



**DEVI
SANSTHAN**

Dignity Education
Vision International

Leave no one behind

FINAL DELIVERABLE

Data Science Society @ UC Berkeley



**DEVI
SANSTHAN**
Dignity Education
Vision International
Leave no one behind

Meet The Team



Project Manager
Bella Chang



Project Manager
Zackary Oon



Consultant
Medha Iyer



Consultant
Lillian Jiang



Consultant
Deepta Jasthi



Consultant
Vennila Annamalai



Consultant
Michael Florip



Agenda

1. Question Generation
2. Analytics
3. Deliverables
4. Appendix



**DEVI
SANSTHAN**
Dignity Education
Vision International
Leave no one behind

Question Generation

Overview

Meet the Team

Agenda

Question Generation

Analytics

Questions

Appendix



Question Generation Overview

Prompt

Project Question

How can create a **streamlined process** for **question generation**, and a **new question bank** for the Literacy Now app?

Additional Questions

- How can we ensure ChatGPT generated questions are...
 - **Unbiased?**
 - **Appropriately challenging?**

Project Timeline

Exploratory Data Analysis

Explore dataset
Filter non-English questions
Frequency analysis

New Question Generation

ChatGPT
Web scraping

Create Pipeline

Python function
Input: question type, file path, expression



**DEVI
SANSTHAN**
Dignity Education
Vision International
Leave no one behind

Question Generation

Text-Based Questions

Meet the Team

Agenda

Question Generation

Analytics

Questions

Appendix



Text-Based Question Insights

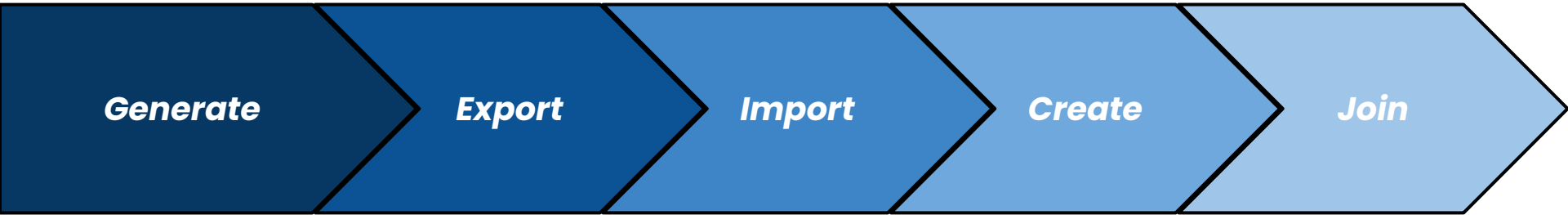
Mathematics

Example: Saba Has Rf 20 and she buys 6 eggs for Rf 18. How much money left with her?

- Rely on a **mix of text- and image-based prompts**
- More text-based questions could be produced with **Natural Language Processing (NLP) models**
- Selected categories: Addition, Subtraction, Multiplication, Division, and Perimeter and Area



Text-Based Question Generation Pipeline





What is ChatGPT?

- Tool driven by **AI technology** that can generate human-like text output
- Trained to understand and generate **natural language**
- Fast response time



 OpenAI



Mitigating Biases

Potential Problems

- Underlying biases
- Difficulty level
- Quality of questions



Our Solutions

Interpretability

Understand how ChatGPT creates its responses

Zero shot generation

Allows an AI language model to generate text it has been trained on



Interpretability

Goal: Check if ChatGPT has **inherent biases** when generating questions

Prompt: What **assumptions** did you use when making these questions? **How** did you come up with them?

ChatGPT Response:

- Assumed target audience of **elementary level** kids in India
- Capable of solving word problems that require **basic arithmetic** operations
- Drew from **common scenarios and objects** to elementary level kids in India
- Clearly phrase questions by using **simple** language and sentence structure
- **Culturally sensitive**

See full output in Appendix (slide 44)



Interpretability (cont.)

Goal: Check if ChatGPT has **inherent biases** when generating questions

How does this check for inherent biases?

- Provides **clear understanding** of what drives the model's predictions
- Makes it easy to **identify biases** that may be present

Conclusion

- Does **not** have any underlying biases when generating questions



Zero-Shot Generated Questions

Goal: Confirm ChatGPT generates **low-risk** and **easily interpretable** questions

Prompt: Come up with 10 scenario multiplication questions for elementary level students

ChatGPT Output: (first two)

- If you have 5 groups of 3 apples, how many apples do you have in total?
- If you have 4 boxes of crayons and each box has 8 crayons, how many crayons do you have in total?



Step 1: Generate

Example of ChatGPT Output and Response



generate questions similar to this: "Tom found 12 coins on the beach. Later, he found 18 more. How many coins did he find in total?". I want all the questions to use addition and have relatively simple numbers to add (total less than 100). Make sure all numbers are whole and each scenario could be understood by an 8-year-old. generate 10 unique questions.



1. Jane has 45 marbles, and she found 15 more in the playground. How many marbles does she have in total?
2. Alex saved up Rf 35 from his allowance, and he received Rf 25 as a gift from his grandmother. How much money does he have in total?
3. Sara has 28 stickers, and she was given 12 more by her friend. How many stickers does she have in total?
4. There are 30 apples in a basket, and 10 more apples were added. How many apples are there in the basket now?
5. Alex has 50 baseball cards, and he traded 8 with his friend. How many baseball cards does he have now?



Step 1: Generate (cont.)

Math Questions

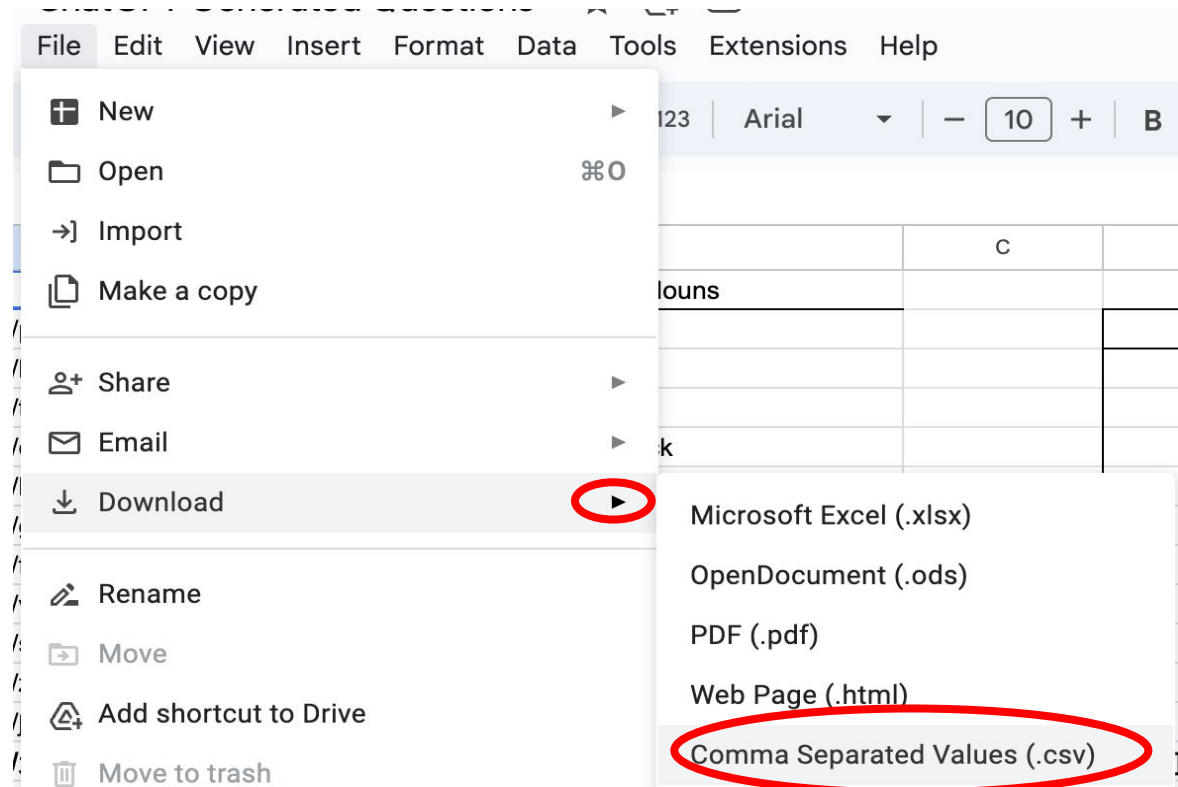
Goal: Create unique simple questions of similar style

Example Prompt

- “I want all the questions to use addition and have **relatively simple numbers to add (total less than 100)**. Make sure all numbers are **whole and positive** and each scenario could be **understood by an 8-year-old**. Generate **10 unique questions**.”



Step 2: Export





Step 3: Generate Solutions

Problem: ChatGPT's solutions were **not** always correct.

Solution: Generate the solutions ourselves.

Steps

1. Extract the numbers from the question
2. Define function that returns the expression to find solution
3. Applied the numbers to that function to get the solution
4. Save expression and solution to their respective columns
5. Create other answer choices



Question Pipeline: Output Table Format

Mathematics

Question	Answer Choices	Answer	Topic	Expression
Ruba buys 5 books for Rf 45 each. How much does she pay?	A) Rf 200 B) Rf 225 C) Rf 225 D) Rf 250	B) Rf 225	Multiplication	5×45
Sara purchases 8 pencils for Rf 10 each. How much does she pay in total?	A) Rf 80 B) Rf 90 C) Rf 100 D) Rf 110	A) Rf 80	Multiplication	8×10



Question Generation

Image-Based Questions



Image-Based Question Insights

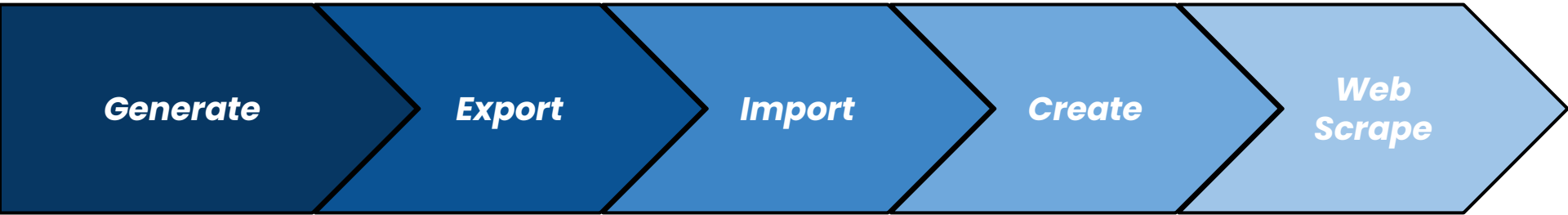
English

Example: Tick the correct statement for the given picture.

- Rely heavily on **image-based questions** and answer choices
- New questions can be **generated with ChatGPT** and **web scraping for images**
- Selected categories
 - Tick the correct word
 - Select the sound word



Image-Based Question Generation Pipeline





Step 1: Generate

ChatGPT Output and Response (Model 3.5)



Generate a table with two columns, the first corresponding to all the English phonemes and the second is a long list of one-syllable nouns that contain that phoneme

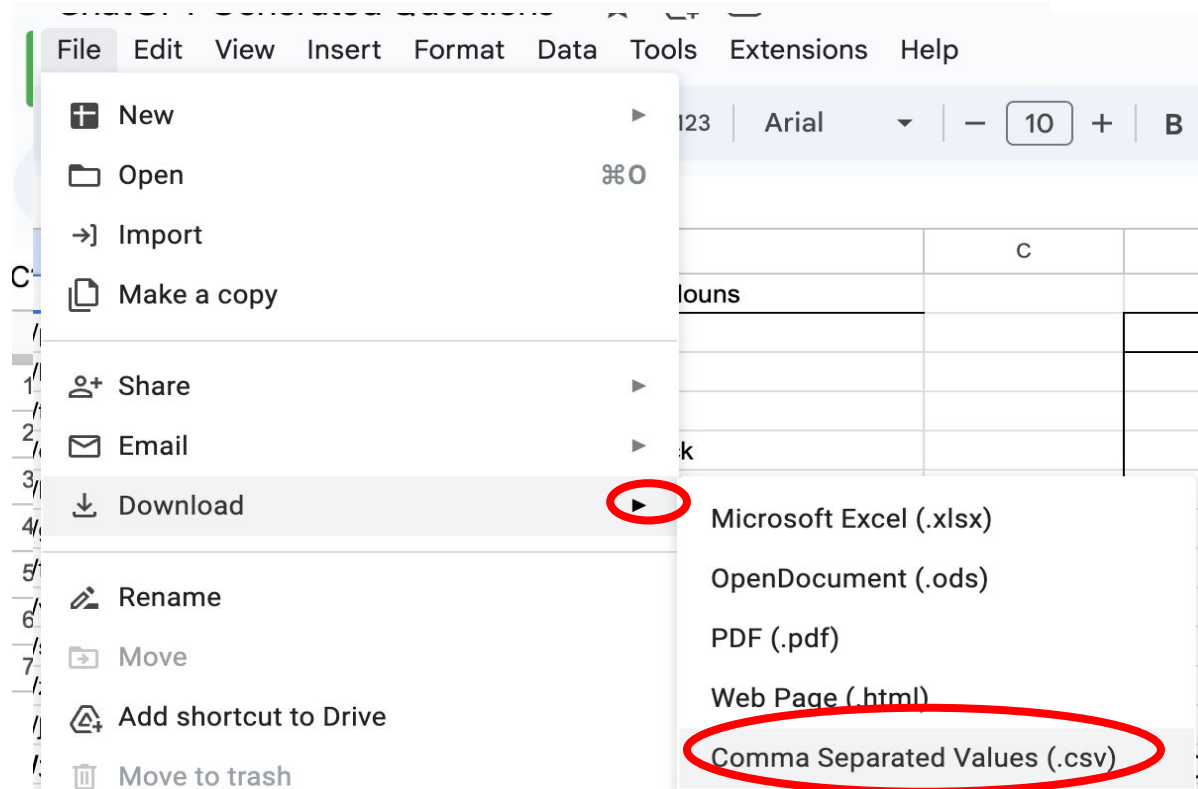


Sure, here is a table with some of the English phonemes and examples of one-syllable nouns that contain each phoneme. Note that English has more phonemes than this table shows, but I've included some of the most common ones:

Phoneme	Example Nouns
/æ/	cat, hat, mat, bat, rat
/ɛ/	bed, red, pen, hen, ten
/ɪ/	pin, bin, win, tin, fin
/ɒ/	dog, hog, log, jog, fog



Step 2: Export





Step 3: Generate Possible Answer Choices

Steps for Each Phoneme

1. Select an **example noun** from the list as an answer
2. **Create the question** for that answer
3. **Randomly sample 3 other answer choices** from the remaining example nouns
4. **Shuffle** the order of the 4 answer choices
5. **Add the newly generated answer choices** to column "Answer Choices"
6. **Add the phoneme** that this question uses to column "Phoneme"

See full output in Appendix (slide 50)



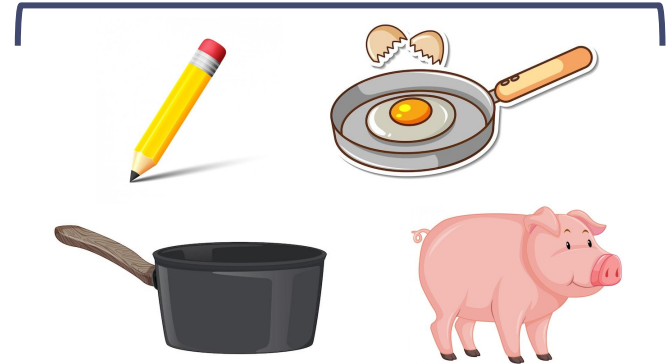
Step 4: Web Scrape Images for Answer Choices

Steps

1. Import **selenium**, **beautiful soup**, and **requests** libraries
2. **Loop through** each object you want an image of
 - a. Scroll to the first image element
 - b. Wait for the first image to present
 - c. Download the image as its name

Example Nouns (/p/ phoneme)

[pen, pan, pot, pig]





Challenges

- Web scraping
 - Works best on a **local computer**
 - Some images are **not downloaded correctly**
- ChatGPT Limitations
 - Prompts must be very specific
 - Accuracy isn't 100%
 - Limited to **text-only** responses (no images, etc.)
- Generating image-based questions
 - Inconsistent images for non-nouns (Example: bite)





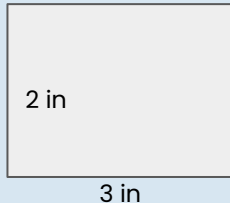
Next Steps & Suggestions

Use our walkthroughs and deliverables to create more questions

Image Based Math Questions

Example: Generate images of shapes with different lengths and ask for its area and perimeter

What is the perimeter of this shape?



Answer Choices:

10
9
13
7

Text based English Questions

Example: Fill in the blanks, complete the sentence

Fill this/that in the sentence:

__ shirt over there is expensive

Answer Choices:

those
that
this
these



Next Steps


Create questions that are **more complex** and **test multiple topics**

```
#example of the create_table function working
```

```
def sub_expr(x):
```

```
    return x[0] + "-" + x[1]
```

```
create_table("subtraction", "/work/CHATGPT/subtraction.csv", sub_expr)
```

	Question object <input type="checkbox"/>	Expression object <input type="checkbox"/>	Solution int64 <input type="checkbox"/>	Category object <input type="checkbox"/>	Answer Choices <input type="checkbox"/>
	Tom had 70 ... 5% Sarah has 3... 5% 18 others 90%	70-25 5% 36-15 5% 18 others 90%	8 - 55 	subtraction 100%	[39, 51, 48, 4... 5% [27, 16, 21, 11] 5% 18 others 90%
10	A classroom had 24 students and 1...	24-10	14	subtraction	[12, 14, 19, 24]
11	There were 70 candies in a jar a...	70-16	54	subtraction	[51, 52, 54, 49]
12	Hannah had 55 stickers and she...	55-30	25	subtraction	[34, 25, 18, 16]



An Example

Create more **complex questions** using our ChatGPT prompt generation tips

- Goal: Generate a question that uses the addition operator then the multiplication operator
 - Example: Medha finds one pencil on the ground and adds it to her collection of 5 pencils. She distributes all the pencils to three of her friends. How many pencils does each friend get
- How?

```
add_then_div_expr(x): #x is an array of all the numbers from the question in order
```

```
Return (x[0] + x[1]) / x[2] #adds x[0] and x[1], then divides by x[2]
```

```
create_table("Addition and Division", filepath, add_then_div_expr)
```



Analytics

Overview

Meet the Team

Agenda

Question Generation

Analytics

Questions

Appendix



Overview [Analytics Team]

Prompt

Project Question: How does accuracy (defined as raw grade - "the numerical or scale grade from activity") differ based on different demographics of a user?

Additional Questions: Is learning being conducted equitably? What are the problems with data collection right now?

Datasets Used

- mdl_student
- mdl_learner
- mdl_grade_grades
- mdl_user_enrollments
- mdl_h5pactivity_attempts

Project Timeline



Cleaning + EDA + Consolidating

Understanding and joining data tables



Visualizations + Hypothesis Testing

Exploring features that could affect students performance



Future Data Collection

How to improve data collection for better analysis in the future



**DEVI
SANSTHAN**
Dignity Education
Vision International
Leave no one behind

Analytics

Visualizing Effectiveness by Features

Meet the Team

Agenda

Question Generation

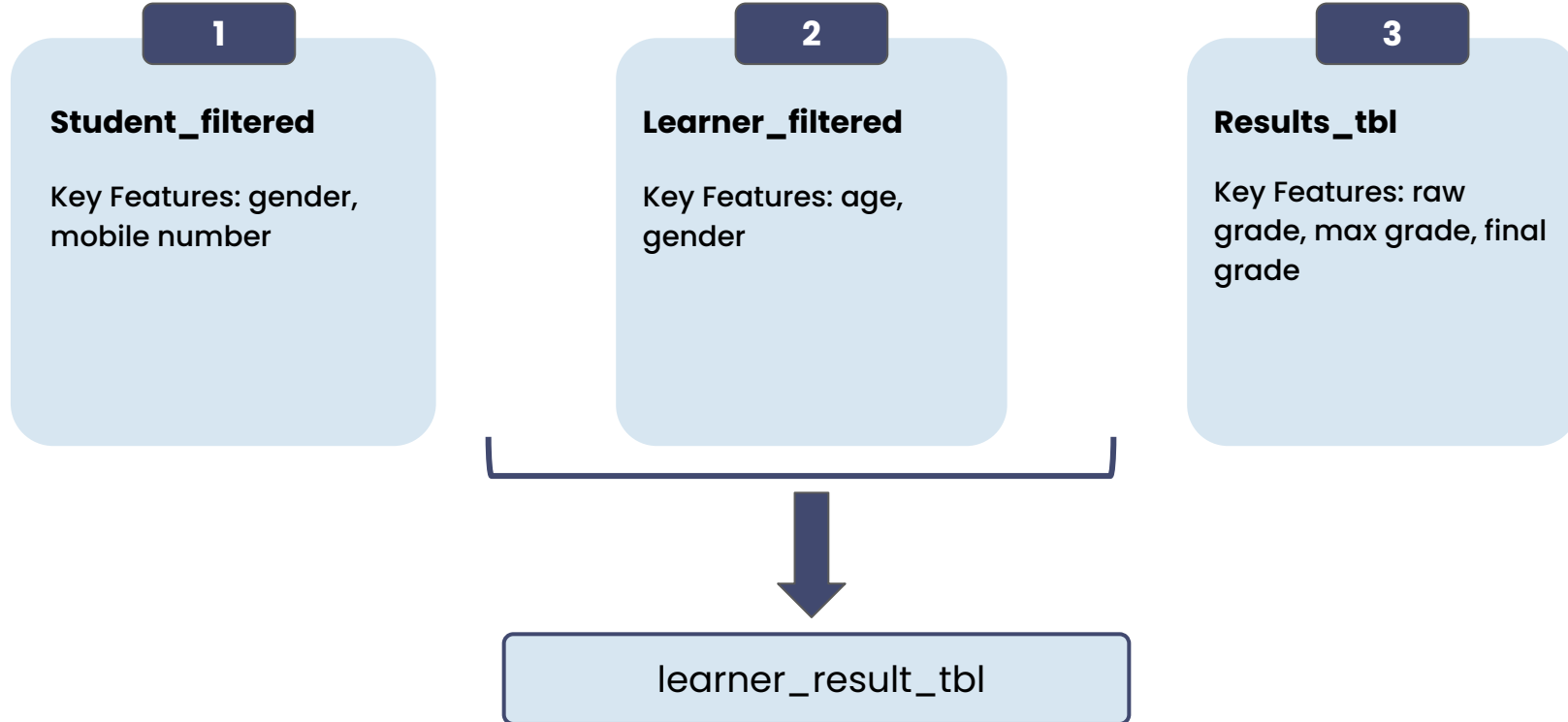
Analytics

Questions

Appendix



Data Cleaning: Table Creation





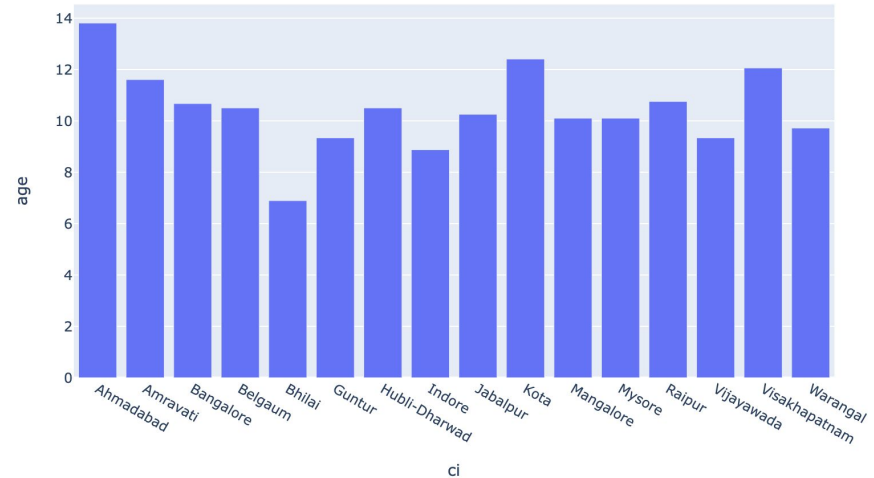
Mean Age of Learners by Location

Mapped **mobile area code** to a **specific city** in order to understand geographic distribution of users

	ci object	mobile_area_co...
	Agra 1.2%	12 6.2%
	Ahmadabad 1.2%	25 3.7%
	79 others 97.5%	56 others 90.1%
20	Coimbatore	42
21	Cuttack	67
22	Dehradun	13
23	Delhi	11
24	Dhanbad	32
25	Faizabad	52
26	Faridabad	12
27	Ghaziabad	12
28	Gorakhpur	55
29	Guntur	86



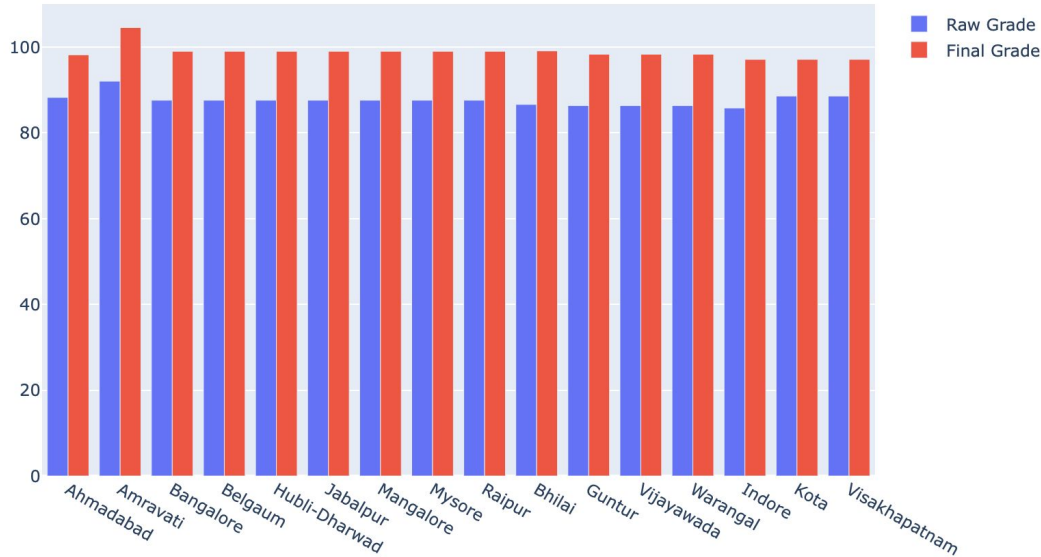
Mean Age of Learners per Location





1 – Using Mean Ages to Visualize App Effectiveness per City

Rough Raw and Final Grade per City



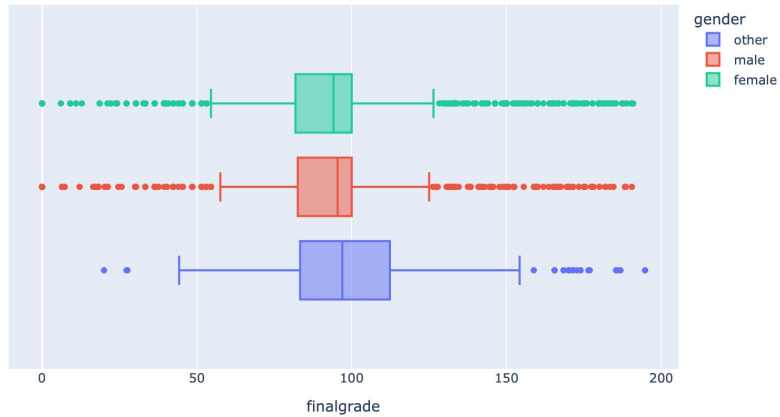
Observations

- Rough estimate of performance for each city
- Do all users score the same?

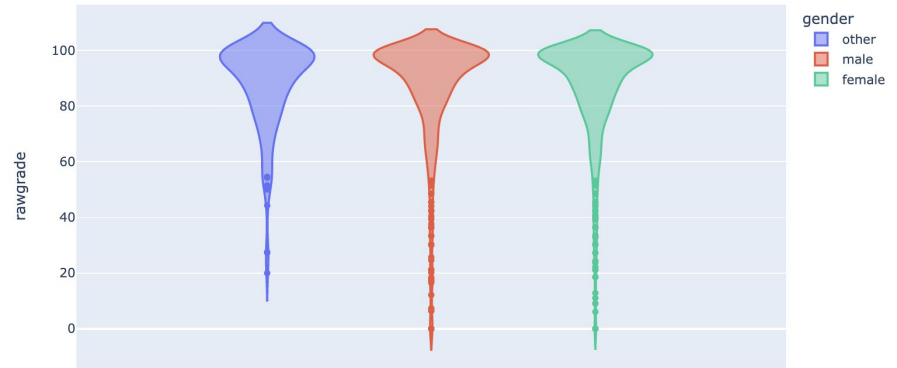


2 – App Effectiveness by Gender

Mean Grade of Learners Sorted by Gender



Rawgrade Filtered by Gender



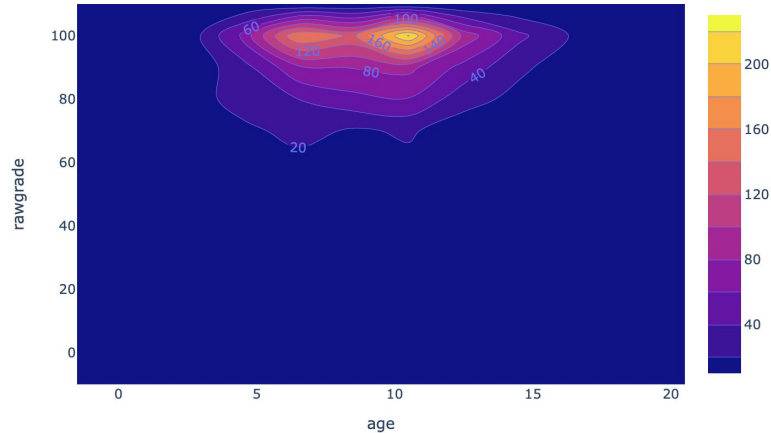
Observations

- Mode: 98-100 across all genders
- Mean: ~95-100 across all genders (lots of outliers in other and male)

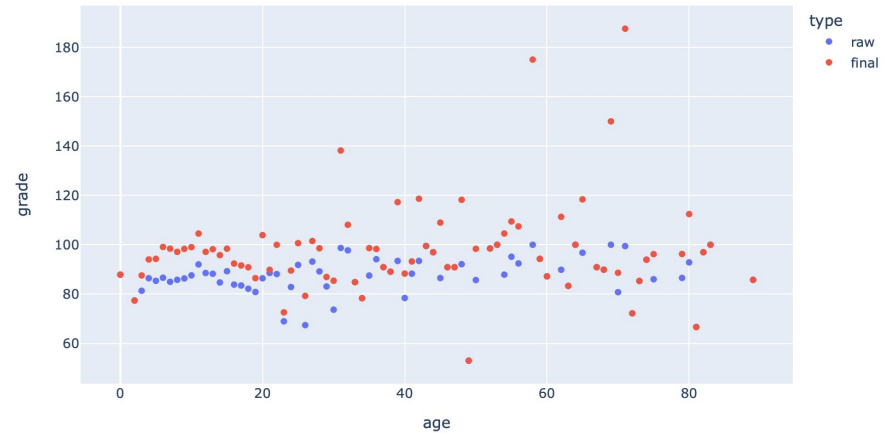


3 – App Effectiveness by Age

Density Plot for rawgrade by age



Mean Grades (Raw and Final) by Age



Observations

- Age 10 and age 7 score around 100 the most frequently
- Mean is around 90-100 for all ages (unsure of max possible score)



Takeaways

- From the visualizations we see that average grade for all ages and genders is about 100
 - This could mean that the question sets are **too easy**
- The question accuracy among genders and ages seems to be relatively evenly distributed
 - There is **no favoritism** towards or bias against a gender in how the questions are phrased/currently generated!
- Raw and final grade seem to have very different ranges
 - Raw: user score in a particular exercise, final: user score overall
 - Raw goes up to a score of 100
 - Final goes up to a score of 200
 - Different scales could be confusing in comparison to one another



Hypothesis Testing with Gender

Question

Is there a significant difference between male learners' raw grades in comparison to female learners?

Null Hypothesis

The mean raw grade between female and male students is roughly the same. Any difference is due to chance.

Alternative Hypothesis

The mean raw grades between male and female students are not equal. A difference is due to something other than random chance.



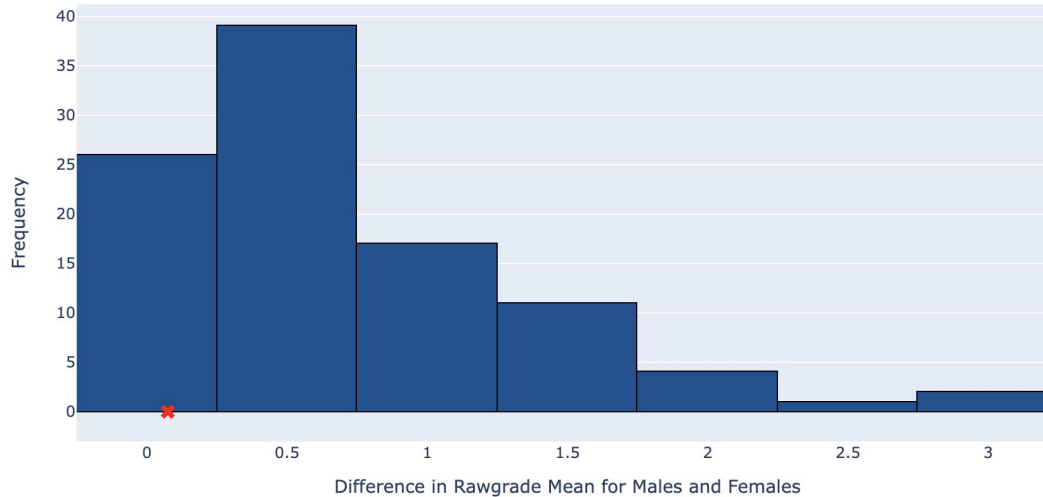
Test

Conducted a bootstrapped hypothesis test to find p-value



Hypothesis Testing with Gender

Histogram of Simulated Test Statistics



P-value: ~ 0.95

- Well above a statistical cut-off of 0.05
- Mean final grade between male and female students roughly the same
- **Distribution** and **attentiveness** towards each group of learners is going smoothly



Analytics

Improving Data Collection



Suggestions

- **Clarify the difference between learner and student** in a way that is more intuitive to outside users
- Have a way to **map students' campus_id to location** in order to geographically consolidate data
- Having **one primary key among all/most tables: 'user_id'**
- Getting **more comprehensive data about learners' scores**: specifically, in each section of the app that they utilize (English, Hindi, math, etc.)
- **Gathering as much demographic data as possible** from the users
- **Standardizing data collection by creating dropdowns** for entering user data



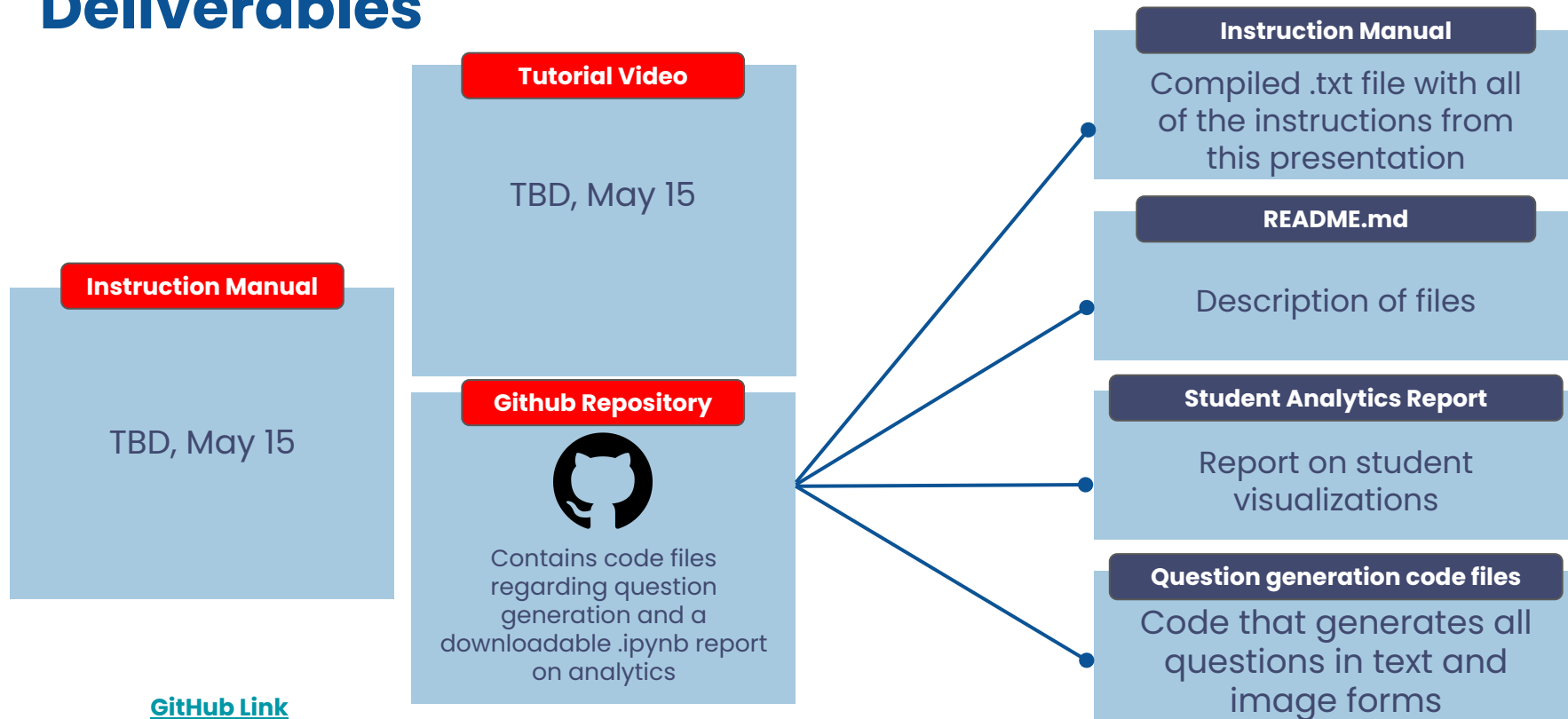
Next Steps*

- Creating a model that can help to **predict user scores based on demographic features** (age, gender, location)
- Pinpointing **further features that may be causing some of the differentiations in grade**, as seen through the large amount of outliers

**These steps can only be possible after taking the suggestions in improving data collection*



Deliverables



[GitHub Link](#)



THANK YOU
Any Questions?



**DEVI
SANSTHAN**
Dignity Education
Vision International
Leave no one behind

Appendix

Meet the Team

Agenda

Question Generation

Analytics

Questions

Appendix



Appendix

[Slide 15, ctd.]

Question: “What assumptions did you use when making these questions? How did you come up with them?”

ChatGPT response: “When making these questions, I made a few assumptions about the target audience of elementary level kids in India, as well as their expected knowledge and skills. Specifically, I assumed that:

Students have a basic understanding of arithmetic operations such as multiplication.

Students are familiar with common objects and activities that are relevant to the Indian context.

Students are capable of solving word problems that require them to apply their knowledge of arithmetic operations to real-world situations.

To come up with the questions, I drew from common scenarios and objects that are likely to be familiar to elementary level kids in India. I also tried to vary the scenarios and numbers used to provide a diverse set of questions that would challenge students' problem-solving skills. Finally, I attempted to phrase the questions in a way that would be clear and understandable for young learners, using simple language and avoiding technical terms or complex sentence structures.

I also tried to be culturally sensitive and respectful by avoiding any topics that could be considered inappropriate or offensive in the Indian context. For example, I did not use any scenarios related to non-vegetarian food or cultural practices that may not be universally understood.”



Text-Based Question Generation Pipeline

[Slide 9, ctd.]

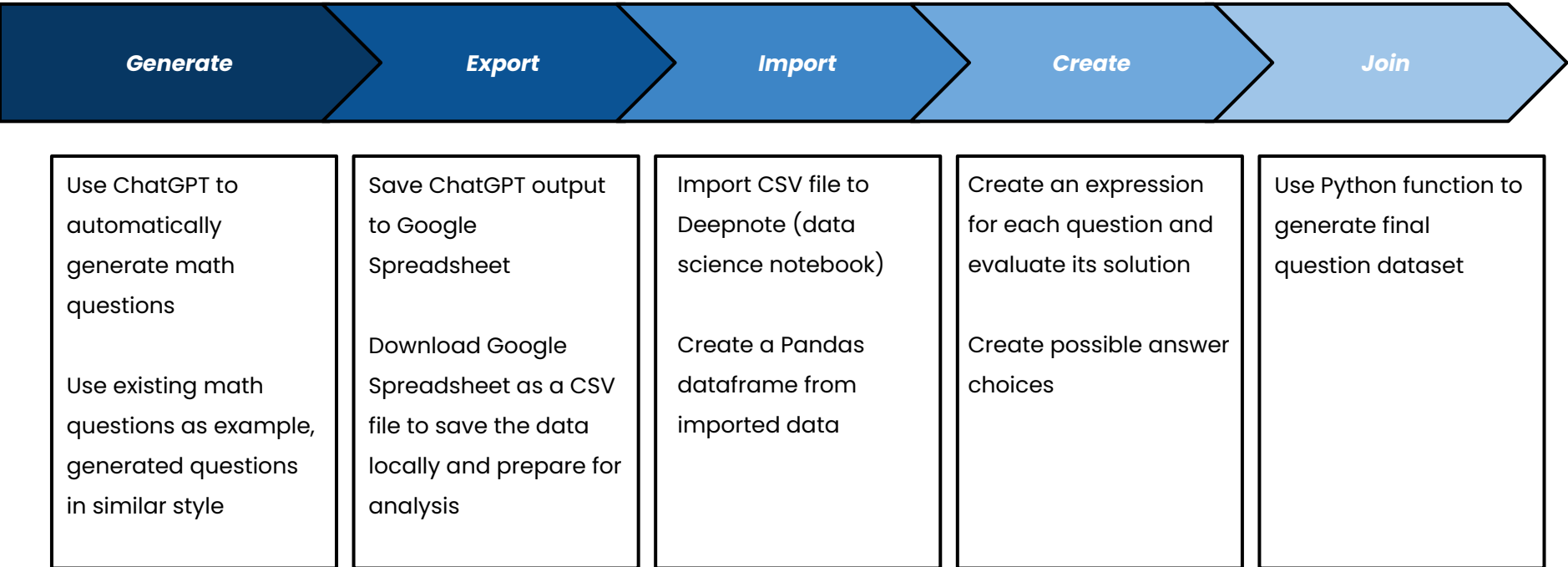
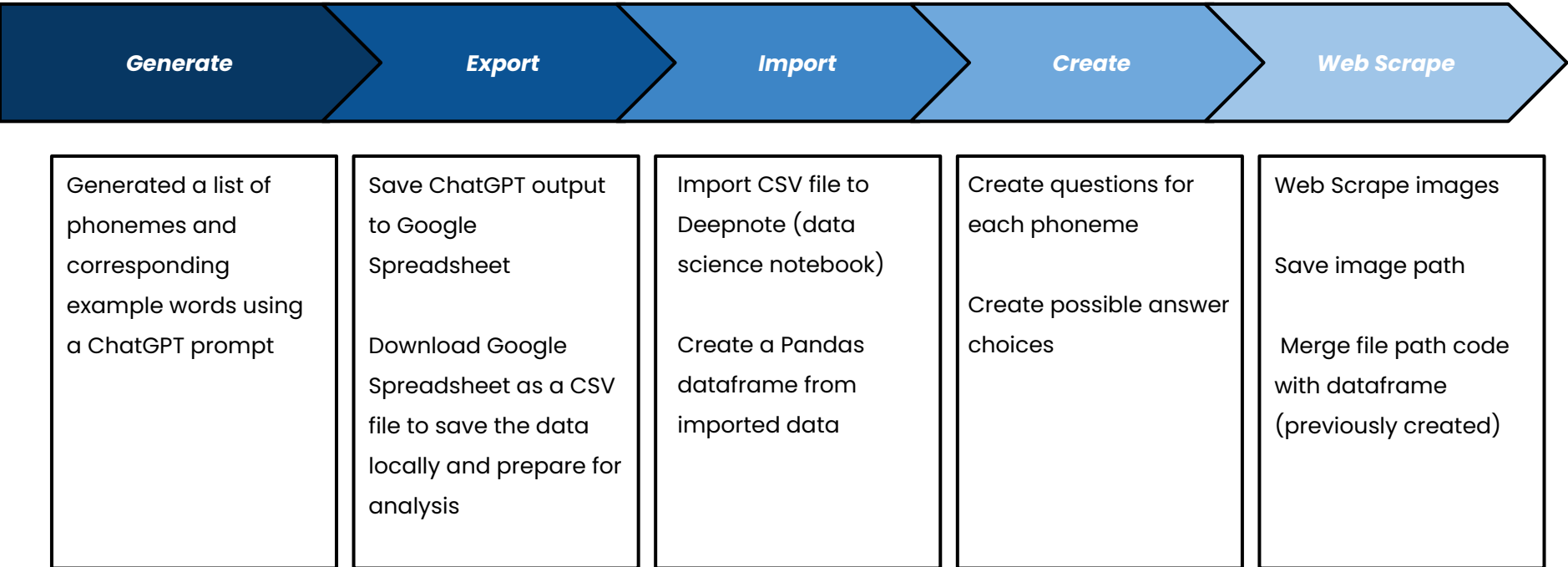




Image-Based Question Generation Pipeline

[Slide 21, ctd.]





Appendix

[Slide 24, code]

```
q_index = 0
for index, row in phoneme_list.iterrows():
    phoneme = row['English Phoneme']
    for word_index in range(len(row['Word List'])):
        generated_english_questions.loc[q_index, "Generated Question"] = "Select the correct " + phoneme + " sound word in the
picture"
        generated_english_questions.loc[q_index, "Answer"] = row['Word List'][word_index]
        answer_choices = random.sample(row['Word List'][:word_index] + row['Word List'][word_index+1:], min(3, len(row['Word
List'])))) + [row['Word List'][word_index]]
        random.shuffle(answer_choices)
        generated_english_questions.loc[q_index, "Answer Choices"] = answer_choices
        generated_english_questions.loc[q_index, "Phoneme"] = phoneme
    q_index += 1
```



Appendix

[Slide 33, ctd.]

In addition to using phone number area codes to approximate students locations, we joined other tables to map age to raw score and gender:

For example: we had two tables: mdl_h5pactivityid and mdl_learner

Using the user_id id column in each of these tables, we were able to join the two tables together and map the ages and genders in the mdl_learner table to the grades in the mdl_h5pactivityid column. This helped us make the visualizations you see on slides 35, 36, and 37.