_enrol\_secondary\_pct*):
  + **Tertiary** (*school\_enrol\_tertiary\_pct*):



**Target Variable:**

* Binary classification: **"Intervention Needed: Yes/No"**
  + A country is flagged for intervention if certain thresholds for education gaps are not met, such as:
    - Low primary or secondary enrollment rates.
    - Insufficient government expenditure.

**Training Strategy:**

* **Data split**: 70% training, 30% testing.
* **Evaluation metrics**:
  + Accuracy, precision, recall, F1-score, and ROC-AUC to assess classification performance.

**2. Interesting Things About My ML**

Feature Selection:

* Retains quantifiable metrics such as government expenditure and enrollment rates for simplicity and relevance.
* Avoids engineered or composite features like an "Education Opportunity Index" (EOI).

Model Interpretability:

* Random Forest generates **feature importance scores**, allowing clear insights into which inputs most influence predictions.

Scalability:

* The model can adapt to include new education metrics or updated country-level data with minimal adjustment.

**3. Results and Insights About My ML**

Key Results:

* **Countries flagged for intervention**: Those with low enrollment rates or insufficient government investment are identified.
* **Feature importance analysis**:
  + Government expenditure and primary completion rates are likely to emerge as the strongest predictors.

Insights Gained:

* **Visualization results**:
  + Scatter plots reveal diminishing returns of government expenditure beyond a certain threshold.
  + Trends over time show enrollment growth in certain regions, stagnation in others.
* **Policy implications**:
  + Insights guide targeted interventions, such as increasing funding or addressing specific regional disparities.
* **Model performance evaluation**:
  + High recall ensures true positives (countries needing intervention) are correctly flagged.
  + Feature importance highlights actionable predictors, aiding policymakers.

### Web App

**Core Features:**

* Input Section:
  + Users input the following metrics:
    - Country (dropdown menu).
    - Year (numeric input).
    - Government expenditure (% of GDP).
    - Completion and enrollment rates (primary, secondary, tertiary).
* Prediction Output:
  + Binary classification: **"Intervention Needed: Yes/No"**
  + Displays prediction confidence level.
* Visualization Panel:
  + **Scatter plots**: Highlight key relationships (e.g., expenditure vs. primary enrollment).
  + **Line graphs**: Show trends over time for selected metrics.
  + **Heatmaps**: Visualize correlations between input features.
* Recommendation Panel:
  + Suggested interventions for flagged countries (e.g., increase expenditure, recruit teachers).
* Report Generation:
  + Option to download insights as PDF/CSV for further analysis.

**User Journey:**

1. Landing Page:
   * Overview of app purpose and a global dashboard of key metrics.
2. Prediction Tool:
   * Users input data to receive predictions and insights.
3. Insights Section:
   * Explore trends, correlations, and feature importance.
4. Take Action:
   * Receive tailored recommendations and global comparisons

## 3. Lack of Access to Personalised Guidance (**Ethan**)

### Scraping

I will be scraping forum titles, post titles, post content replies & likes and post timestamp. After data collection, I will perform data exploration and investigate potentially sparse features like post likes and do the necessary transformations.

**Data Scraping Target:** Obtain a minimum of **50k** unique posts from career threads and educational threads, while maintaining a balanced dataset.

Websites to Scrape from:

1. [A-levels - The Student Room](https://www.thestudentroom.co.uk/forumdisplay.php?f=80) (Total 235k posts)
2. [GCSEs - The Student Room](https://www.thestudentroom.co.uk/forumdisplay.php?f=85) (Total 130k posts)
3. [Studying, revision and exam support - The Student Room](https://www.thestudentroom.co.uk/forumdisplay.php?f=635) (Total 119k posts)
4. All sub-forums in this: [Careers and Jobs - The Student Room](https://www.thestudentroom.co.uk/forumdisplay.php?f=201) (Total 134k posts)

Due to memory concerns, obtaining data for half a million posts could be unfeasible, so there may be a need to aggressively filter out older posts as they may be less relevant by modern standards. The remaining data should still provide relevant data while striking a balance between educational and career oriented posts during training.

I will also be utilising asynchronous requests to combat time constraints.

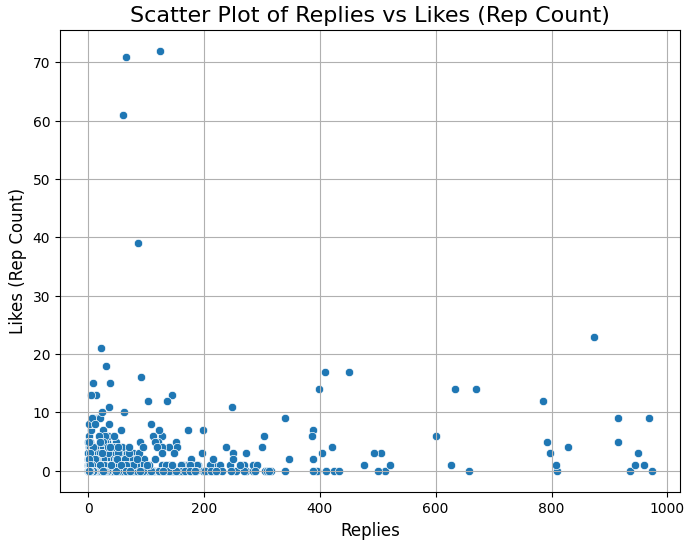
### EDA



*Figure 3.2*



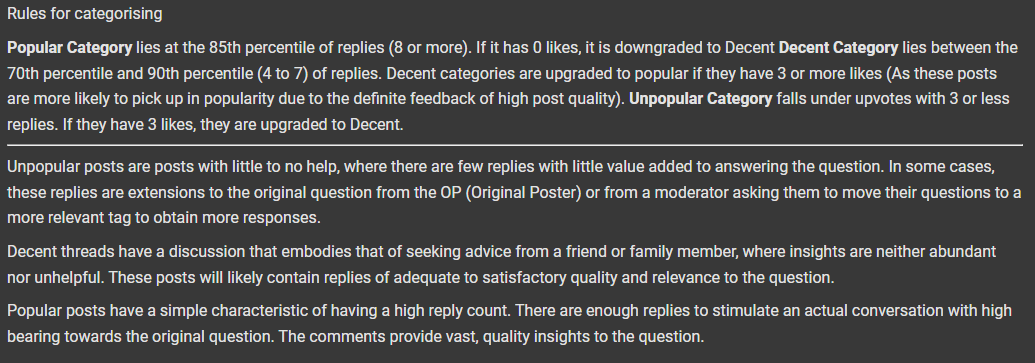
*Figure 3.3*



*Figure 3.4*

Figures 3.2 to 3.4 show how the average thread with 0 likes look like. There are enough replies to supply quality, personalised guidance for the original poster. This means that likes are not directly linked to the engagement a post receives. Likes on this forum tend to be under-utilised, proven by the lack of them despite a large number of replies. This leads to the idea that a thumbs up indicates a truly great post.

With data from Figure 3.4, we can obtain [posts with high likes and posts with low likes and high replies.](https://docs.google.com/document/d/1_2sq9Iwd0WGXQD3pITKPrLVJ-1DSHRft9phGxIevnwk/edit?usp=sharing) From this, we can see that Highly liked posts tend to be more informative and broadly relevant, while Low Likes with High Replies posts are often more specialized, discussion-driven, and focused on community building.



*Figure 3.5*

For Data Processing, I will essentially tokenize the post title and content, standardise the dates as ‘Days since post’ (Some are 1 hour ago, and others are actual dates etc…), check for duplicates, make general improvements to data quality and as mentioned, create a new metric for engagement that combines both replies and likes. Figure 3.5 shows the potential rules for the new metric.

### Model

**Input + Model Call:**

We will call the deep learning model **thrice**. Once with a timestamp of < 2 days, the other with a timestamp of 7 > n > 2 days, and the final prediction with a timestamp of > 7 days. This allows for predictions on how well the post will do over time. Borders are not fixed and could change after EDA.

**Additional Countermeasures:**

On top of the Deep Learning model, we will be connecting post content to a generative AI model on a Fast API application. Open AI’s GPT 4o mini will be used in combination with the [Langchain](https://github.com/langchain-ai/langchain) library to obtain a [structured output](https://python.langchain.com/docs/concepts/structured_outputs/#:~:text=LangChain%20provides%20a%20method%2C%20with_structured_output%28%29%2C%20that%20automates%20the,for%20all%20model%20providers%20that%20support%20structured%20output.) of whether the post content

1. Is ridiculous/Out of line
2. Leaks PII (Personally Identifiable Information)
3. Is relevant to the post title
4. Is relevant to education or careers/job experiences

This ensures that unnecessary/spam calls to the deep learning model are averted. Moreover, the AI model itself is inexpensive to call, reducing overall costs. Hosting the FastAPI application is also inexpensive, by utilising Google Cloud’s generous free requests for servers, or by hosting it on services like Render that comes with a free web service that you can constantly [ghost ping](https://console.cron-job.org/) to prevent the instance from automatically spinning down with inactivity.

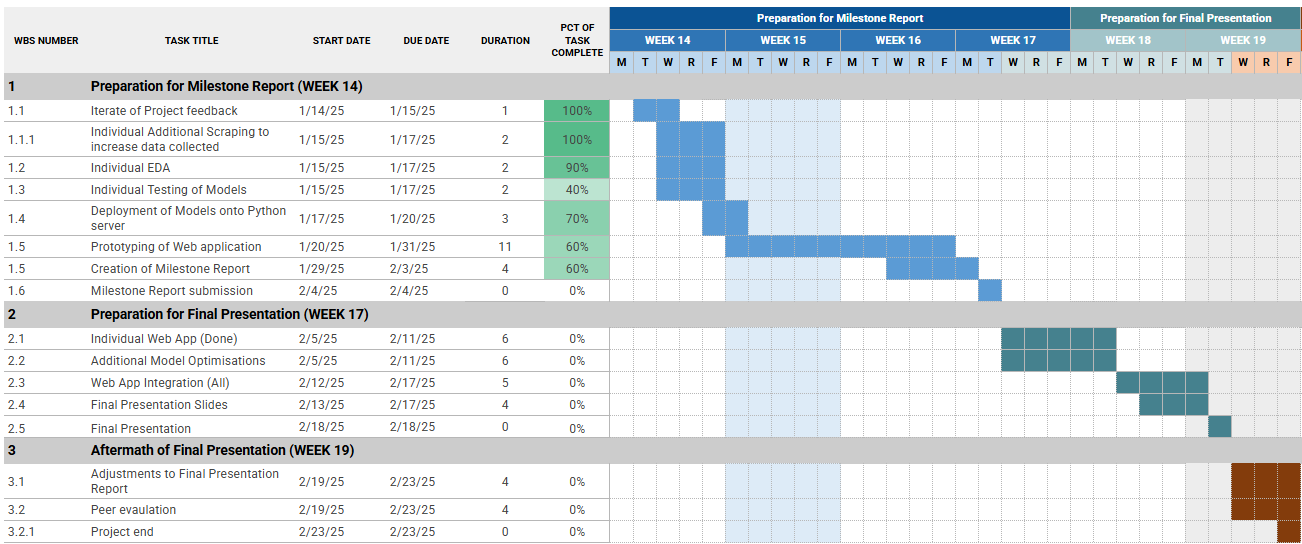
For the actual predictive model, I will first create a traditional ML model. I will be using Logistic Regression, Naive Bayes and Support Vector Machine each utilising Count Vectorizer and TFIDF Vectorizer, and finally DistilBert. Confusion matrices will also be plotted for a quick overview of how the model performs on both training and testing sets. If the minimally viable model is successful, I will move on to Deep Learning models, or maybe even utilise the DistilBert model and optimise it further.

### Web App

Interface for post input and output box (For Gen AI output and Engagement Prediction)

# Milestones & Deliverables

## Overall Gantt chart



[Gantt chart for Proposal](https://docs.google.com/spreadsheets/d/1YwhLa5v1ll4Cu3XsiquJzx67aWK8S0r3gJzFydp5zIg/edit?usp=sharing)

# 

## Conclusion

In conclusion, our project is a catalyst for transformative change, tackling the heart of social inequality and empowering individuals across the globe. By harnessing the power of Deep Learning, we don’t simply provide personalised career and educational insights – we break down barriers that have long held people back. Our AI chatbot goes beyond salary predictions, offering real-time cost-of-living insights that shatter outdated perceptions of value, helping individuals make informed choices, regardless of their background or location. Through educational analysis, we pinpoint areas where resources are lacking – be it in school enrolment or teacher availability, offering data-driven solutions to enhance access to quality education.

Our post engagement prediction model, meanwhile, revolutionises how users interact with forums. By increasing post visibility and boosting engagement, we maximise the chances of receiving personalised, actionable guidance. This ensures that no question goes unanswered, and no opportunity is missed. With every click and prediction, we are not just providing information – we are catalysing a movement towards a more inclusive, fairer society where opportunities are within reach for everyone. This is not just progress; it is a reimagining of what is possible, driven by data and a commitment to a sustainable future.

# Reference links (for any research done)

## Income Inequality across geography (Wei Heng)

### 

### Global Income Inequality (Wei Heng)

[1]: Rizwana Yasmeen, Wasi Ul Hassan Shah,

Unlocking the power of solar photovoltaics and ICT trade on income inequality across developed and developing economic spectrum,

Solar Energy,

Volume 274,

2024,

112558,

ISSN 0038-092X,

https://doi.org/10.1016/j.solener.2024.112558.

(https://www.sciencedirect.com/science/article/pii/S0038092X24002524)

[2]: [Reduce inequality within and among countries - United Nations Sustainable Development](https://www.un.org/sustainabledevelopment/inequality/)

