

Lesson 07 Demo 05

Generating Synthetic and Augmented Datasets Using Generative AI Tools

Objective: To generate synthetic and augmented datasets using generative AI tools for validating data integrity, enhancing test coverage, and ensuring robustness during the testing phase of the Software Development Life Cycle

Tools required: Mostly AI and ChatGPT 4

Prerequisites: Basic knowledge and difference of synthetic and augmented dataset

Steps to be followed:

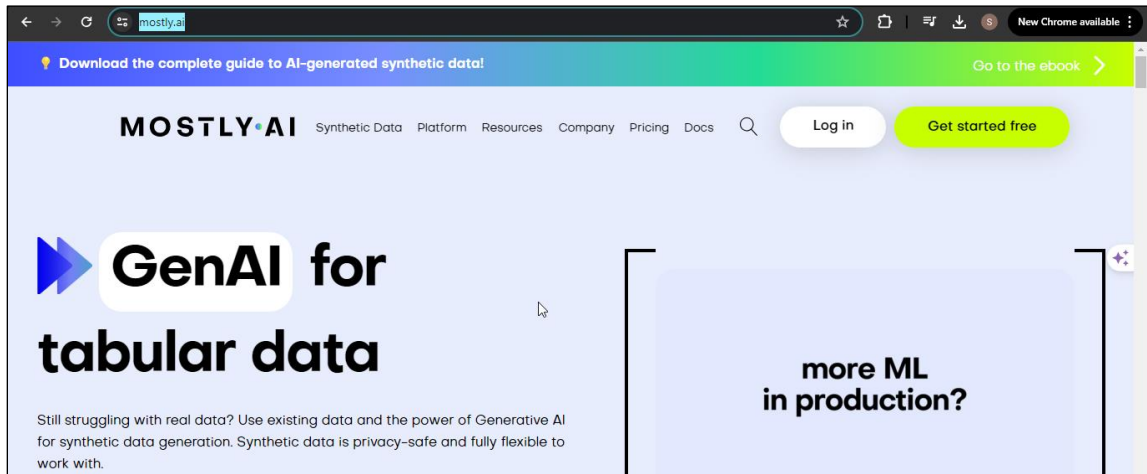
1. Generate the synthetic dataset using Mostly AI
2. Generate the synthetic and augmented dataset using ChatGPT

Note: Please note that all the GenAI tools used in this exercise can produce varied outputs even when presented with similar prompts. Thus, you may get different outputs for the same prompt.

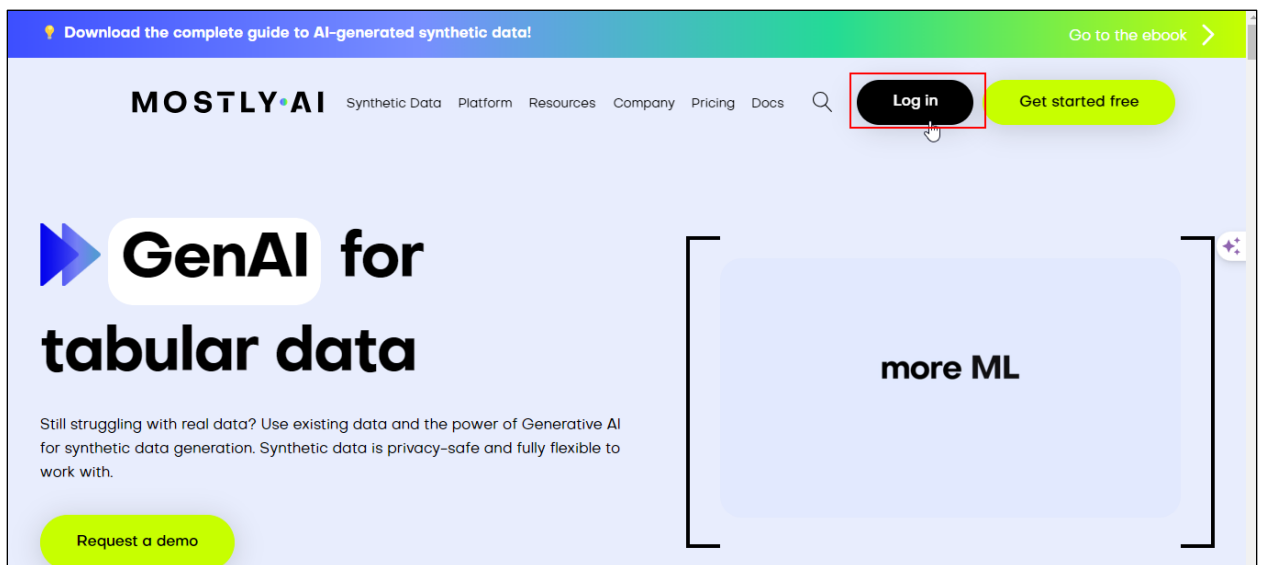
Note: Mostly AI may offer some customization options, primarily focusing on generating realistic synthetic data from scratch rather than augmenting existing datasets. ChatGPT 4, on the other hand, is a versatile tool that can be used to generate synthetic data and manipulate and augment existing datasets effectively.


Step 1: Generate the synthetic dataset using Mostly AI

1.1 Navigate to <https://mostly.ai/>



1.2 Click on the **Log in** button and enter the credentials






MOSTLY AI

Sign in to start generating synthetic data

* E-mail address

☐ Remember me

Sign in

 Log In with Google

Don't have an account yet?

Sign up here

Note: Create a new account by clicking on the **Sign up here** button if you do not have an account


1.3 Click on the **Upload file** button

MOSTLY AI

HomeGeneratorsSynthetic datasetsConnectors

5.0 credits

AK


Welcome,  🤖


Data innovation through Generative AI: train your generator to craft synthetic datasets.


Latest generators


Sample Census Data Generator	✓ Ready	1 month ago	MA
Sample Baseball Data Generator	✓ Ready	1 month ago	MA

Train a generator with your own data

 Upload file >


 Connect to source >

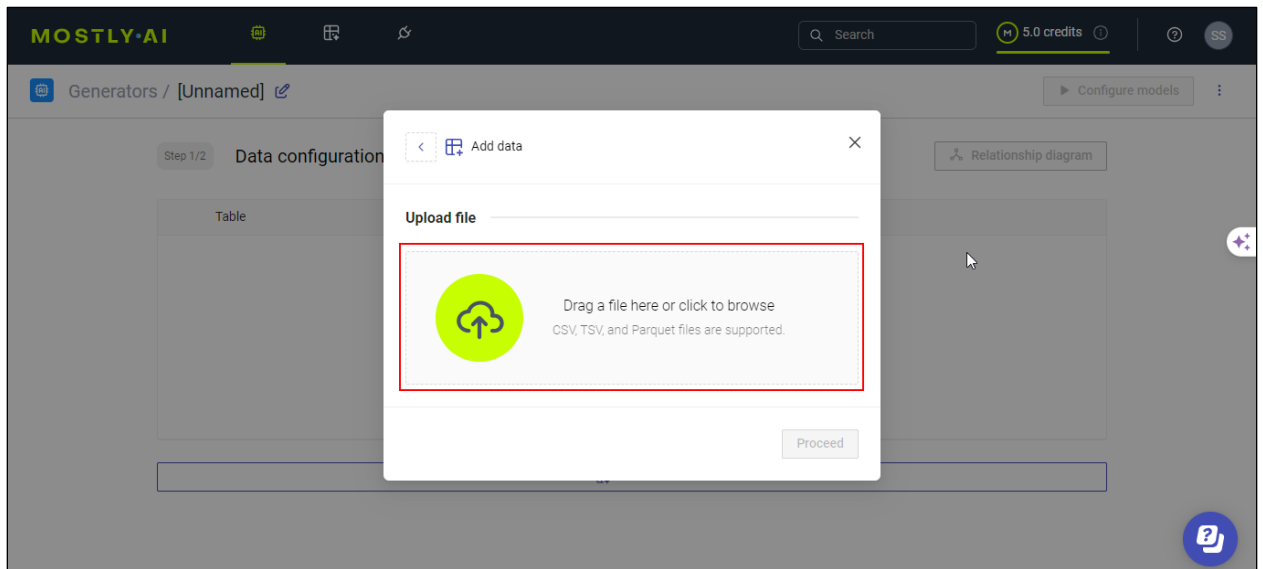
 Get API key >

 How to start?

Explore the available generators and **start generating data.**

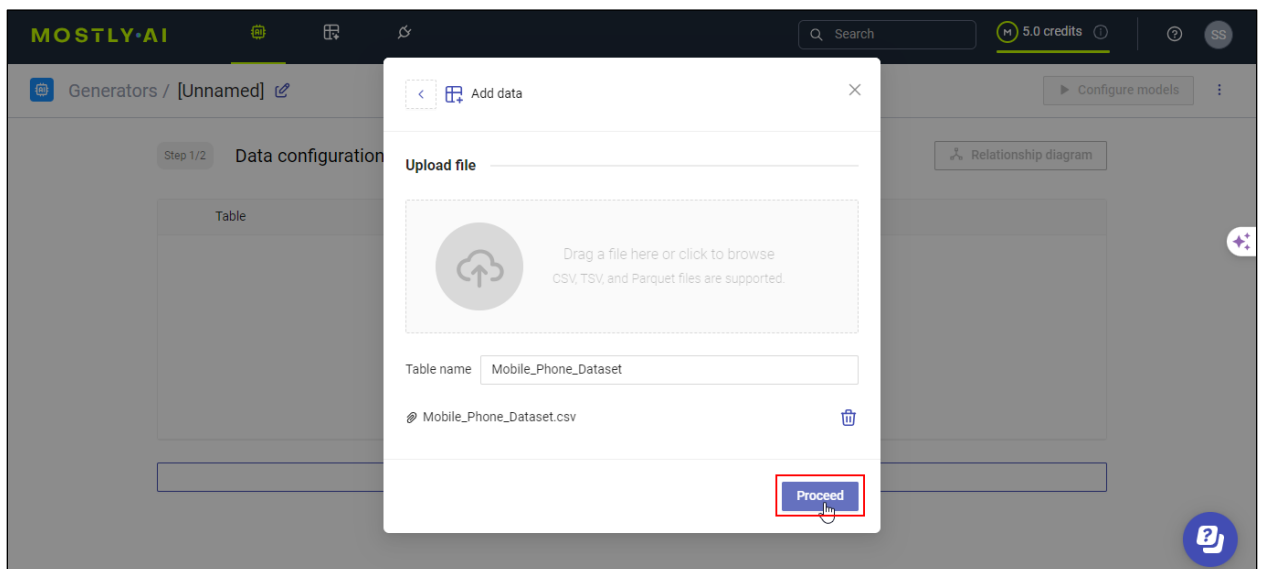
+ New synthetic dataset



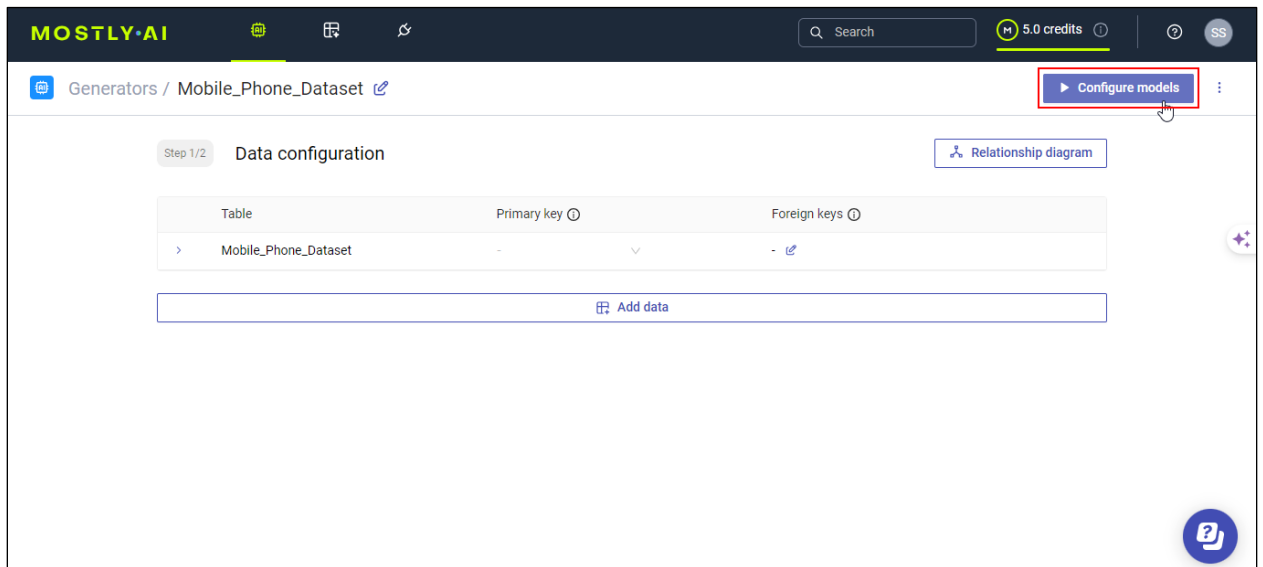


Note: You may use any sample data set; if you do not have any data set, use ChatGPT to create a demo dataset of the desired number of rows and columns.
Use the following prompt:
Create a dataset of 100 rows for a list of mobile phones, including year of launch, model, price, country, camera configuration, and discount offer

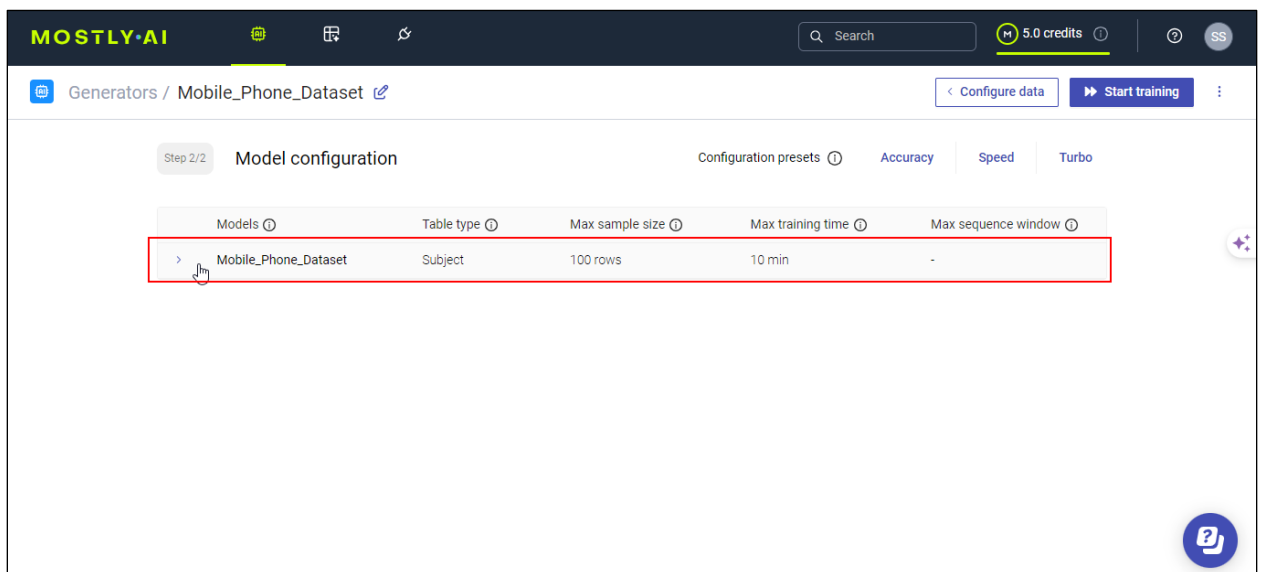
1.4 Click on the **Proceed** button



1.5 Click on the **Configure models** button, as shown in the screenshot below:



1.6 Click on the **Mobile_Phone_Dataset** list and make the desired changes accordingly, as shown in the screenshots below:



Generators / Mobile_Phone_Dataset [🔗](#)

[< Configure data](#)
[▶▶ Start training](#)

Step 2/2

Model configuration

Configuration presets ⓘ

Accuracy

Speed

Turbo

Models ⓘ	Table type ⓘ	Max sample size ⓘ	Max training time ⓘ	Max sequence window ⓘ
▼ Mobile_Phone_Dataset	Subject	100 rows	10 min	-

Max sample size

100

rows

Max training time

10

mins

Max sequence window

Not applicable for subject tables

Max training epochs ⓘ

100

epochs

Model size ⓘ

Medium

▼

Batch size ⓘ

Auto

▼

Flexible generation ⓘ

On

Off

Value protection ⓘ

On

Off

Rare category replacement method ⓘ

Constant

▼

?

Note: Customize the dataset parameters such as column names and data distribution to match specific requirements

1.7 Click on the **Start training** button

Generators / Mobile_Phone_Dataset [🔗](#)

[< Configure data](#)
[▶▶ Start training](#)

Step 2/2

Model configuration

Configuration presets ⓘ

Accuracy

Speed

Turbo

Models ⓘ	Table type ⓘ	Max sample size ⓘ	Max training time ⓘ	Max sequence window ⓘ
▼ Mobile_Phone_Dataset	Subject	100 rows	10 min	-

Max sample size

100

rows

Max training time

10

mins

Max sequence window

Not applicable for subject tables

Max training epochs ⓘ

100

epochs

Model size ⓘ

Medium

▼

Batch size ⓘ

Auto

▼

Flexible generation ⓘ

On

Off

Value protection ⓘ

On

Off

Rare category replacement method ⓘ

Constant

▼

?

1.8 Scroll down to the **Training status** section

Generators / Mobile_Phone_Dataset [Share](#) [Generate synthetic data](#)

Training status In progress

Model	Step	Progress	Duration
Mobile_Phone_Dataset	Fetch training data	<div><div></div></div>	1s
Mobile_Phone_Dataset	Analyze training data	<div><div></div></div>	2s
Mobile_Phone_Dataset	Encode training data	Queued	-
Mobile_Phone_Dataset	Train AI model	Queued	-
Mobile_Phone_Dataset	Generate data for model report	Queued	-
Mobile_Phone_Dataset	Analyze data for model report	Queued	-

Configuration

Overview
Data insights
Model samples
Training status
Configuration

1.9 Examine the generated synthetic data in the **Model samples** section to ensure it meets your requirements

Generators / Mobile_Phone_Dataset [Share](#) [Generate synthetic data](#)

Model samples

Mobile_Phone_Dataset

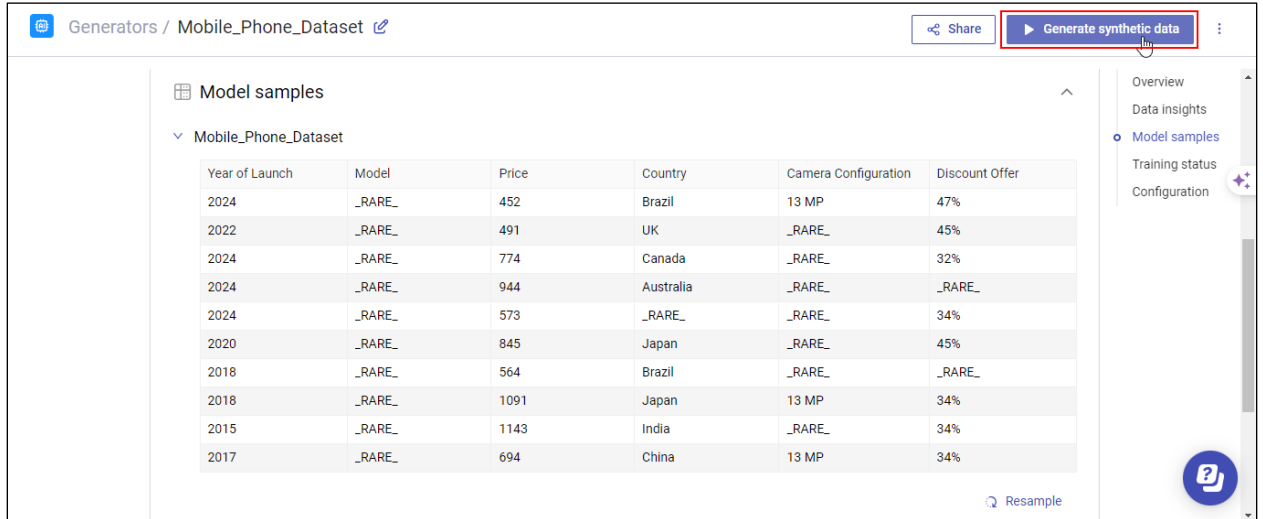
Year of Launch	Model	Price	Country	Camera Configuration	Discount Offer
2024	_RARE_	452	Brazil	13 MP	47%
2022	_RARE_	491	UK	_RARE_	45%
2024	_RARE_	774	Canada	_RARE_	32%
2024	_RARE_	944	Australia	_RARE_	_RARE_
2024	_RARE_	573	_RARE_	_RARE_	34%
2020	_RARE_	845	Japan	_RARE_	45%
2018	_RARE_	564	Brazil	_RARE_	_RARE_
2018	_RARE_	1091	Japan	13 MP	34%
2015	_RARE_	1143	India	_RARE_	34%
2017	_RARE_	694	China	13 MP	34%

[Resample](#)

Overview
Data insights
Model samples
Training status
Configuration

Note: You may click **Resample** if the sample data does not meet your requirements.

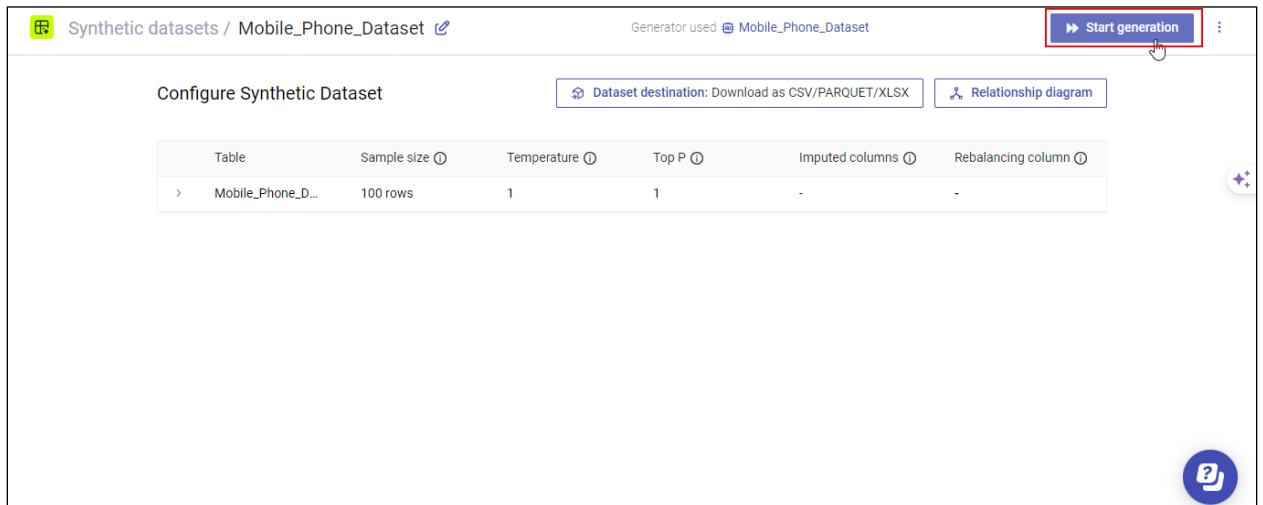
1.10 Click on the **Generate synthetic data** button



The screenshot shows the 'Generators / Mobile_Phone_Dataset' interface. At the top right, there is a 'Share' button and a 'Generate synthetic data' button, which is highlighted with a red box. Below the header, there is a 'Model samples' section with a table of data. On the right side, there is a sidebar with navigation links: Overview, Data insights, Model samples (selected), Training status, and Configuration. At the bottom right, there is a 'Resample' button and a help icon.

Year of Launch	Model	Price	Country	Camera Configuration	Discount Offer
2024	_RARE_	452	Brazil	13 MP	47%
2022	_RARE_	491	UK	_RARE_	45%
2024	_RARE_	774	Canada	_RARE_	32%
2024	_RARE_	944	Australia	_RARE_	_RARE_
2024	_RARE_	573	_RARE_	_RARE_	34%
2020	_RARE_	845	Japan	_RARE_	45%
2018	_RARE_	564	Brazil	_RARE_	_RARE_
2018	_RARE_	1091	Japan	13 MP	34%
2015	_RARE_	1143	India	_RARE_	34%
2017	_RARE_	694	China	13 MP	34%

1.11 Click on the **Start generation** button



The screenshot shows the 'Synthetic datasets / Mobile_Phone_Dataset' interface. At the top right, there is a 'Start generation' button, which is highlighted with a red box. Below the header, there is a 'Configure Synthetic Dataset' section with a table of configuration parameters. At the bottom right, there is a help icon.

Table	Sample size	Temperature	Top P	Imputed columns	Rebalancing column
> Mobile_Phone_D...	100 rows	1	1	-	-

You will see the following interface:

Synthetic datasets / Mobile_Phone_Dataset

Generator used Mobile_Phone_Dataset

Share
Download synthetic dataset

Generated by **Syed Sharoz** • Created on April 28, 2024 at 19:39 • 0

Overall accuracy
Data points
Description

-
-
Click here to edit description...

Used credits
-

Data insights
Data insights will be available once the generation status is Ready.

Data samples
Data samples will be available once the generation status is Ready.

Generation status Queued

- Overview
- Data insights
- Data samples
- Generation status
- Configuration

1.12 Click on the **Download synthetic dataset** dropdown and select the desired option to download or directly export the synthetic dataset, as shown in the screenshots below:

Synthetic datasets / Mobile_Phone_Dataset

Generator used Mobile_Phone_Dataset

Share
Download synthetic dataset

Data insights

Table	Accuracy				Distances	Reports
	Overall	Univariate	Bivariate	Coherence		
Mobile_Phone_Dataset	28.2% (83.6%)	42.8%	13.7%	-	2.75 (2.79)	Model Data

Data samples

Mobile_Phone_Dataset

Year of Launch	Model	Price	Country	Camera Configuration	Discount Offer
2023	_RARE_	596	_RARE_	_RARE_	34%
2023	_RARE_	1094	Japan	13 MP	34%

- Overview
- Data insights
- Data samples
- Generation status
- Configuration

Synthetic datasets / Mobile_Phone_Dataset [Generator used Mobile_Phone_Dataset](#) [Share](#) [Download synthetic dataset](#)

Data insights

Table	Accuracy				Distances	Reports
	Overall	Univariate	Bivariate	Coherence		
Mobile_Phone_Dataset	28.2% (83.6%)	42.8%	13.7%	-	2.75 (2.79)	Model Data

Data samples

Mobile_Phone_Dataset

Year of Launch	Model	Price	Country	Camera Configuration	Discount Offer
2023	_RARE_	596	_RARE_	_RARE_	34%
2023	_RARE_	1094	Japan	13 MP	34%

Download as CSV
Download as PARQUET
Download as XLSX

Generation status
Configuration

Step 2: Generate the synthetic and augmented dataset using ChatGPT

2.1 Navigate to the <https://chat.openai.com> website and log in to your account

ChatGPT

chat.openai.com/auth/login

ChatGPT

Suggest fun activities
for a team-building day with remote employees

Get started

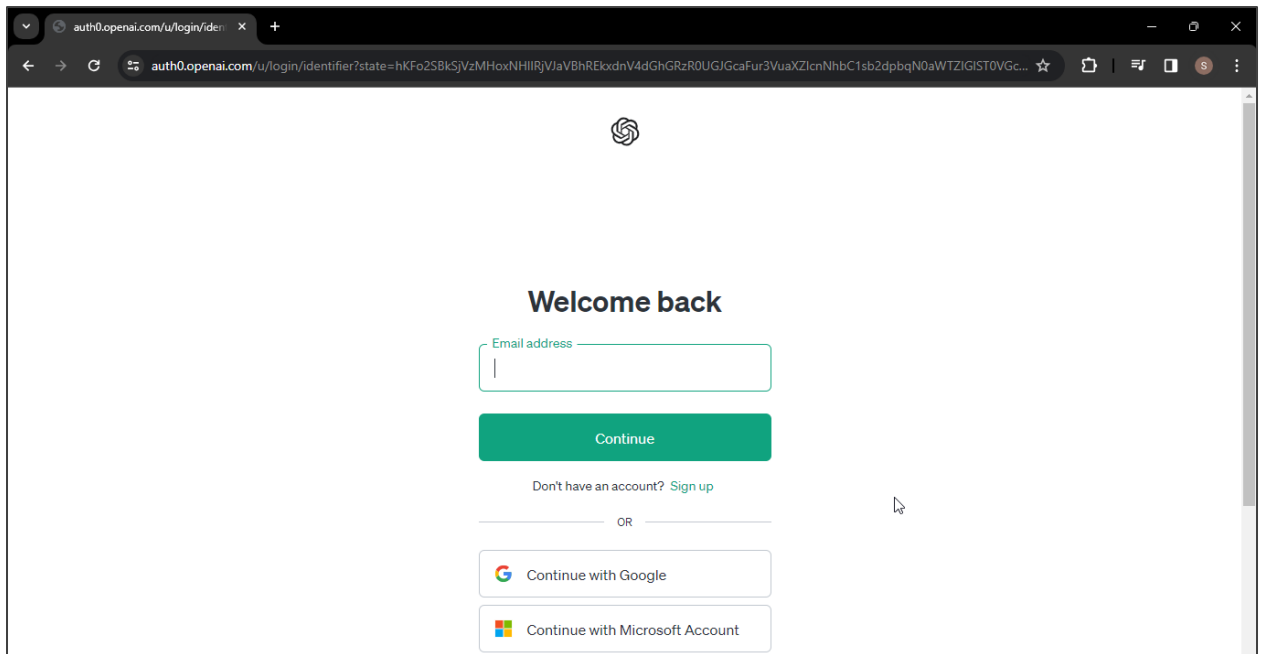
[Log in](#) [Sign up](#)

OpenAI

[Terms of use](#) | [Privacy policy](#)

[Get citation](#)

Note: Sign up if you do not have an account



2.2 Generate the synthetic data using the following prompt:

Create synthetic data including the following columns:

Card Number

Cardholder name

Expiry date

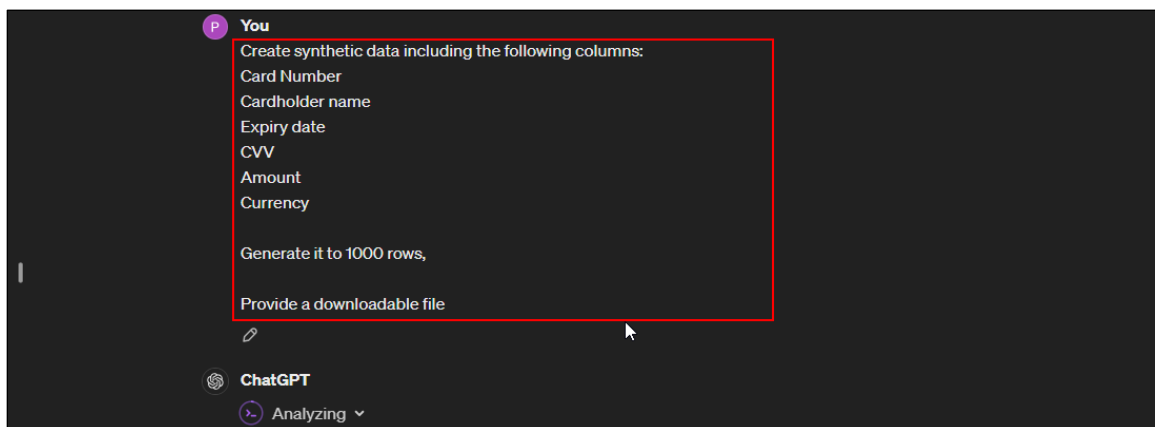
CVV

Amount

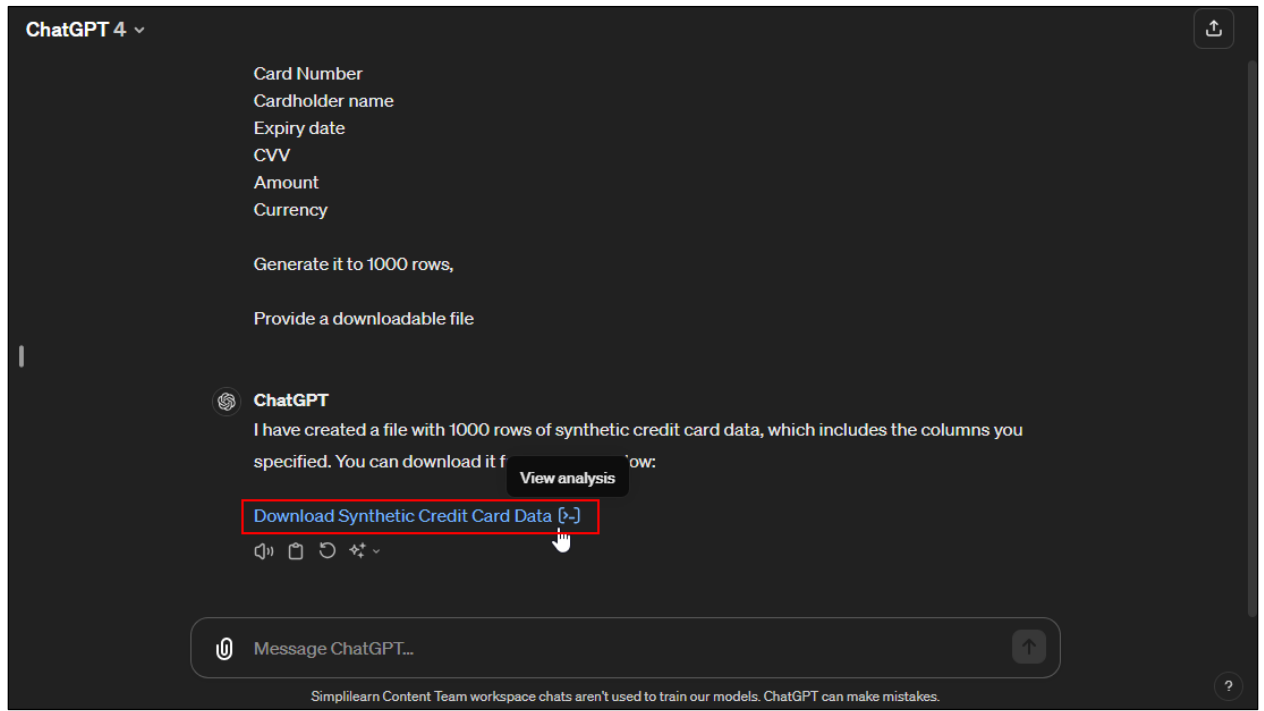
Currency

Generate it to 1000 rows,

Provide a downloadable file



2.3 Click on the **Download Synthetic Credit Card Data** link, as shown in the screenshot below:



The generated synthetic data is as shown below:

File Home Insert Page Layout Formulas Data Review View Automate Help

Clipboard Font Alignment Number Styles Cells Editing Add-ins

Apptos Narrow 11 A A

B I U Bold Italic Underline

General

Conditional Formatting Format as Table Cell Styles Insert Delete Format Cells

Sort & Filter Find & Select

Add-ins Analyze Data

L14

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Card Number	Cardholder Name	Expiry Date	CVV	Amount	Currency												
2	6.51771E+15	Robert Castaneda	Sep-33	804	1168.87	CAD												
3	3.44476E+14	Kimberly Maldonado	Nov-24	678	914.11	CAD												
4	4.63897E+12	Karen Smith	Sep-27	56	406.01	EUR												
5	4.26949E+12	Evelyn Alexander	Dec-28	731	1840.83	USD												
6	1.80048E+14	Sylvia Williams	Sep-27	140	2636.74	EUR												
7	3.74615E+14	Virginia Sanchez	Jul-25	483	2654.04	EUR												
8	4.3741E+15	Mrs. Renee Clarke MD	Oct-28	182	1923.12	CAD												
9	6.01192E+15	Sara McGrath	Mar-26	136	309.39	JPY												
10	3.4528E+14	Cindy Bishop	Nov-33	664	1810.91	CAD												
11	2.2669E+15	John Andrews	Mar-34	297	3190.83	GBP												
12	1.80088E+14	Jason Frye	Jul-25	322	2745.97	USD												
13	4.97088E+12	Brittney Jones	Dec-26	748	3167.06	CAD												
14	4.34366E+15	Lori Larson	Oct-24	436	717.99	CAD												
15	1.80048E+14	Dana White	Apr-34	64	1266.99	CAD												
16	2.13132E+14	Maurice Morgan	Jun-33	887	3949.08	CAD												
17	4.10998E+12	Hunter Gibson	Jun-26	448	1272.29	GBP												
18	4.9801E+15	Cheryl Gibson	Oct-25	439	3852.04	GBP												
19	3.82958E+13	Anna Williams	Jun-24	948	582.03	EUR												
20	4.48771E+18	Ashley Wheeler	Mar-30	496	198.8	CAD												
21	3.47058E+14	Jared Lawrence	Dec-32	326	4578.65	USD												
22																		

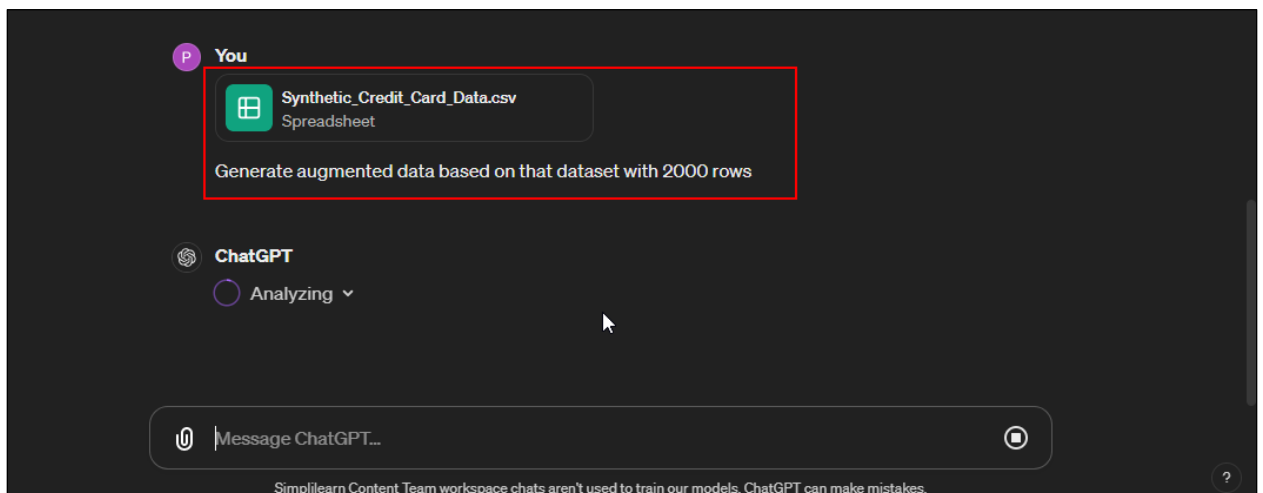
Synthetic_Credit_Card_Data

Ready Accessibility: Unavailable

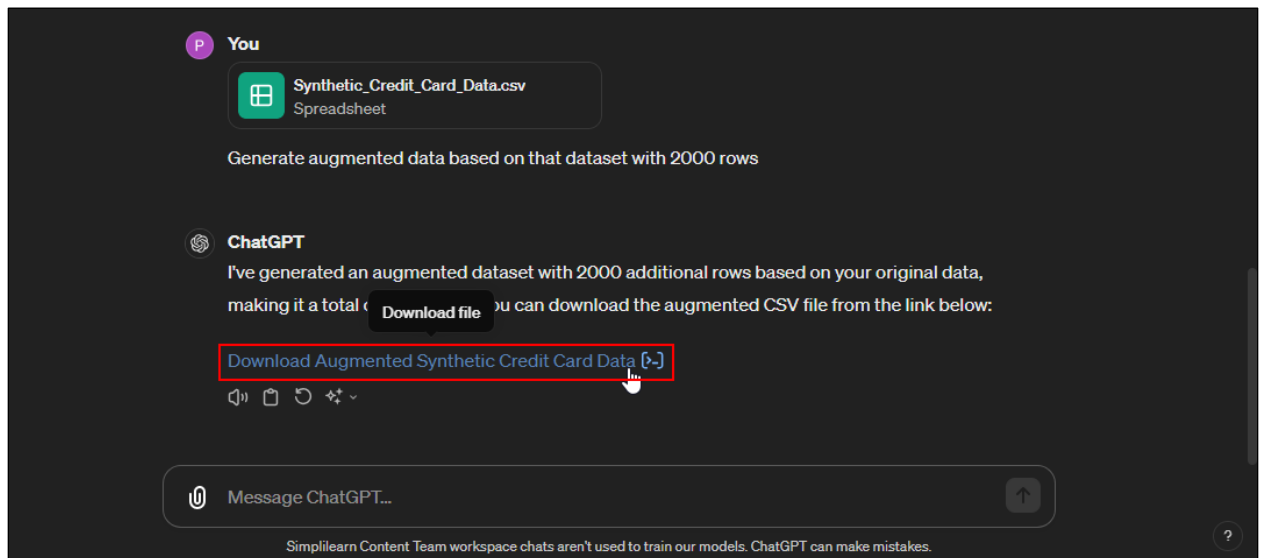
	A	B	C	D	E	F
989	0.40121E+14	Mrs. Ashley Gattiaguer MD	Oct-27	387	4137.79	CAD
990	1.80004E+14	Brendan West	May-28	800	2381.98	USD
991	4.92962E+15	James Castillo	Nov-31	3134	1724.13	GBP
992	6.01129E+15	Brianna Khan	Jun-33	244	2920.28	GBP
993	3.79551E+14	Gary Wiley	Jul-27	776	2877.49	JPY
994	3.78367E+14	James Santiago	Feb-26	325	2090.85	CAD
995	2.30697E+14	Jennifer Rodriguez	Aug-32	467	1258.65	GBP
996	4.26249E+12	Melinda Wells	May-29	927	4904.39	GBP
997	6.39082E+11	Natalie Heath	Oct-28	680	3296.22	JPY
998	4.727E+12	David Powell	Jul-27	832	2233.91	JPY
999	2.27236E+15	Kelli Silva	Feb-32	551	193.43	USD
1000	4.58762E+15	Ryan Barrera	Aug-29	17	3209.57	USD
1001	3.7643E+14	Kathleen Larsen				

2.4 Upload the synthetic dataset and generate the augmented dataset using the following prompt:

Generate augmented data based on that dataset with 2000 rows



2.5 Click on the **Download Augmented Synthetic Credit Card Data** link



The generated augmented data is as shown below:

The screenshot shows a Microsoft Excel spreadsheet titled 'Augmented_Synthetic_Credit_Card'. The spreadsheet contains 12 rows of data, each representing a credit card record. The columns are labeled A through R, with data primarily in columns A through F. The data includes card numbers, names, expiration dates, and monetary values in various currencies (GBP, JPY, USD, EUR). The status bar at the bottom indicates 'Ready', 'Accessibility: Unavailable', and summary statistics: 'Average: 1.30649E+15', 'Count: 6', 'Sum: 5.22596E+15'.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
992	6.01129E+15	Brianna Khan	Jun-33	3134	1724.13	GBP												
993	3.79551E+14	Gary Wiley	Jul-27	244	2920.28	GBP												
994	3.78367E+14	James Santiago	Feb-26	776	2877.49	JPY												
995	2.30697E+15	Jennifer Rodriguez	Oct-32	325	2090.85	CAD												
996	4.26249E+12	Melinda Wells	Aug-32	467	1258.65	GBP												
997	6.39082E+11	Natalie Heath	May-29	927	4904.39	GBP												
998	4.727E+12	David Powell	Oct-28	680	3296.22	JPY												
999	2.27236E+15	Kelli Silva	Jul-27	832	2233.91	JPY												
1000	4.58762E+15	Ryan Barrera	Feb-32	551	193.43	USD												
1001	3.7643E+14	Kathleen Larsen	Aug-29	17	3209.57	USD												
1002	5.22596E+15	Chris Green	Nov-32	13	4852.04	JPY												
1003	2.13188E+14	Kathleen Robertson	Jul-26	919	1175.12	JPY												
1004	4.85652E+15	Crystal Charles	Feb-34	658	2913.45	JPY												
1005	3.71493E+14	Shelby Holmes	Apr-32	358	4237.41	CAD												
1006	4.32471E+12	Morgan McIntosh	Sep-24	994	1941.48	GBP												
1007	5.03846E+11	Sabrina Bryant	Sep-26	242	4145.68	CAD												
1008	6.01168E+15	Mr. Frank Small	Aug-31	841	2270.45	CAD												
1009	4.80837E+15	Sharon Rodriguez	May-27	272	1091.05	CAD												
1010	3.71958E+14	Nicole Jenkins	Sep-25	381	3962.79	USD												
1011	2.26568E+15	Julian Hardy	Jul-27	3	4929.59	JPY												
1012	3.58558E+15	David Barnett	Aug-25	8391	1416.12	EUR												

By following these steps, you have successfully generated synthetic and augmented datasets using generative AI tools for validating data integrity, enhancing test coverage, and ensuring robustness during the testing phase of the Software Development Life Cycle.