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:

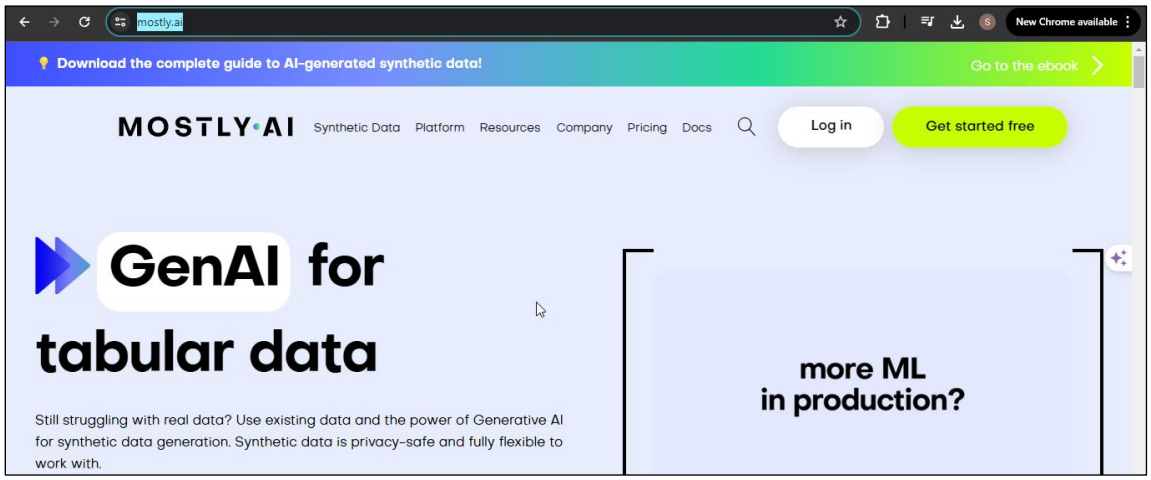
1. Generate the synthetic dataset using Mostly AI
2. Generate the synthetic and augmented dataset using ChatGPT

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.

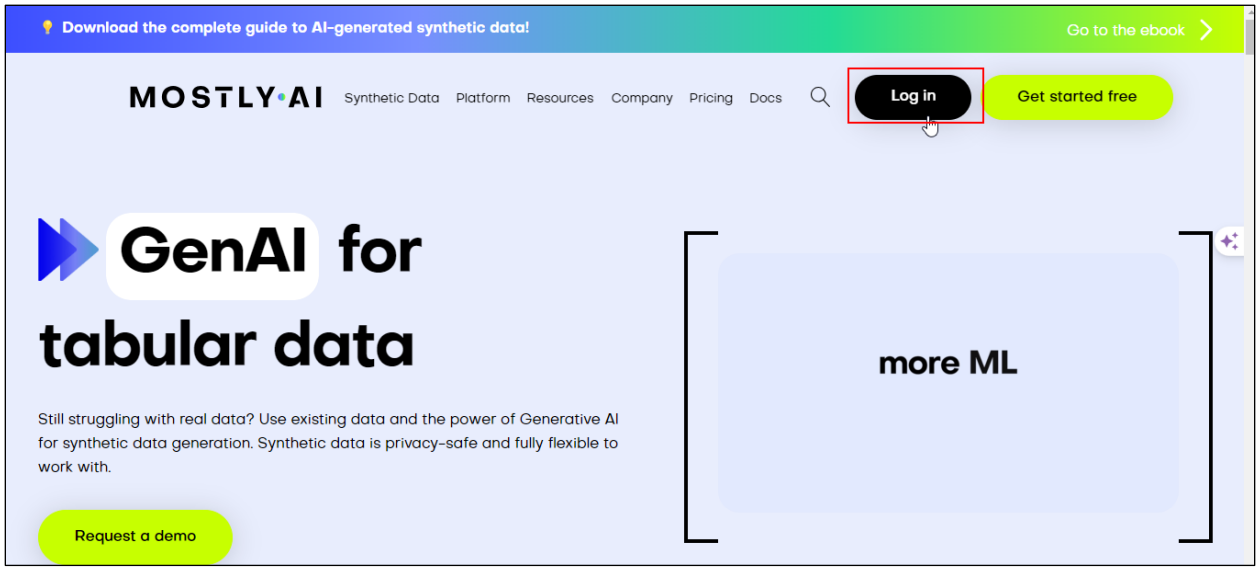
Note: Mostly AI may offer some customization options, primarily focusing on generating realistic synthetic data from scratch rather than augmenting existing datasets.
ChatGPT 4, on the other hand, is a versatile tool that can be used to generate synthetic data and manipulate and augment existing datasets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

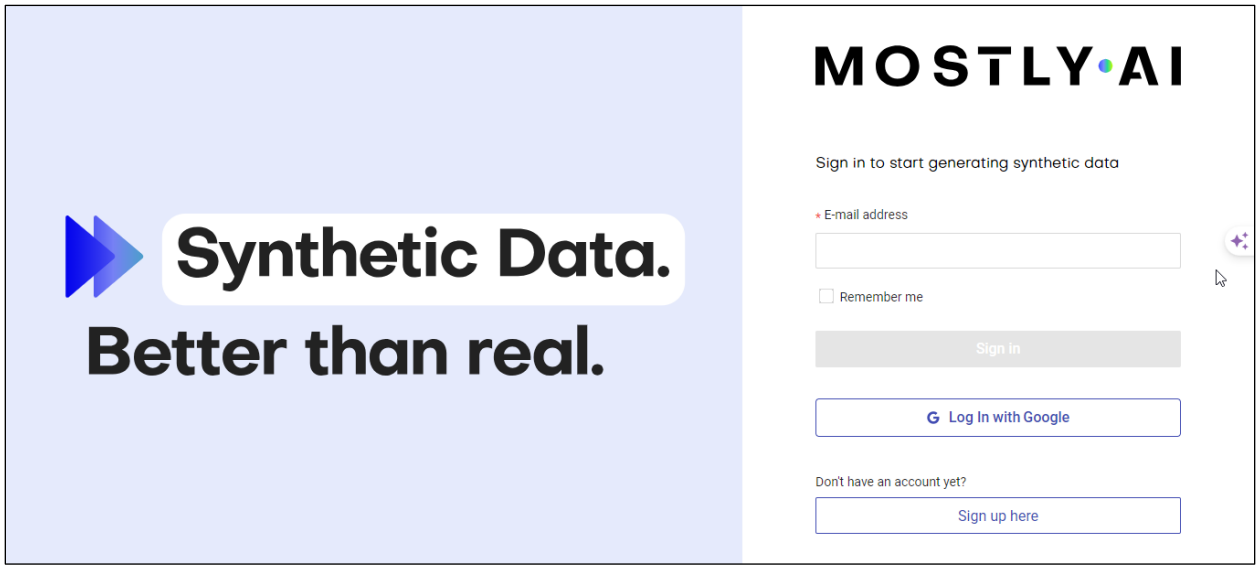
Step 1: Generate the synthetic dataset using Mostly AI

1.1 Navigate to **https://mostly.ai/**



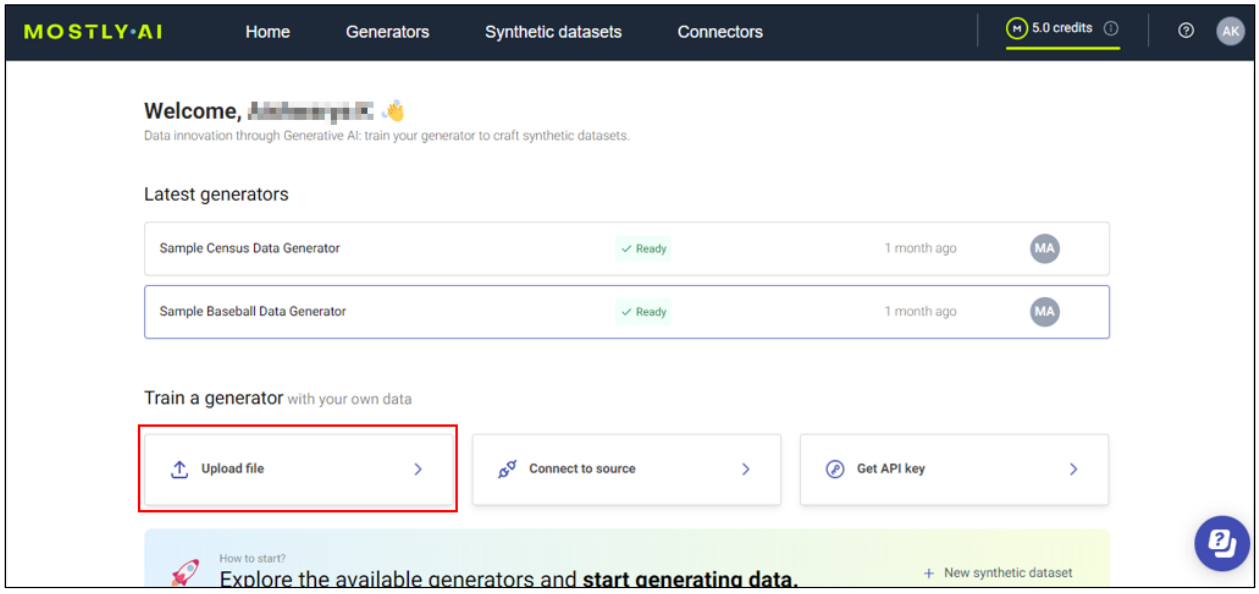
1.2 Click on the **Log in** button and enter the credentials

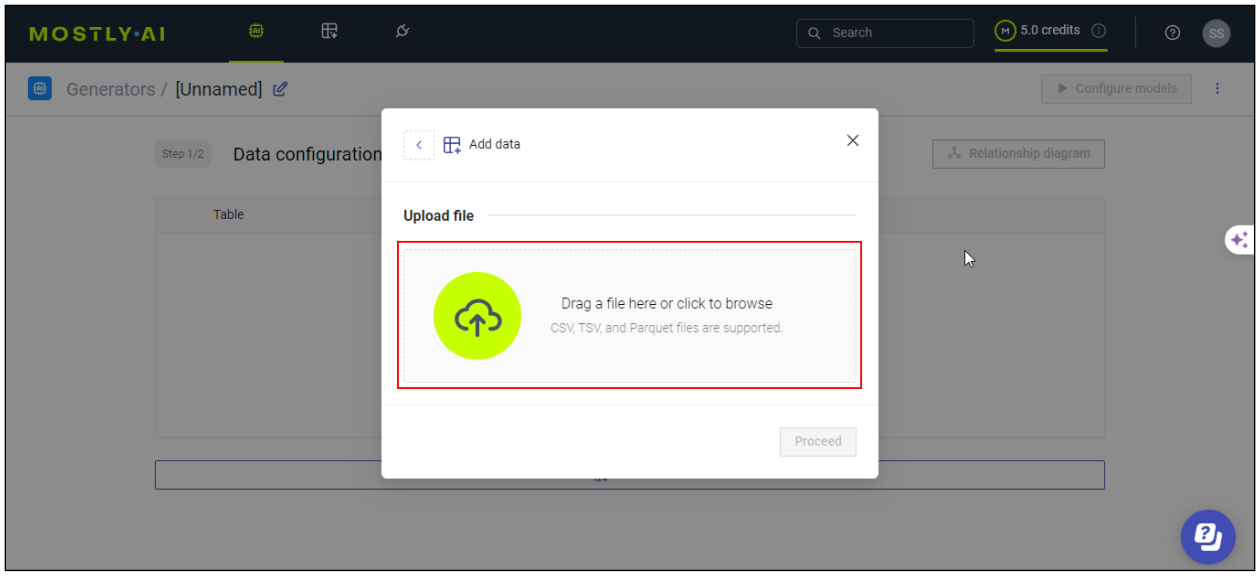




Note: Create a new account by clicking on the **Sign up here** button if you do not have an account

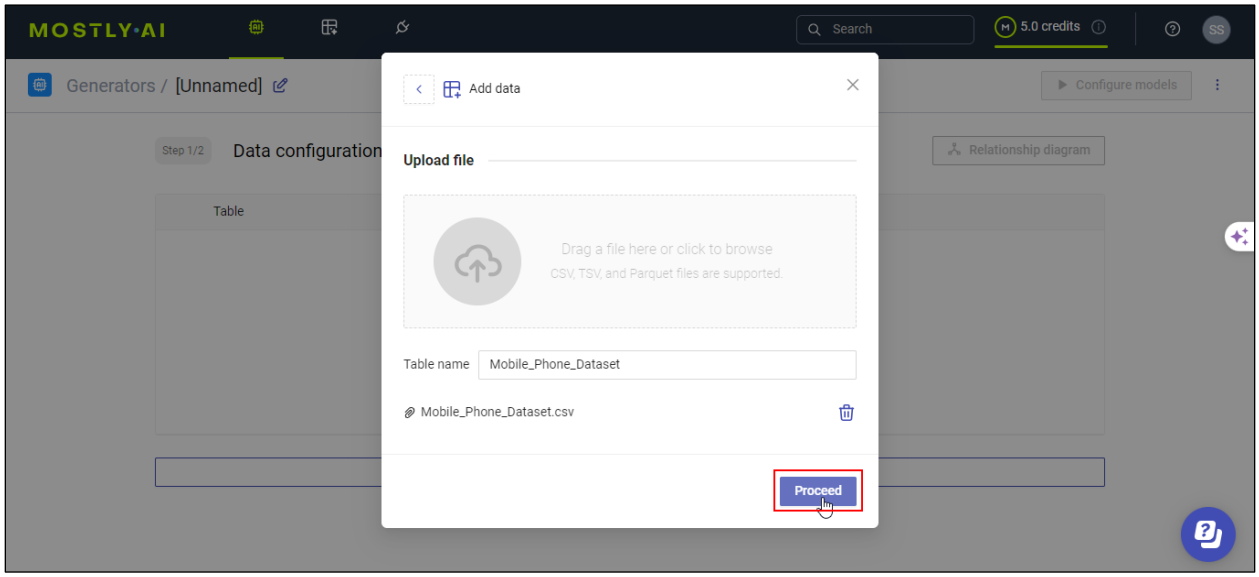
1.3 Click on the **Upload file** button



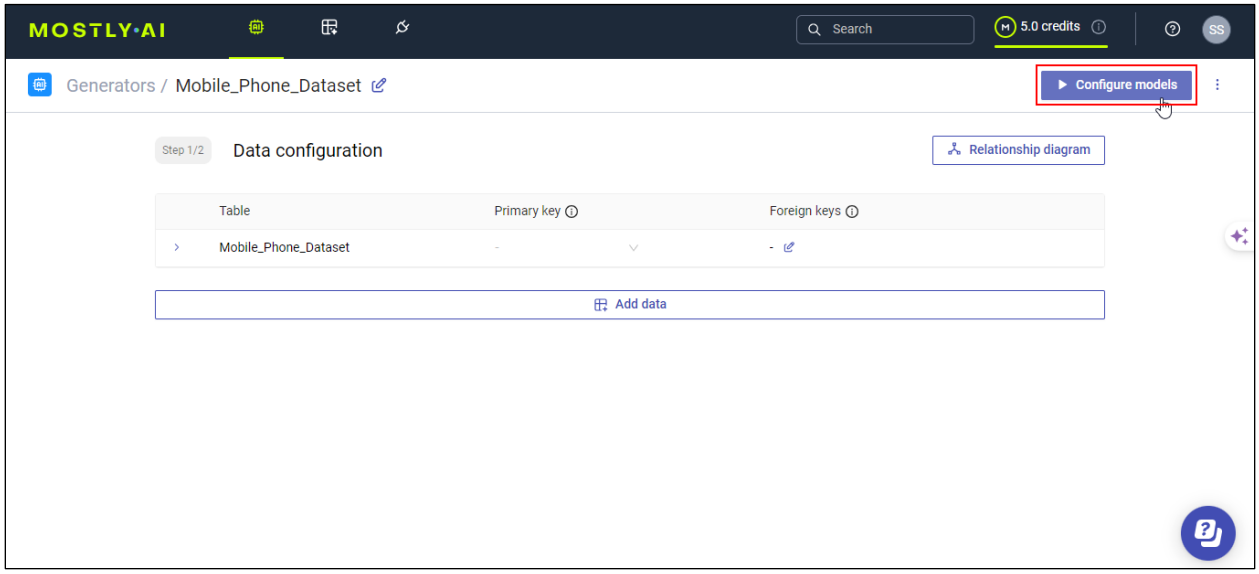


Note: You may use any sample data set; if you do not have any data set, use ChatGPT to create a demo dataset of the desired number of rows and columns.
Use the following prompt:
Create a dataset of 100 rows for a list of mobile phones, including year of launch, model, price, country, camera configuration,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

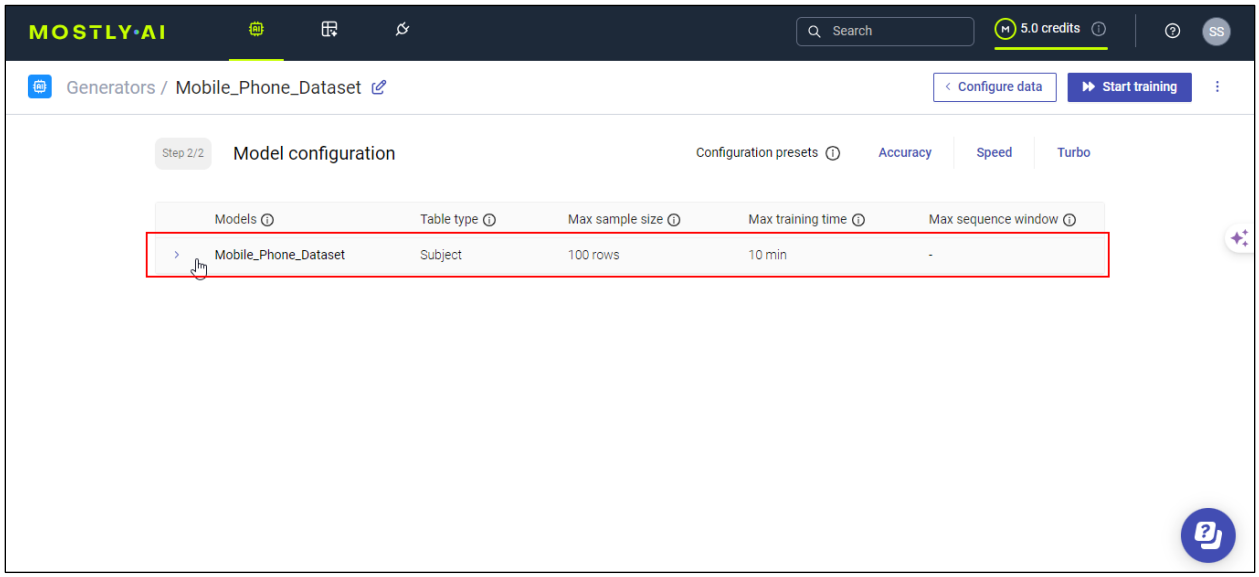
1.4 Click on the **Proceed** button

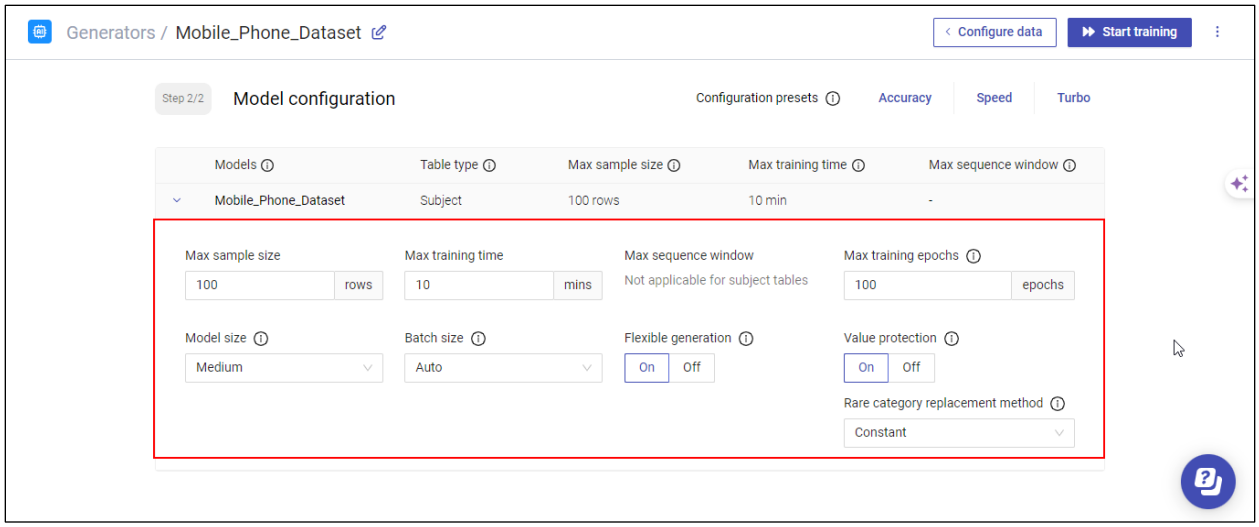


1.5 Click on the **Configure models** button, as shown in the screenshot below:



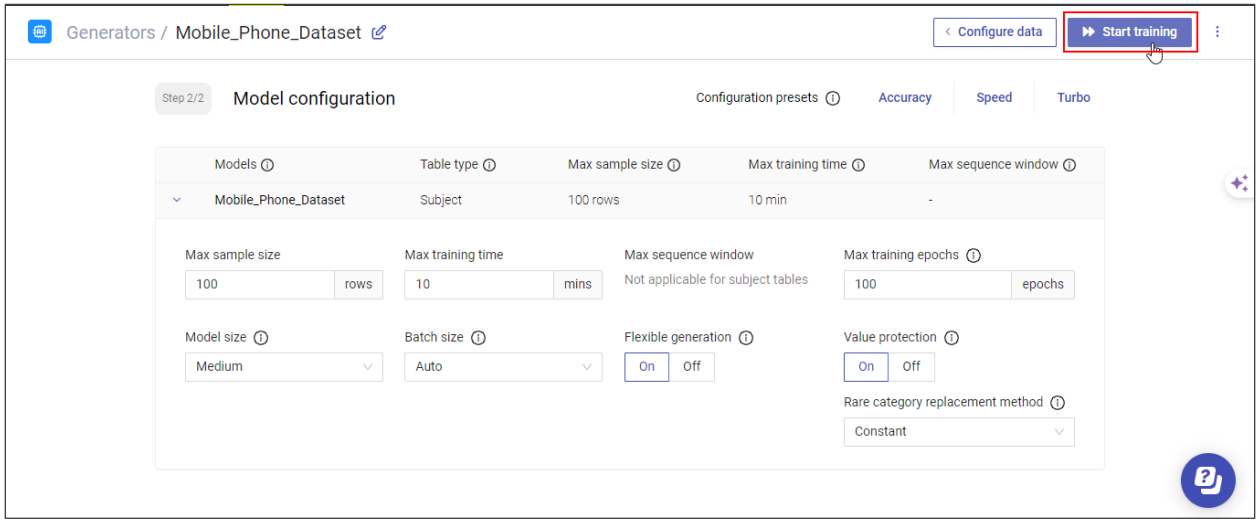
1.6 Click on the **Mobile_Phone_Dataset** list and make the desired changes accordingly, as shown in the screenshots below:



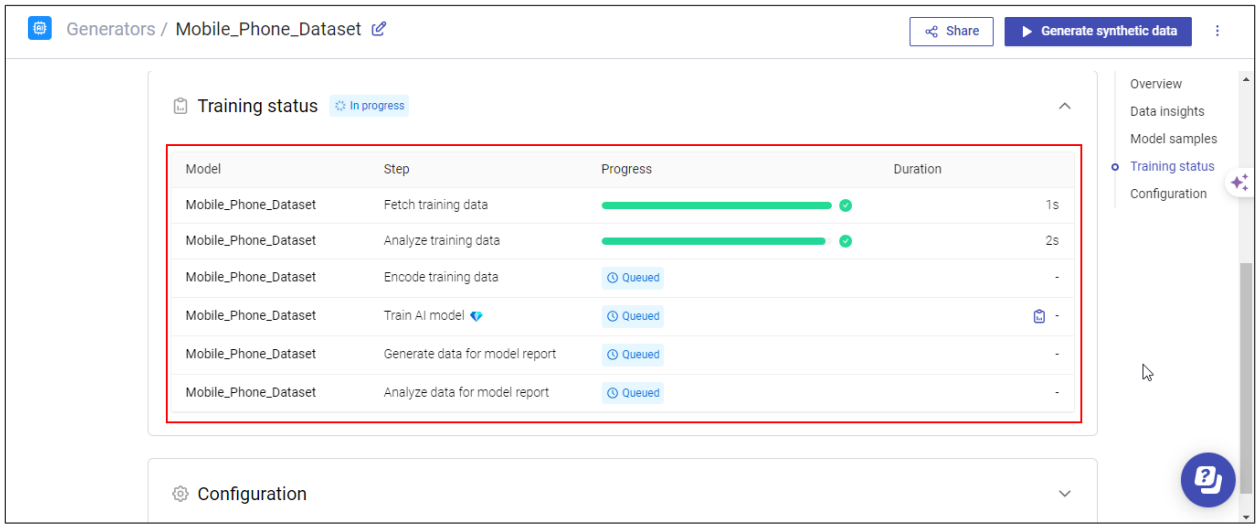


Note: Customize the dataset parameters such as column names and data distribution to match specific requirements

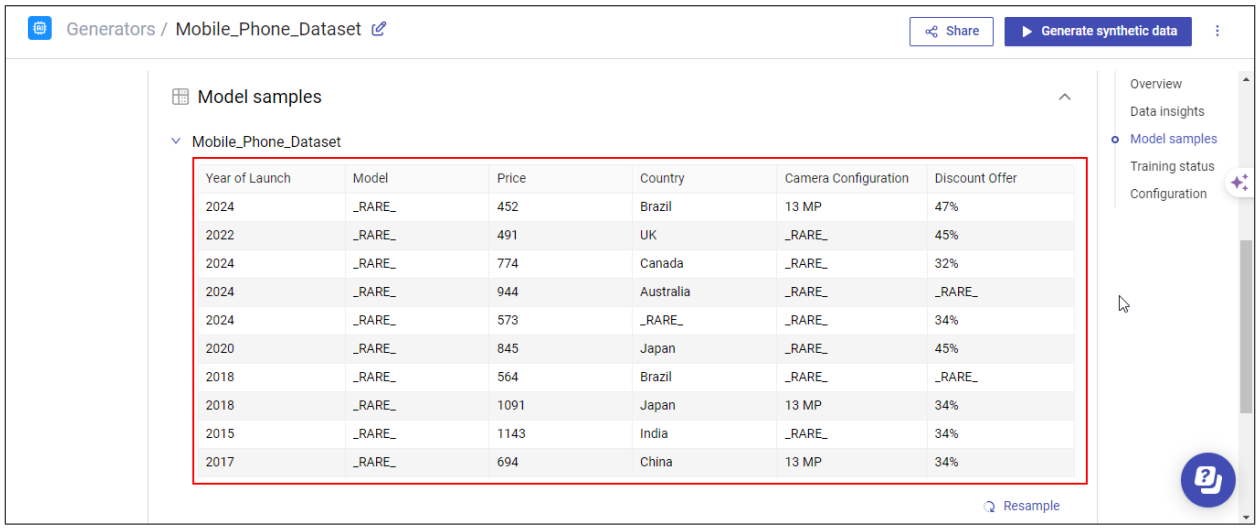
1.7 Click on the **Start training** button



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



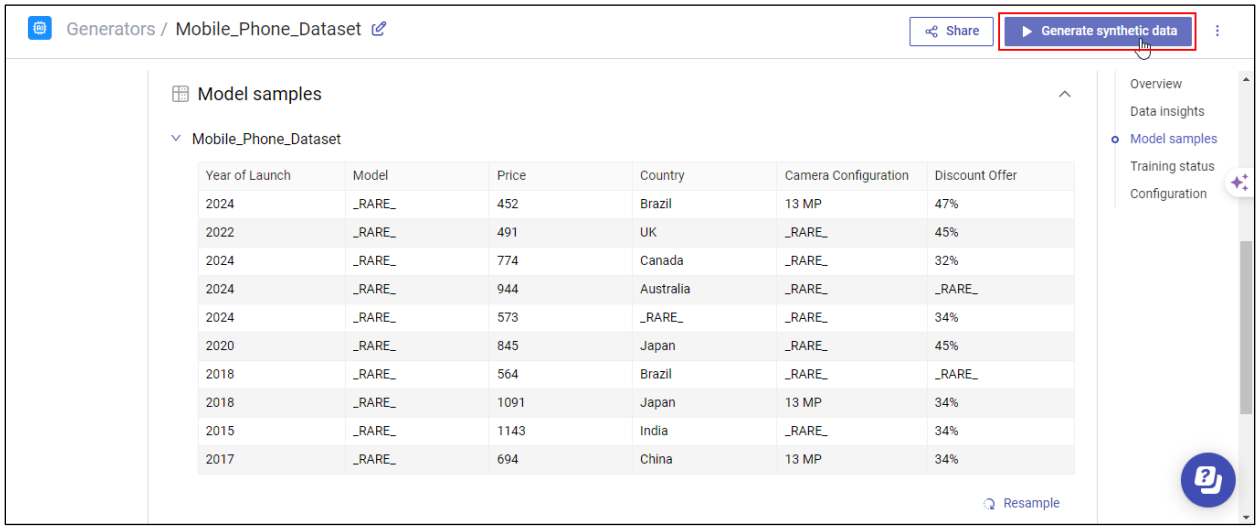
1.9 Examine the generated synthetic data in the **Model samples** section to ensure it meets your requirements



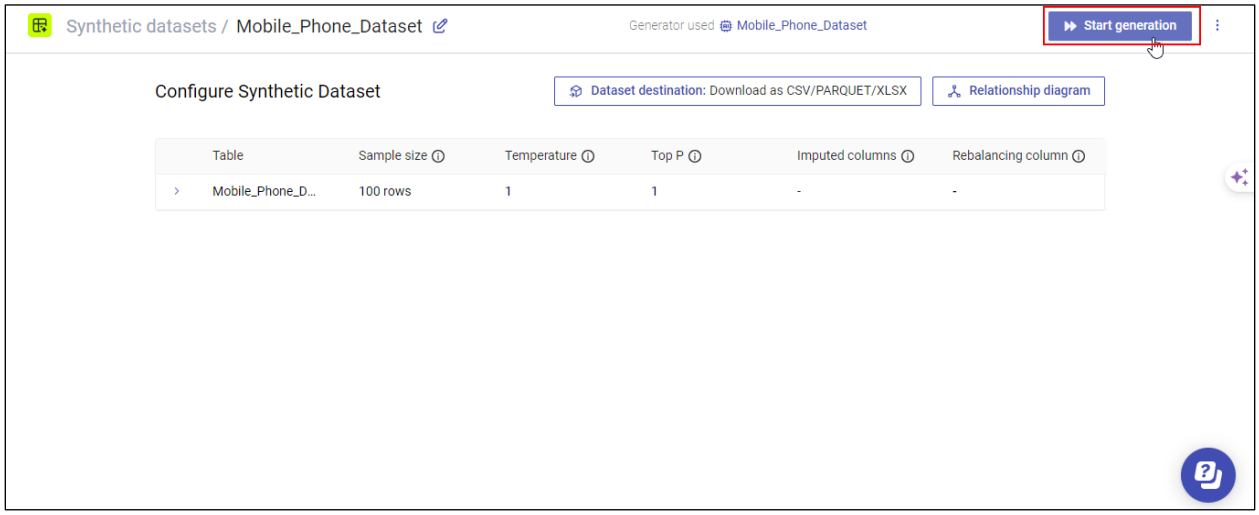
Year of Launch	Model	Price	Country	Camera Configuration	Discount Offer
2024	_RARE_	452	Brazil	13 MP	47%
2022	_RARE_	491	UK	_RARE_	45%
2024	_RARE_	774	Canada	_RARE_	32%
2024	_RARE_	944	Australia	_RARE_	_RARE_
2024	_RARE_	573	_RARE_	_RARE_	34%
2020	_RARE_	845	Japan	_RARE_	45%
2018	_RARE_	564	Brazil	_RARE_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

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

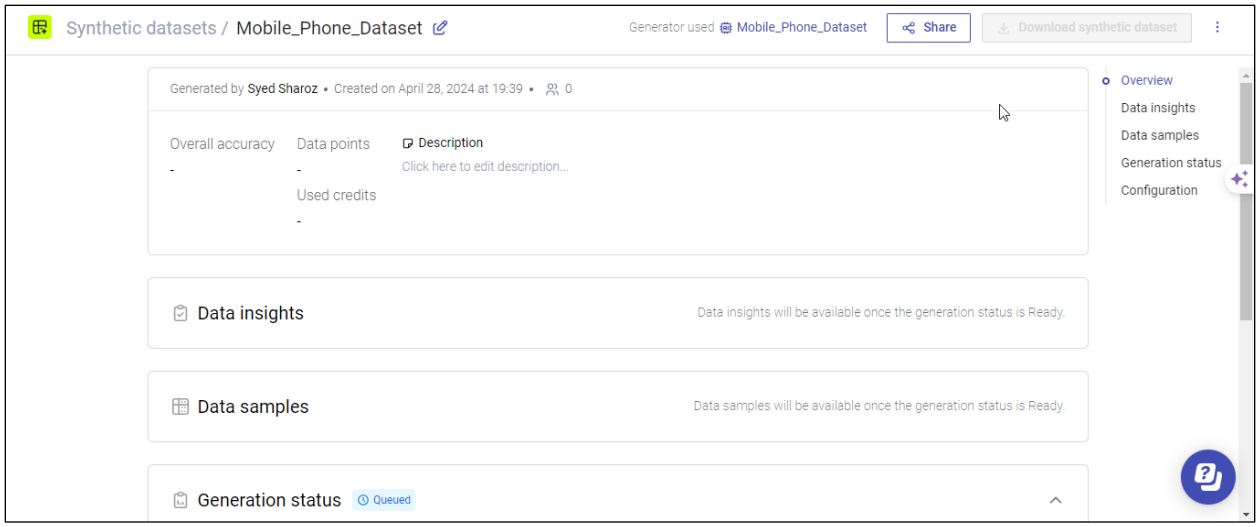
1.10 Click on the **Generate synthetic data** button



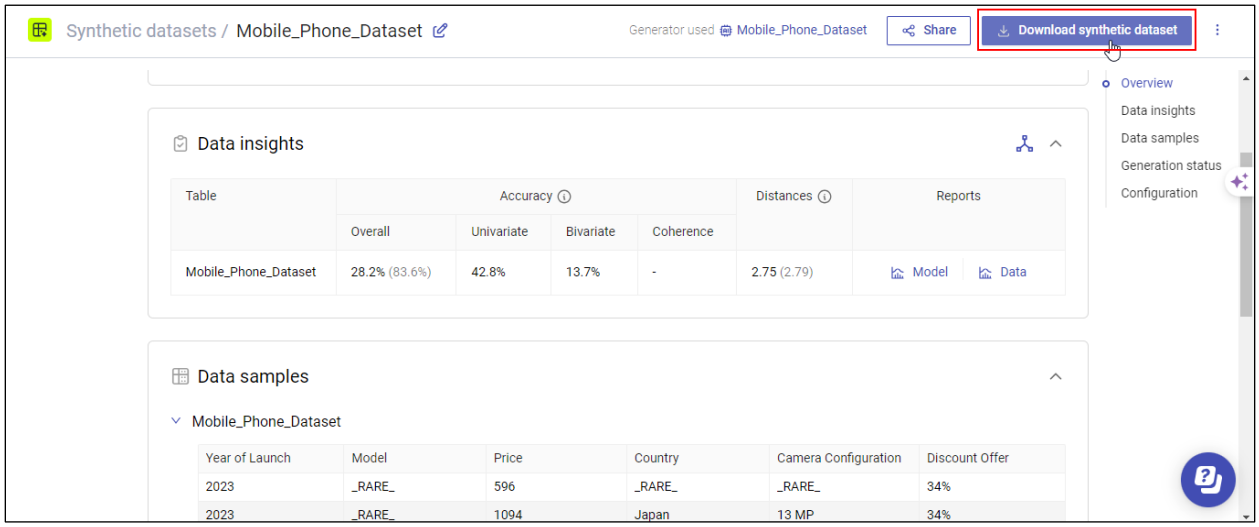
1.11 Click on the **Start generation** button

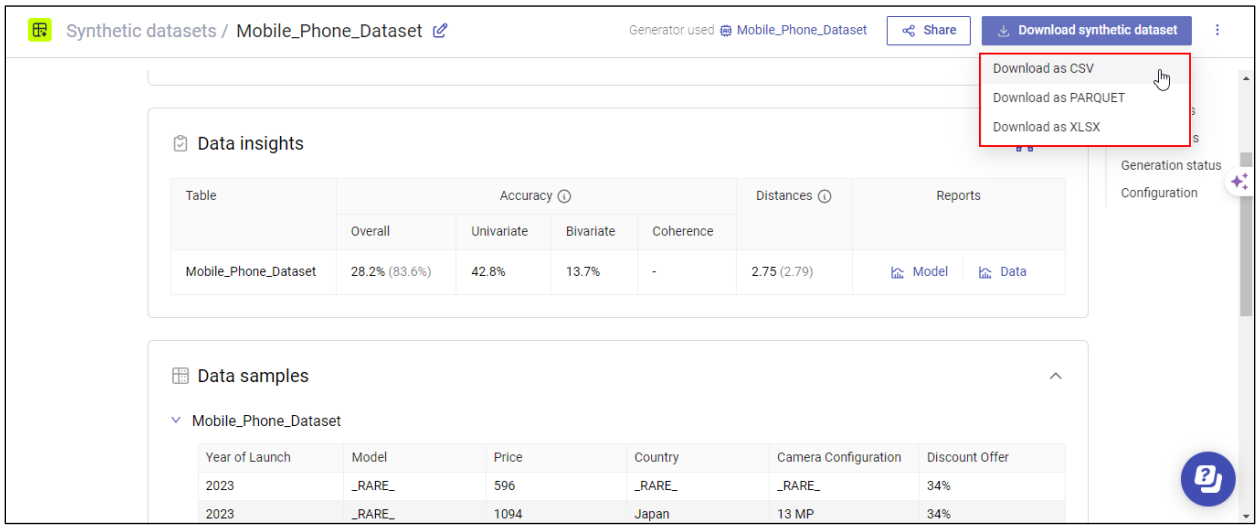


You will see the following interface:



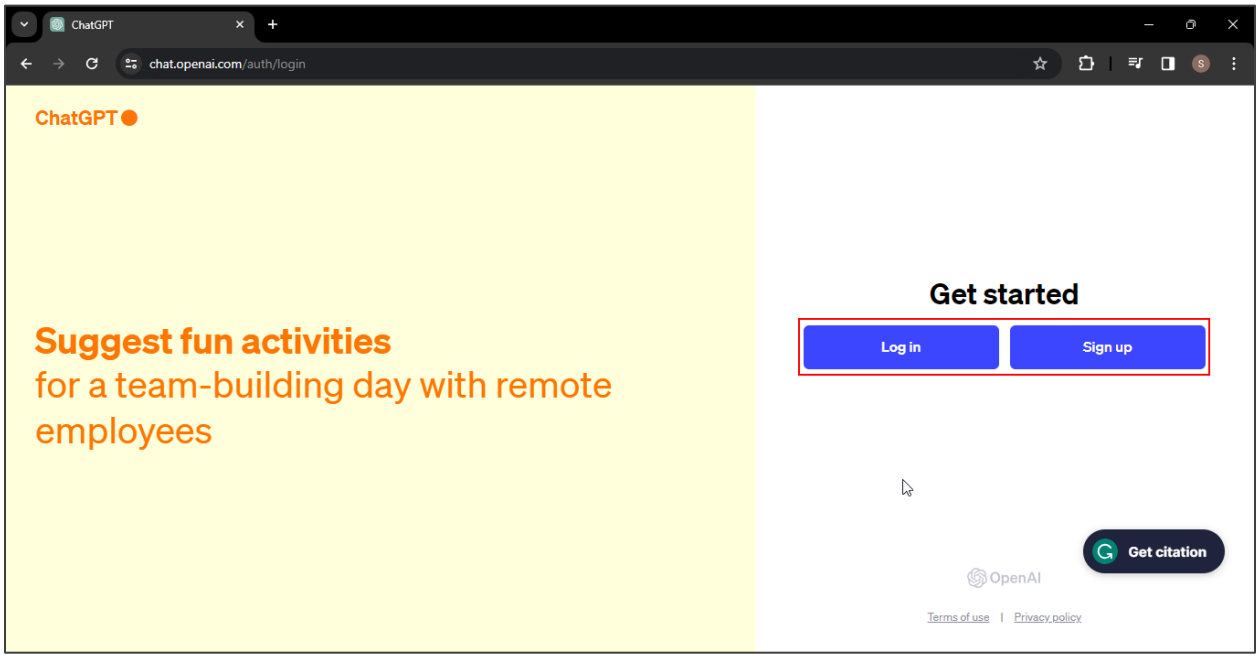
1.12 Click on the **Download synthetic dataset** dropdown and select the desired option to download or directly export the synthetic dataset, as shown in the screenshots below:



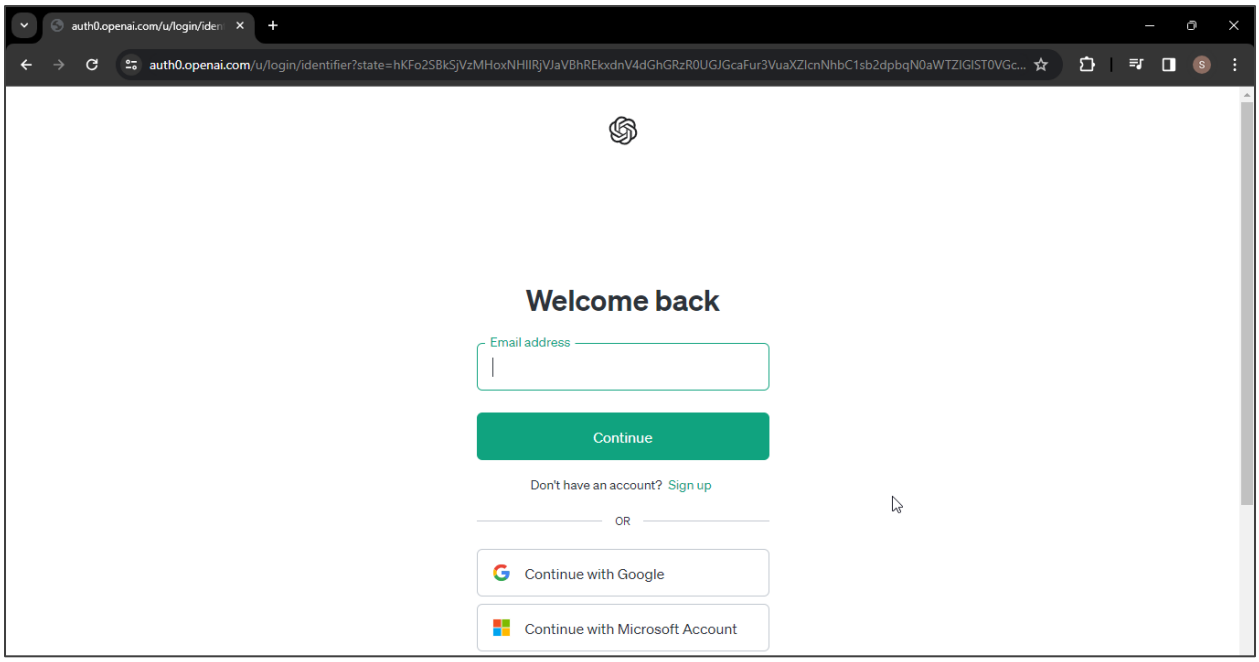


Step 2: Generate the synthetic and augmented dataset using ChatGPT

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



Note: Sign up if you do not have an account

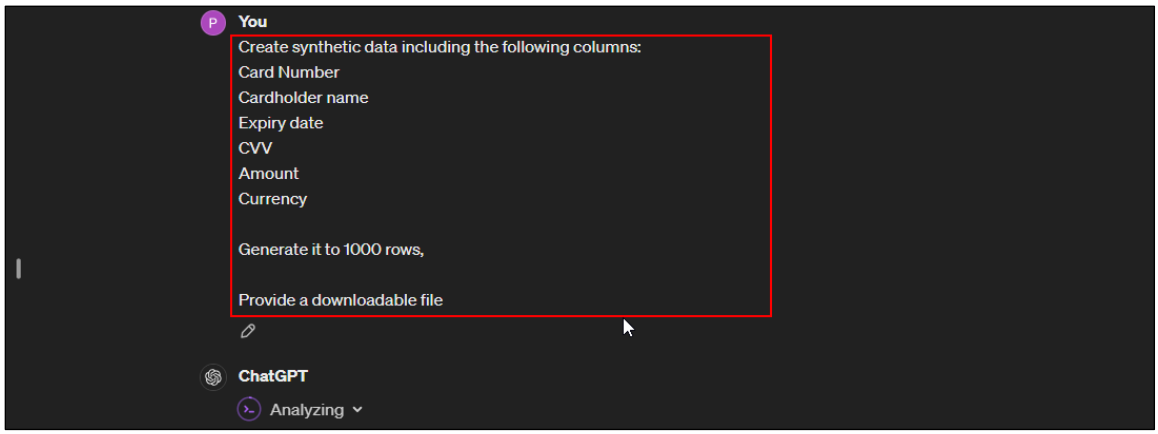


2.2 Generate the synthetic data using the following prompt:

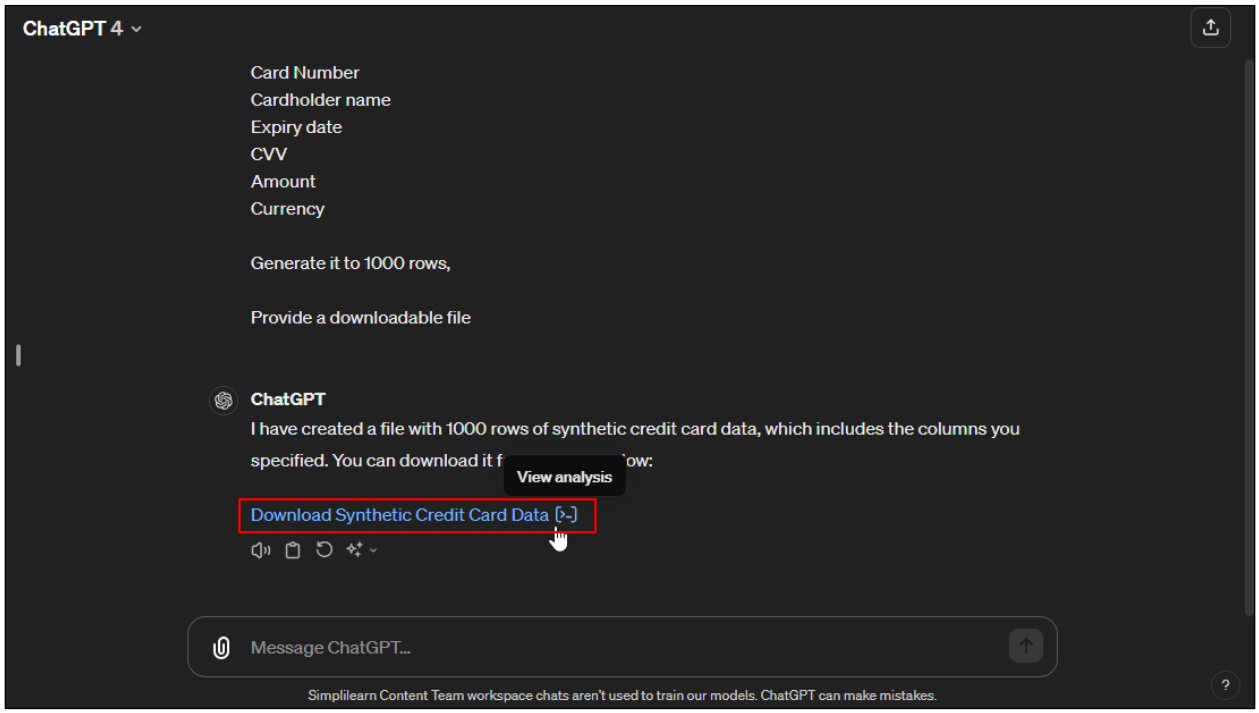
Create synthetic data including the following columns:
Card Number
Cardholder name
Expiry date
CVV
Amount
Currency

Generate it to 1000 rows,

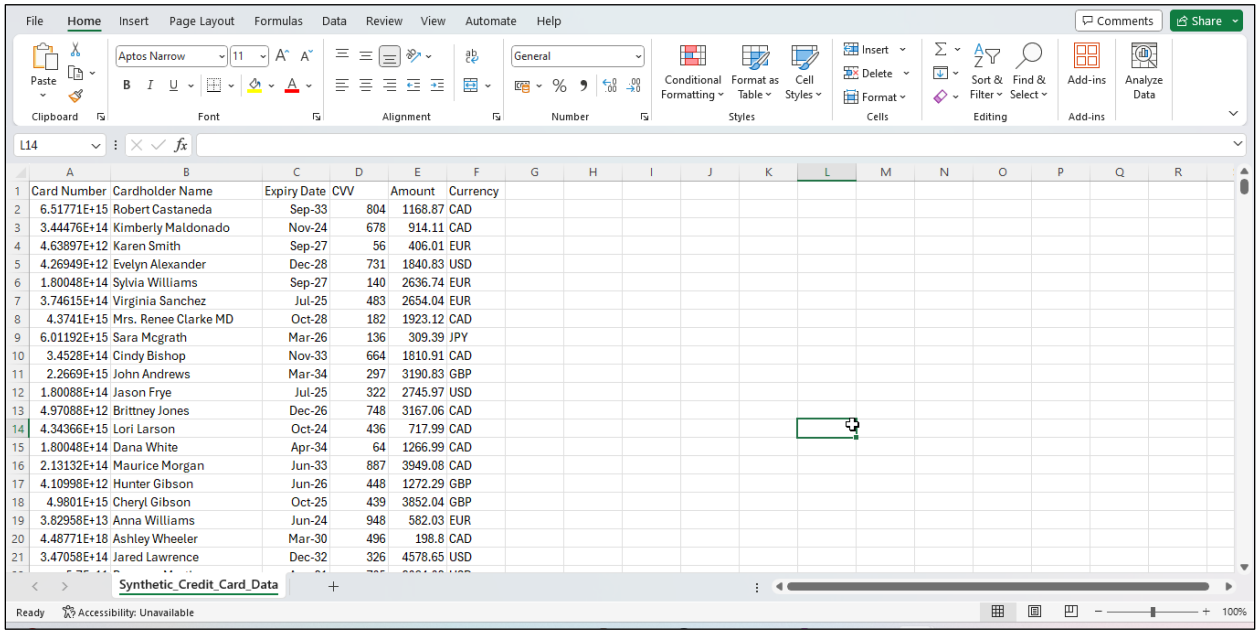
Provide a downloadable file



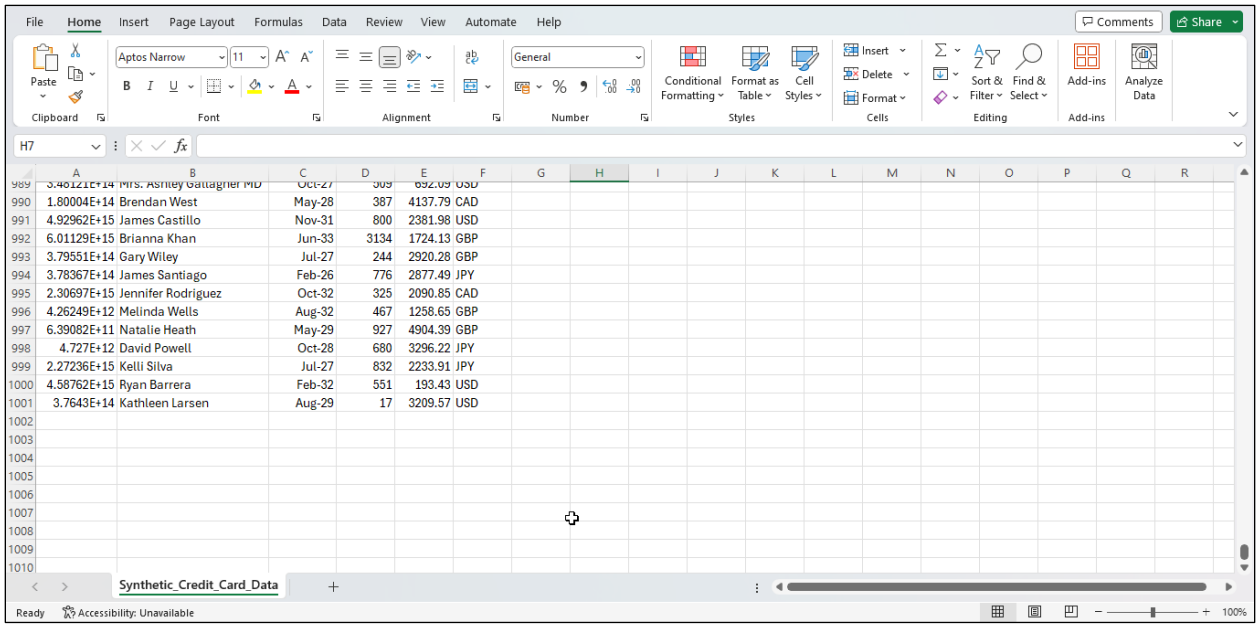
2.3 Click on the **Download Synthetic Credit Card Data** link, as shown in the screenshot below:



The generated synthetic data is as shown below:

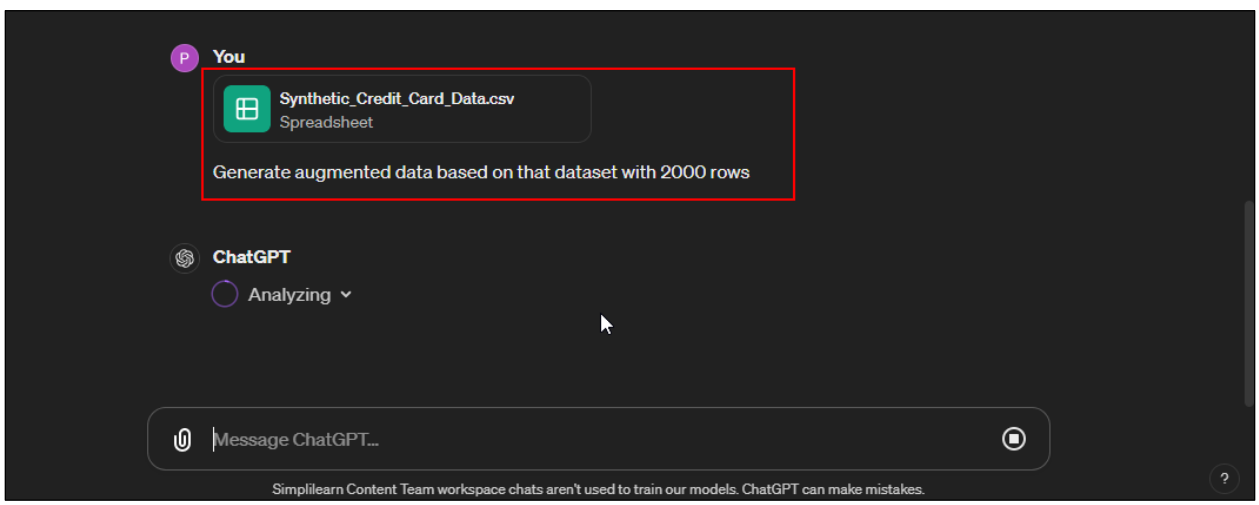


Card Number	Cardholder Name	Expiry Date	CVV	Amount	Currency
6.51771E+15	Robert Castaneda	Sep-33	804	1168.87	CAD
3.44476E+14	Kimberly Maldonado	Nov-24	678	914.11	CAD
4.63897E+12	Karen Smith	Sep-27	56	406.01	EUR
4.26949E+12	Evelyn Alexander	Dec-28	731	1840.83	USD
1.80048E+14	Sylvia Williams	Sep-27	140	2636.74	EUR
3.74615E+14	Virginia Sanchez	Jul-25	483	2654.04	EUR
4.3741E+15	Mrs. Renee Clarke MD	Oct-28	182	1923.12	CAD
6.01192E+15	Sara Mcgrath	Mar-26	136	309.39	JPY
3.4528E+14	Cindy Bishop	Nov-33	664	1810.91	CAD
2.2669E+15	John Andrews	Mar-34	297	3190.83	GBP
1.80088E+14	Jason Frye	Jul-25	322	2745.97	USD
4.97088E+12	Brittney Jones	Dec-26	748	3167.06	CAD
4.34366E+15	Lori Larson	Oct-24	436	717.99	CAD
1.80048E+14	Dana White	Apr-34	64	1266.99	CAD
2.13132E+14	Maurice Morgan	Jun-33	887	3949.08	CAD
4.10998E+12	Hunter Gibson	Jun-26	448	1272.29	GBP
4.9801E+15	Cheryl Gibson	Oct-25	439	3852.04	GBP
3.82958E+13	Anna Williams	Jun-24	948	582.03	EUR
4.48771E+18	Ashley Wheeler	Mar-30	496	198.8	CAD
3.47058E+14	Jared Lawrence	Dec-32	326	4578.65	USD

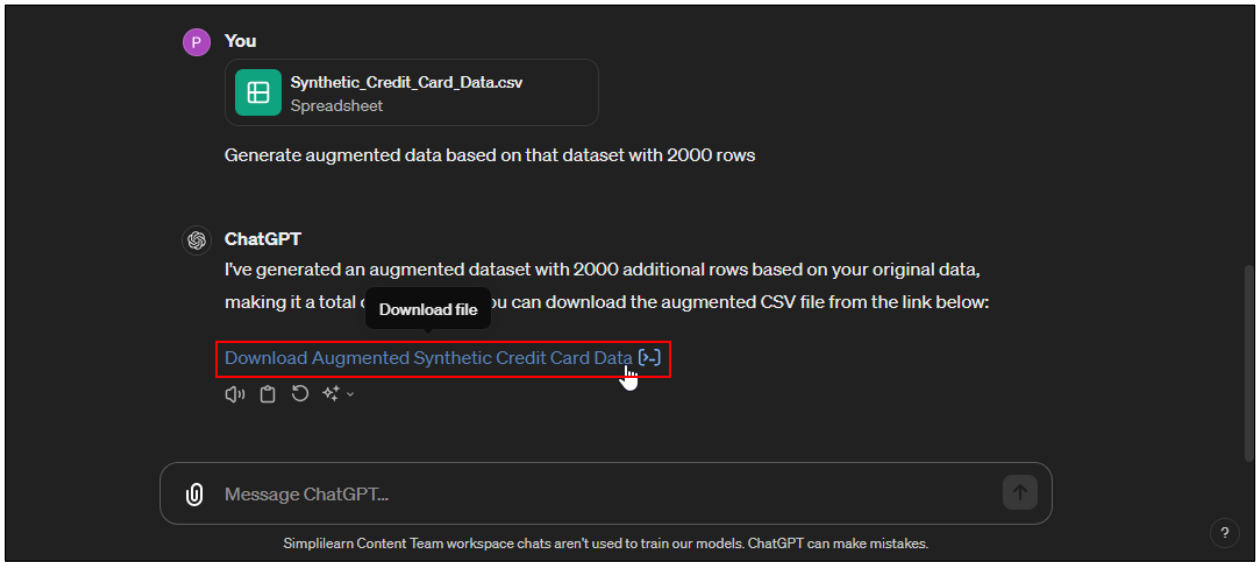


989	3.46121E+14	Mrs. Ashley Gallagher MD	Oct-27	509	692.09 USD
990	1.80004E+14	Brendan West	May-28	387	4137.79 CAD
991	4.92962E+15	James Castillo	Nov-31	800	2381.98 USD
992	6.01129E+15	Brianna Khan	Jun-33	3134	1724.13 GBP
993	3.79551E+14	Gary Wiley	Jul-27	244	2920.28 GBP
994	3.78367E+14	James Santiago	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

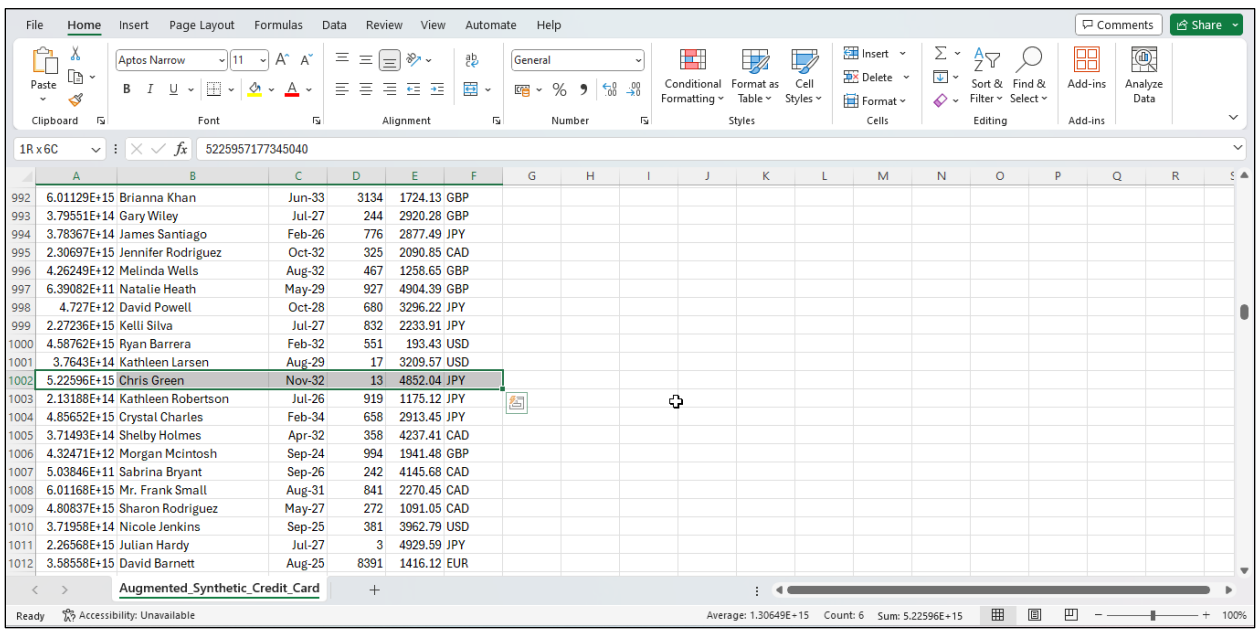
2.4 Upload the synthetic dataset and generate the augmented dataset using the following prompt:
Generate augmented data based on that dataset with 2000 rows



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



The generated augmented data is as shown below:



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.