PERCONA

AI for SQL and Data Analytics

Current State and the Future

Peter Zaitsev,
Founder at Percona
21 Feb, 2024

ConFoo.CA
DEVELOPER CONFERENCE

Who are You ?

- **Application Developer**
- **Data Scientist, Business User ?**
- **AI Expert**
- **Database Operations Expert**

© 2024 Percona

AI Basics

Training – Learning on Existing data

Inference – Using that learning on new data

Feedback Loop to improve

© 2024 Percona

Classical Inference

Classifying

Taking Action

Transformation

Prediction

© 2024 Percona

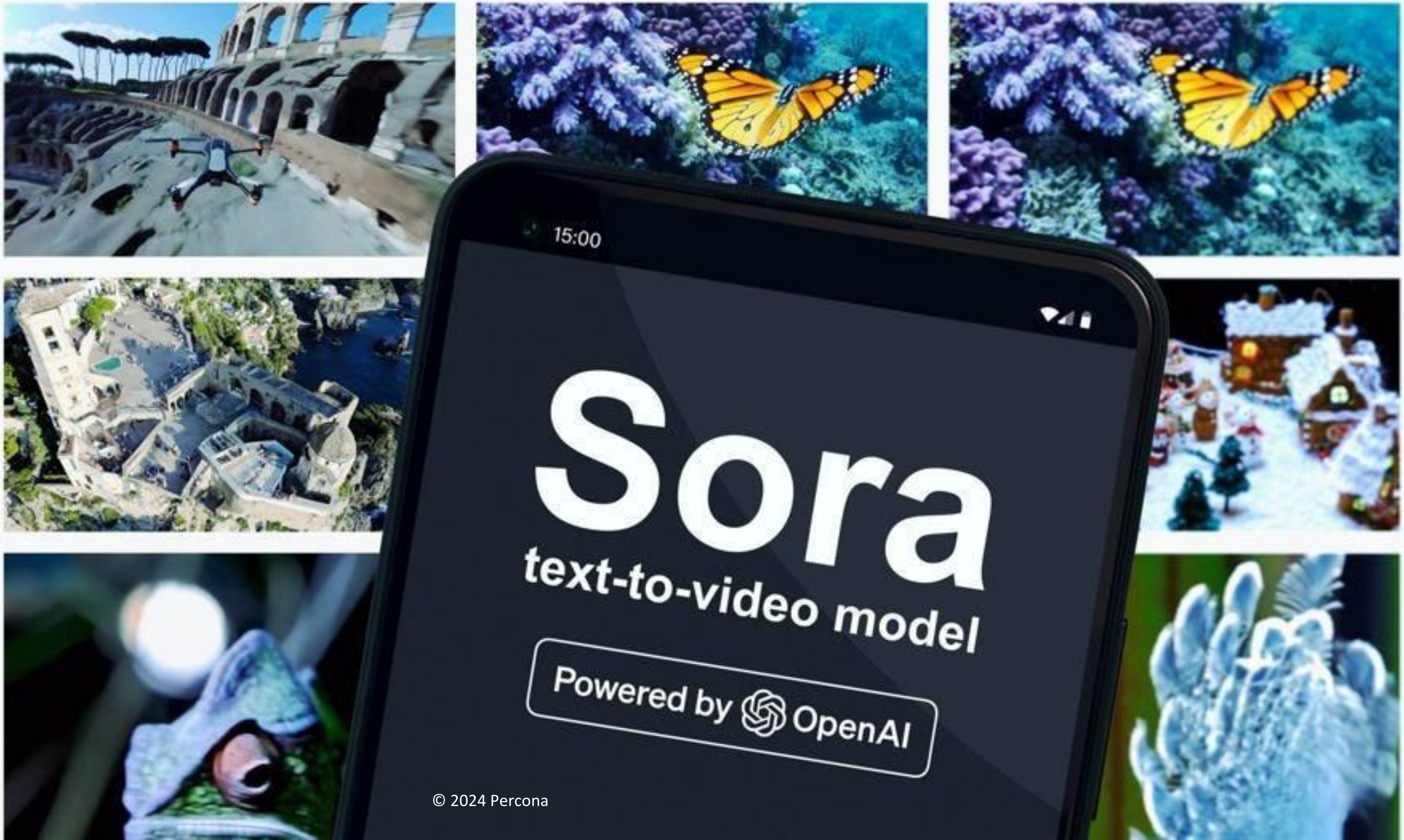
ChatGPT

Bard

Generative AI

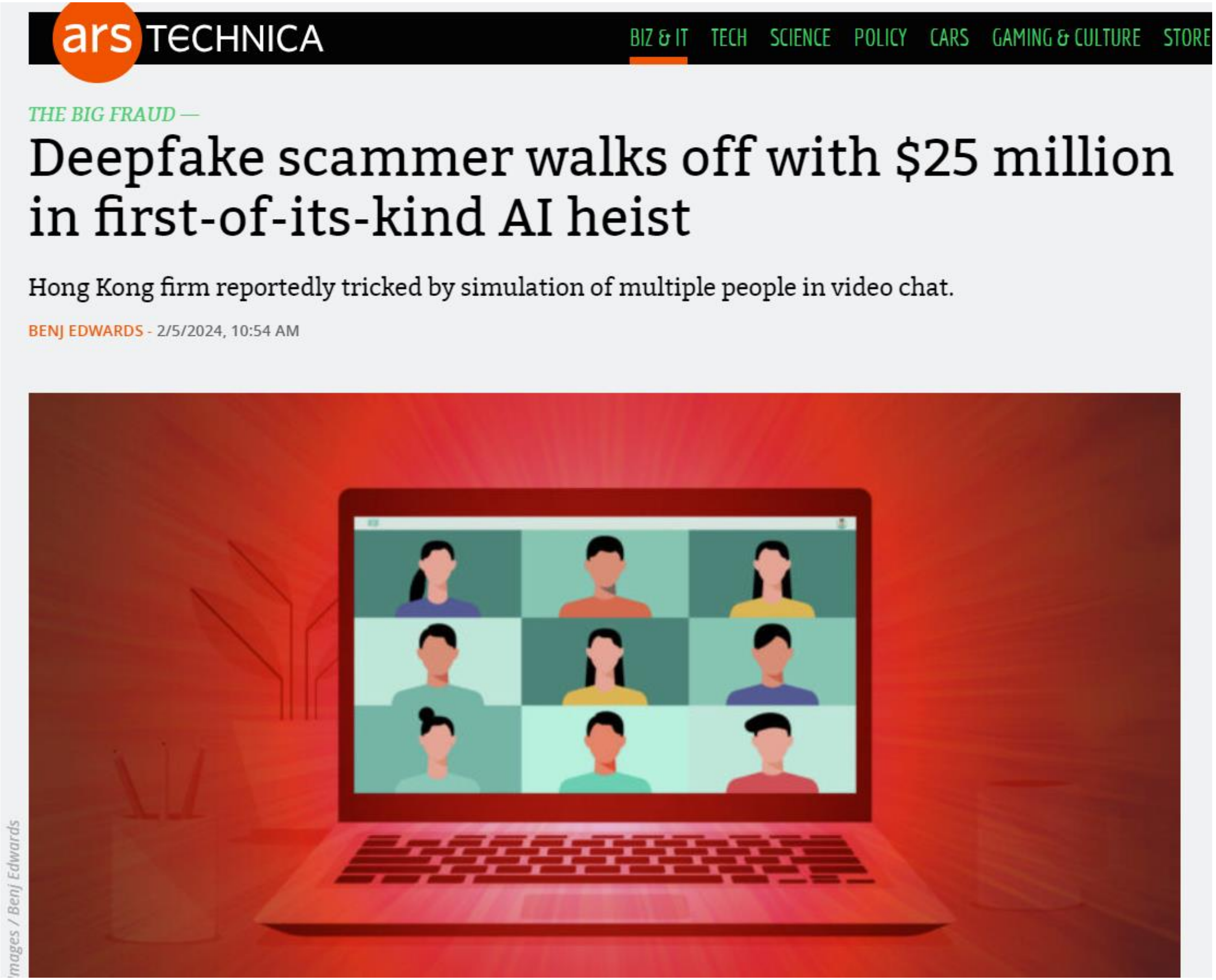
Generate content (text, image, video) based on human language prompt

© 2024 Percona



© 2024 Percona

For Good and for Evil



THE BIG FRAUD —

Deepfake scammer walks off with $25 million in first-of-its-kind AI heist

Hong Kong firm reportedly tricked by simulation of multiple people in video chat.

BENJ EDWARDS - 2/5/2024, 10:54 AM

Prompt as API

“Natural” API allows to start using system with very little training

© 2024 Percona

Be Warned: Extremely fast pace of Innovation

© 2024 Percona

AI in Data Industry

- **Business User Focus**
- **Data Analyst Focus**
- **Developer Focus**
- **Operator Focus**

© 2024 Percona

Tools

General Purpose Tools

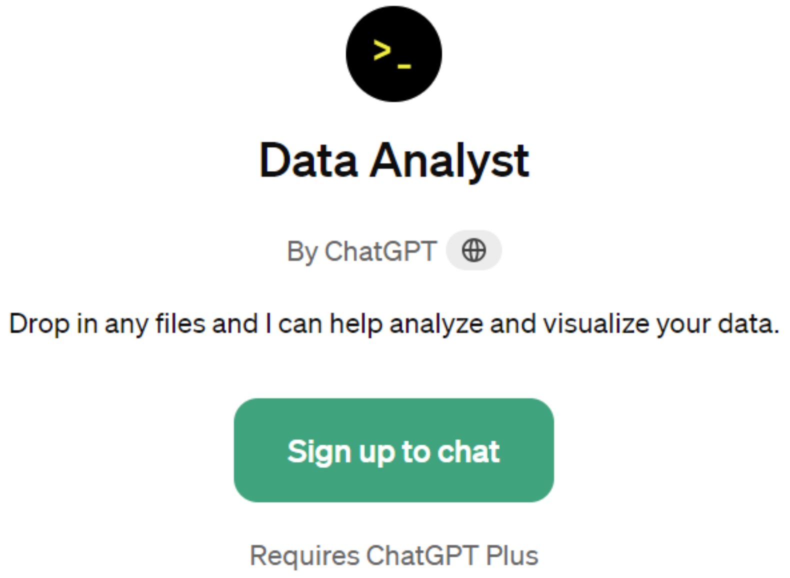
Specialized Tools

© 2024 Percona

Business User Focus

© 2024 Percona

AI Analyses



Data Analyst

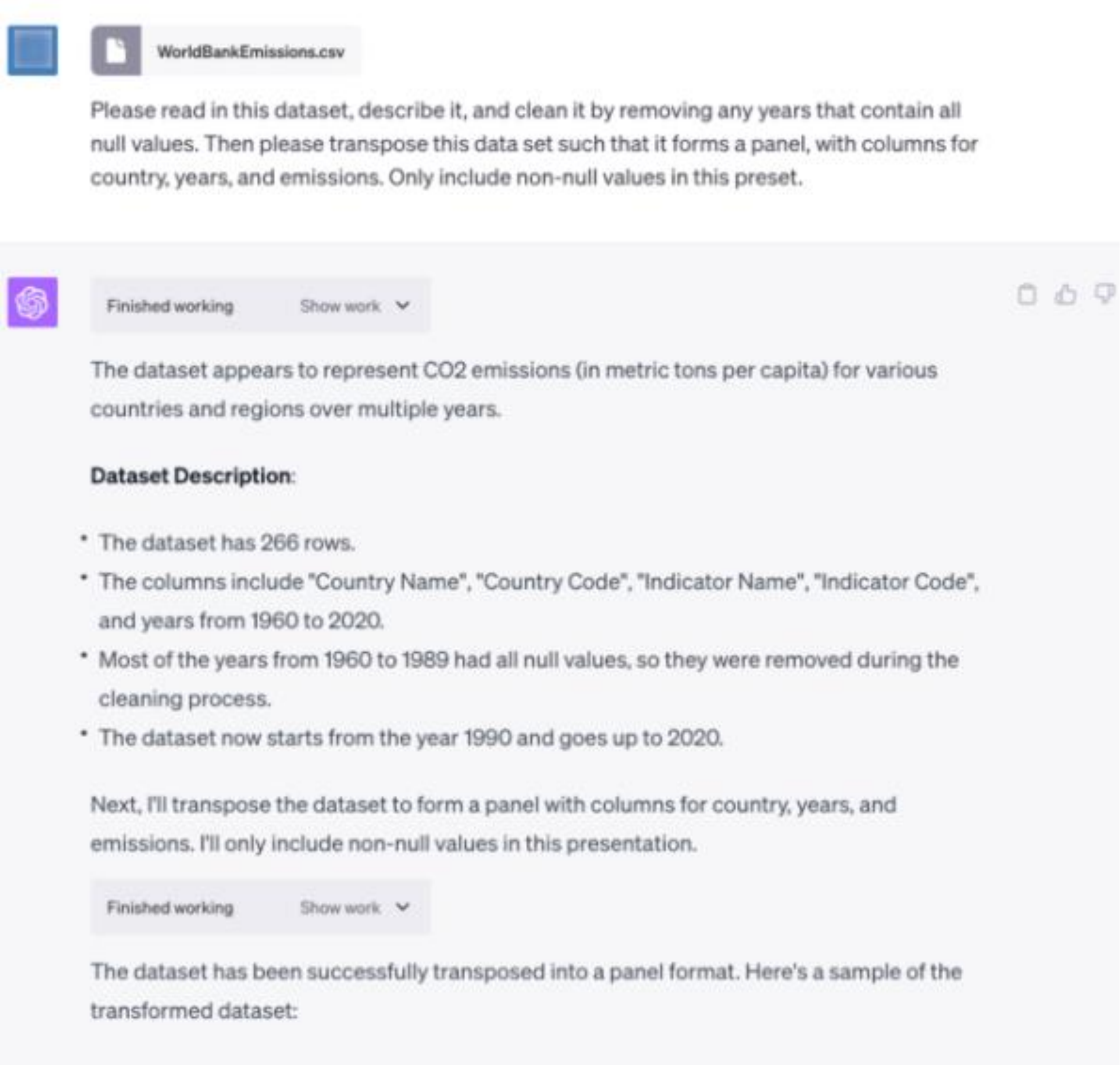
By ChatGPT

Drop in any files and I can help analyze and visualize your data.

Requires ChatGPT Plus

© 2024 Percona

Example



Please read in this dataset, describe it, and clean it by removing any years that contain all null values. Then please transpose this data set such that it forms a panel, with columns for country, years, and emissions. Only include non-null values in this preset.

The dataset appears to represent CO2 emissions (in metric tons per capita) for various countries and regions over multiple years.

Dataset Description:

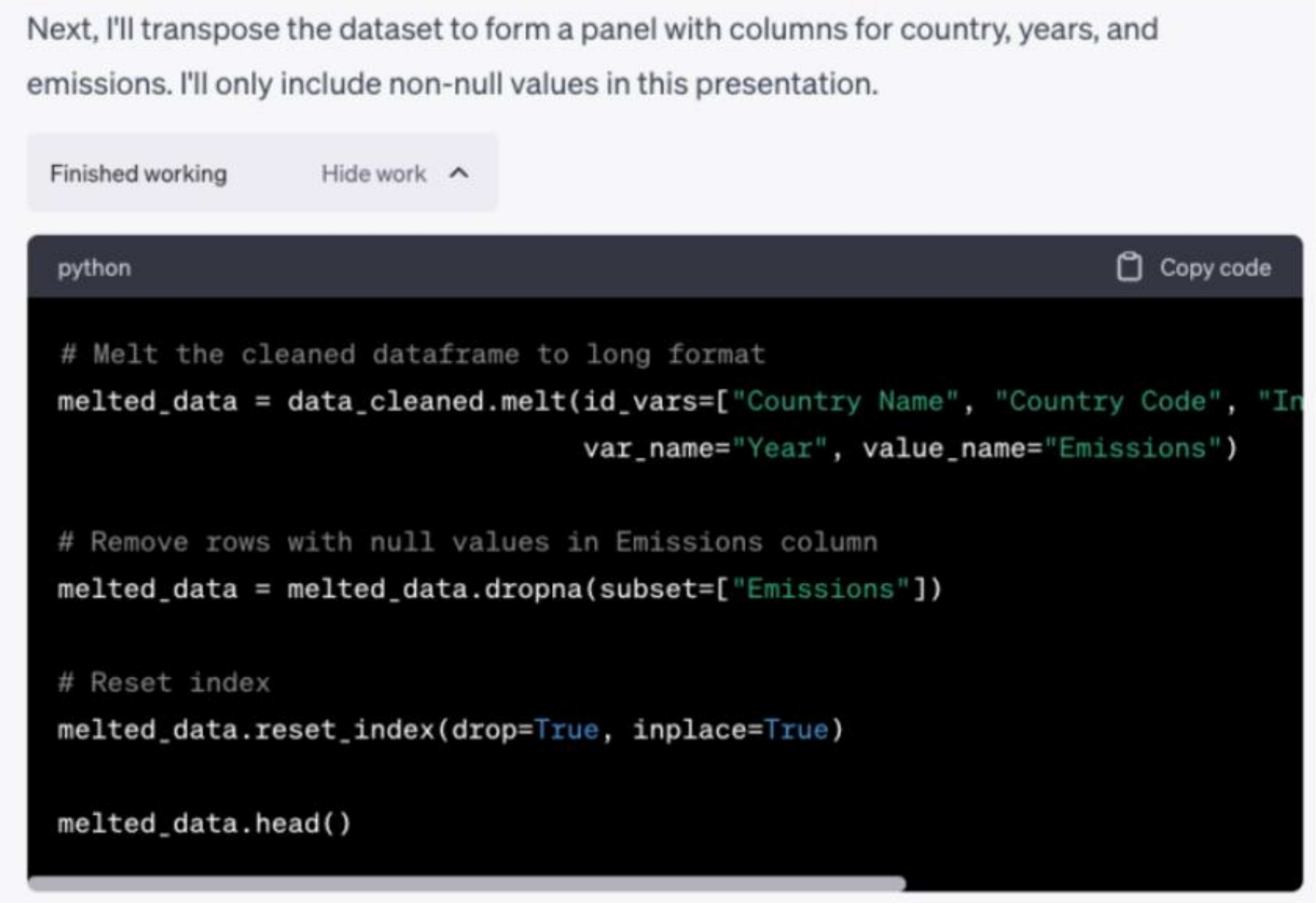
- The dataset has 266 rows.
- The columns include "Country Name", "Country Code", "Indicator Name", "Indicator Code", and years from 1960 to 2020.
- Most of the years from 1960 to 1989 had all null values, so they were removed during the cleaning process.
- The dataset now starts from the year 1990 and goes up to 2020.

Next, I'll transpose the dataset to form a panel with columns for country, years, and emissions. I'll only include non-null values in this presentation.

The dataset has been successfully transposed into a panel format. Here's a sample of the transformed dataset:

© 2024 Percona

Show the Work

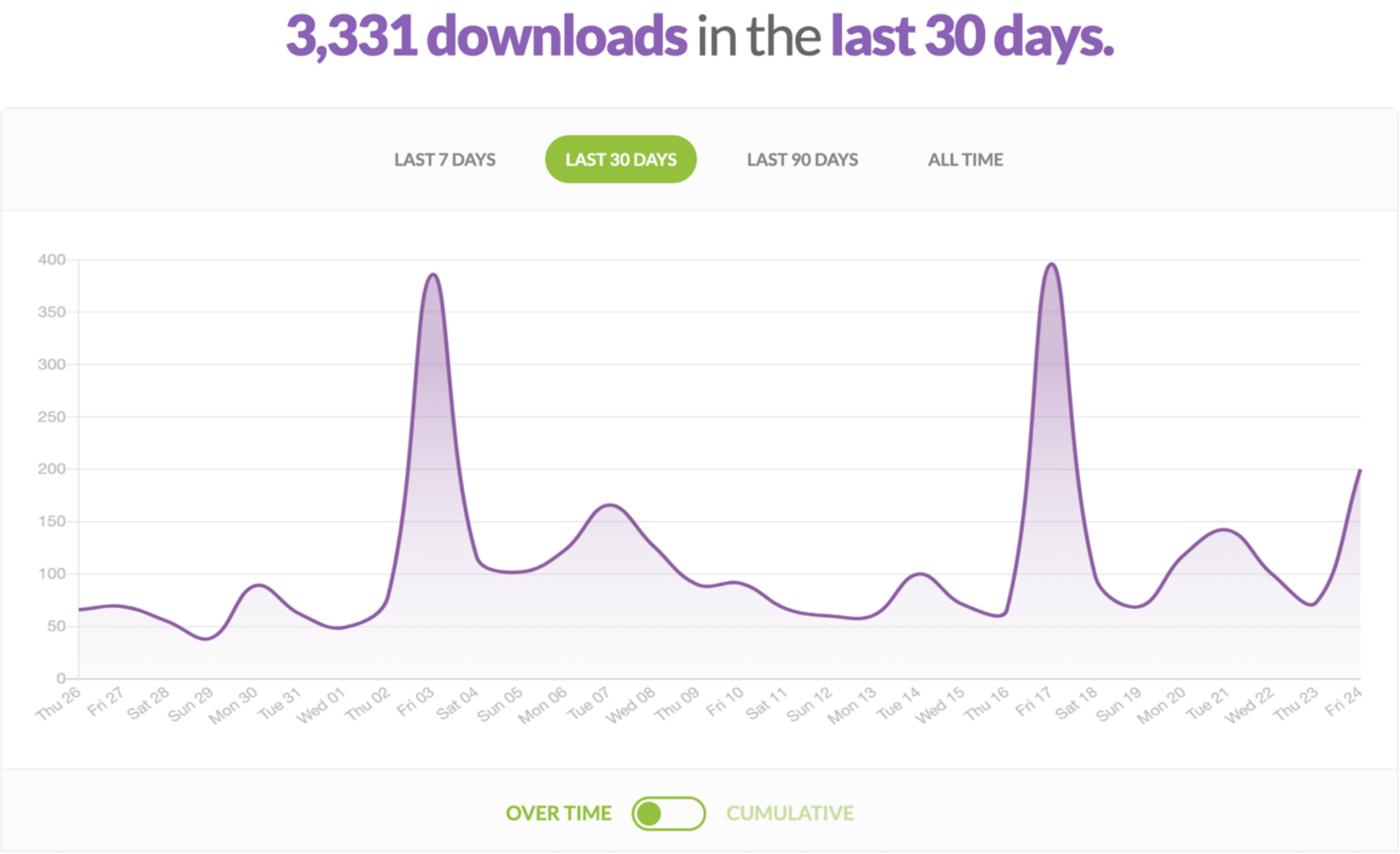


Next, I'll transpose the dataset to form a panel with columns for country, years, and emissions. I'll only include non-null values in this presentation.

```python
# Melt the cleaned dataframe to long format
melted_data = data_cleaned.melt(id_vars=["Country Name", "Country Code", "In
                                var_name="Year", value_name="Emissions")

# Remove rows with null values in Emissions column
melted_data = melted_data.dropna(subset=["Emissions"])

# Reset index
melted_data.reset_index(drop=True, inplace=True)

melted_data.head()
```

© 2024 Percona

Analyze Charts

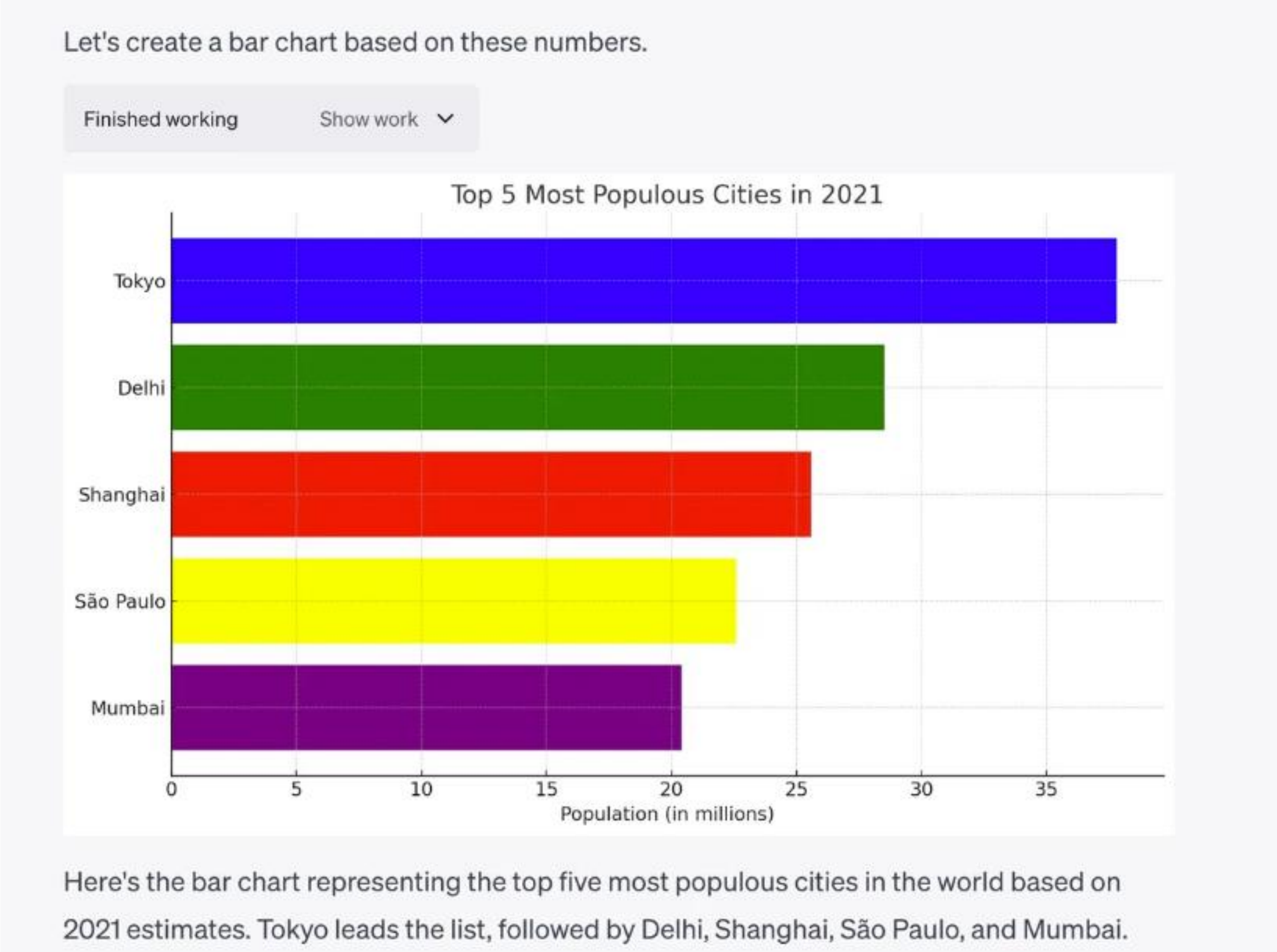


3,331 downloads in the **last 30 days.**

https://www.descript.com/blog/article/using-chatgpt-data-analysis-to-interpret-charts--diagrams

© 2024 Percona

Generate Charts and Diagrams



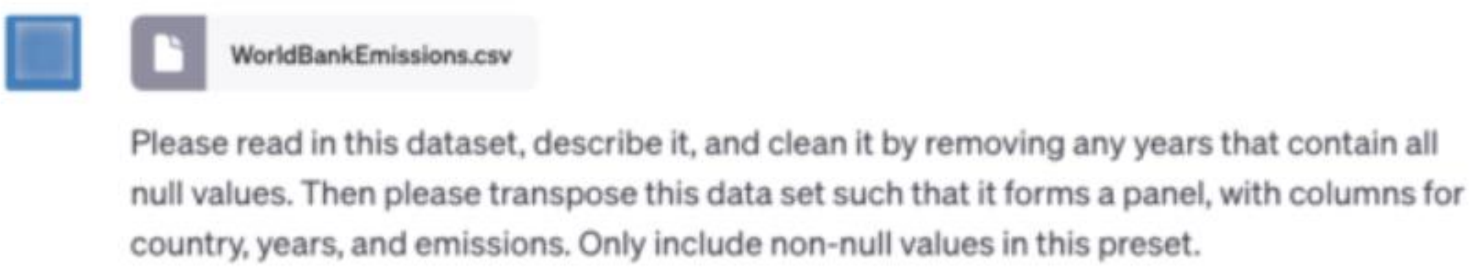
Let's create a bar chart based on these numbers.

Here's the bar chart representing the top five most populous cities in the world based on 2021 estimates. Tokyo leads the list, followed by Delhi, Shanghai, São Paulo, and Mumbai.

https://www.zdnet.com/article/how-to-use-chatgpt-to-make-charts-and-tables/

© 2024 Percona

Data Clean-up



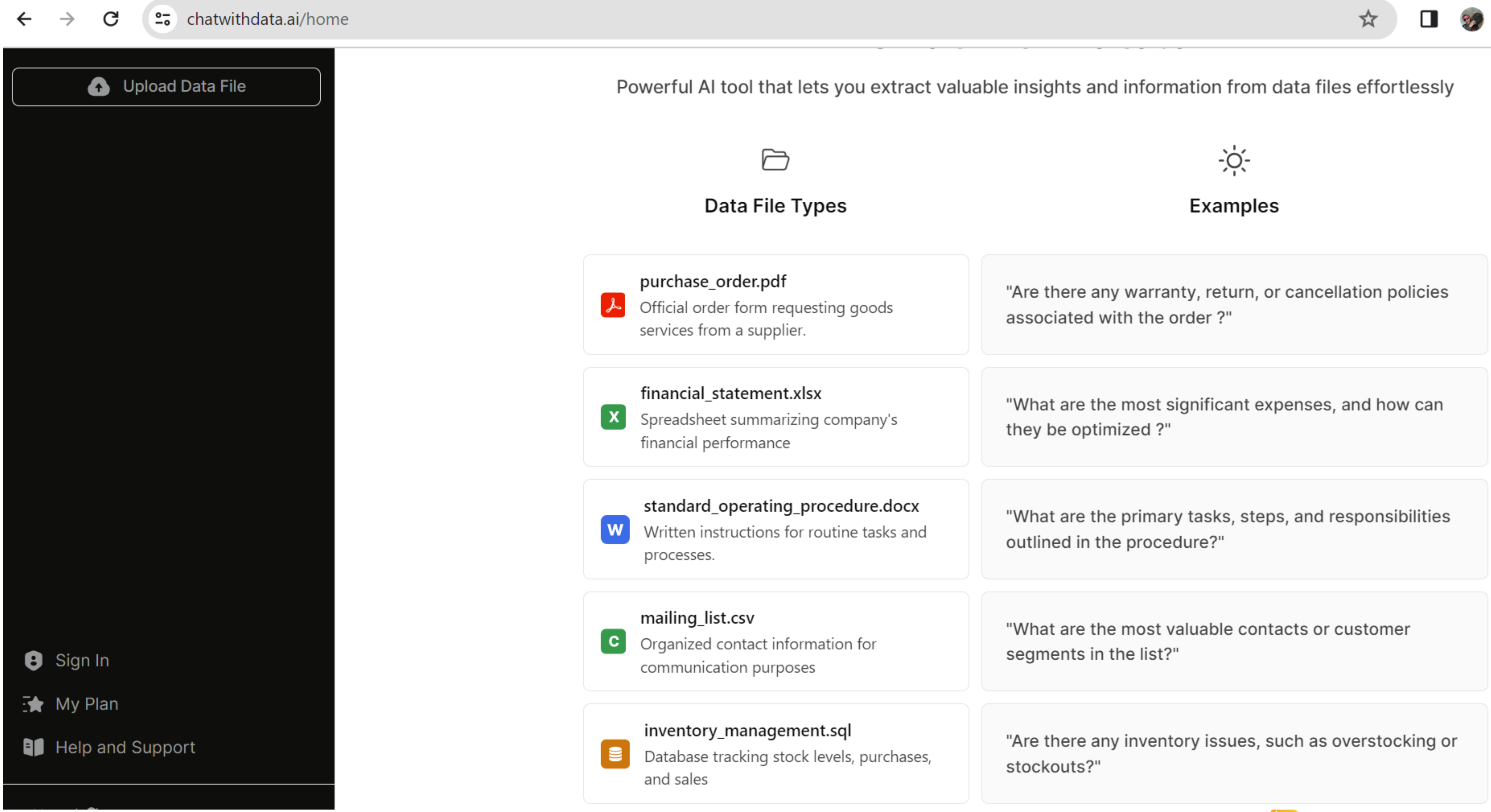
Please read in this dataset, describe it, and clean it by removing any years that contain all null values. Then please transpose this data set such that it forms a panel, with columns for country, years, and emissions. Only include non-null values in this preset.

https://mitsloanedtech.mit.edu/ai/tools/data-analysis/how-to-use-chatgpts-advanced-data-analysis-feature/

> **Prompt**: Today we'll perform data cleaning and preprocessing on a real-world dataset. Write the code to do the following: 1. Fetch the bank marketing dataset from the UCI machine learning repository. Use libraries like requests or urllib. 2. Download the zip file, unzip the contents, and read in the bank-full.csv file into a pandas dataframe (call it data).

https://www.kdnuggets.com/2023/08/harnessing-chatgpt-automated-data-cleaning-preprocessing.html

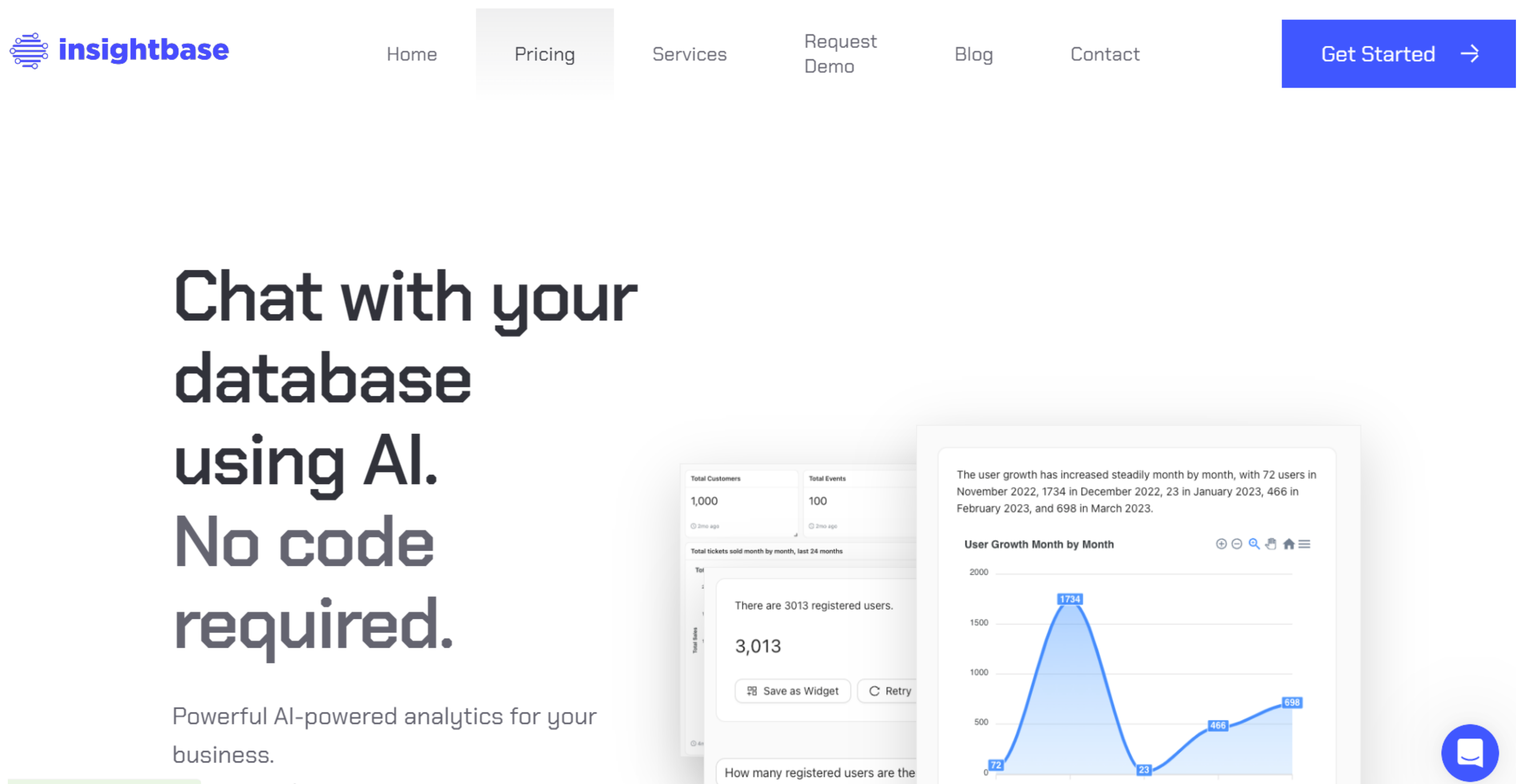
© 2024 Percona



Powerful AI tool that lets you extract valuable insights and information from data files effortlessly

Data File Types	Examples
purchase_order.pdf Official order form requesting goods services from a supplier.	"Are there any warranty, return, or cancellation policies associated with the order ?"
financial_statement.xlsx Spreadsheet summarizing company's financial performance	"What are the most significant expenses, and how can they be optimized ?"
standard_operating_procedure.docx Written instructions for routine tasks and processes.	"What are the primary tasks, steps, and responsibilities outlined in the procedure?"
mailing_list.csv Organized contact information for communication purposes	"What are the most valuable contacts or customer segments in the list?"
inventory_management.sql Database tracking stock levels, purchases, and sales	"Are there any inventory issues, such as overstocking or stockouts?"

© 2024 Percona



Chat with your database using AI. No code required.

Powerful AI-powered analytics for your business.

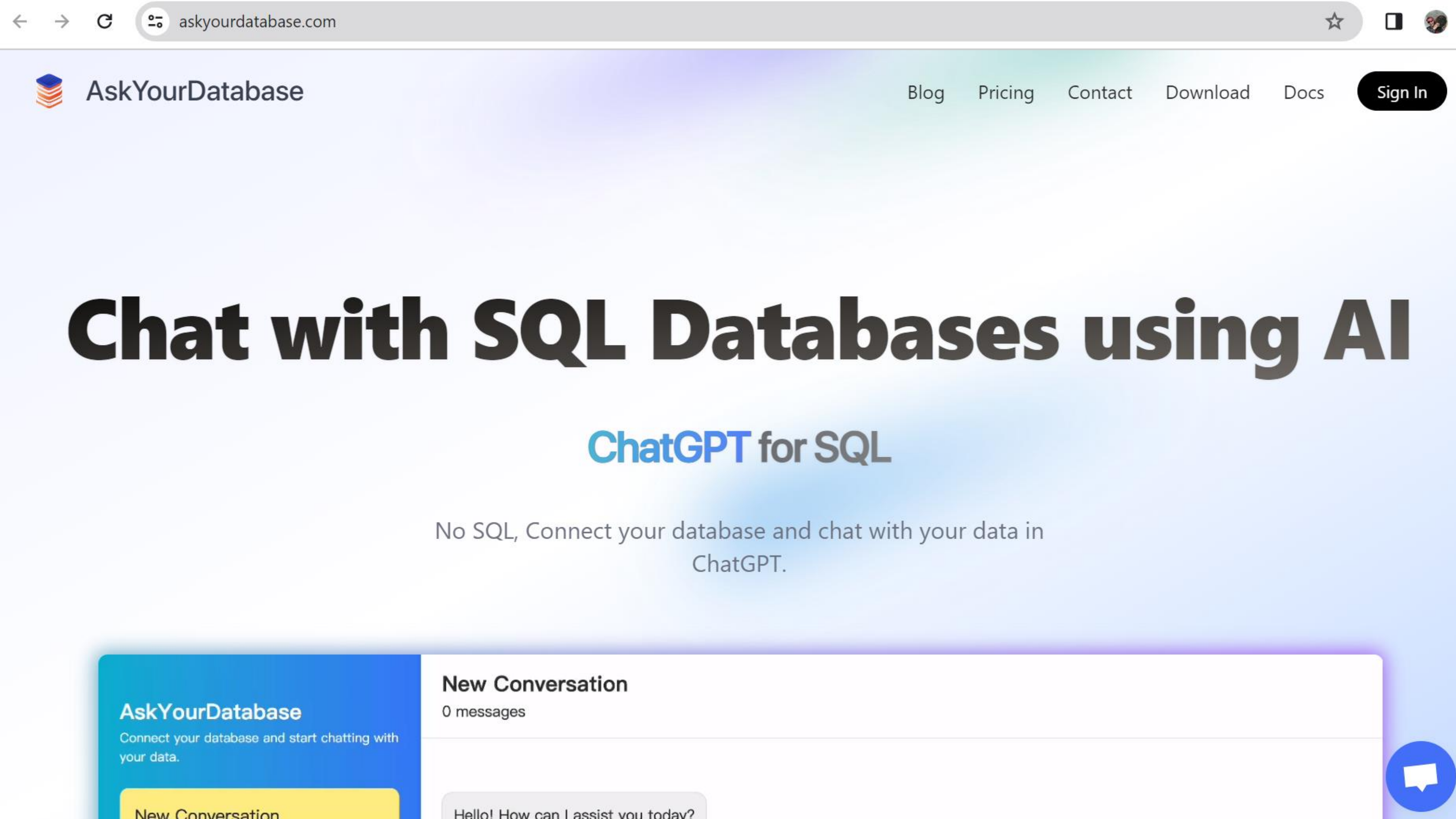
There are 3013 registered users.

3,013

How many registered users are the

The user growth has increased steadily month by month, with 72 users in November 2022, 1734 in December 2022, 23 in January 2023, 466 in February 2023, and 698 in March 2023.

© 2024 Percona



Chat with SQL Databases using AI

ChatGPT for SQL

No SQL, Connect your database and chat with your data in ChatGPT.

AskYourDatabase

Connect your database and start chatting with your data.

New Conversation

0 messages

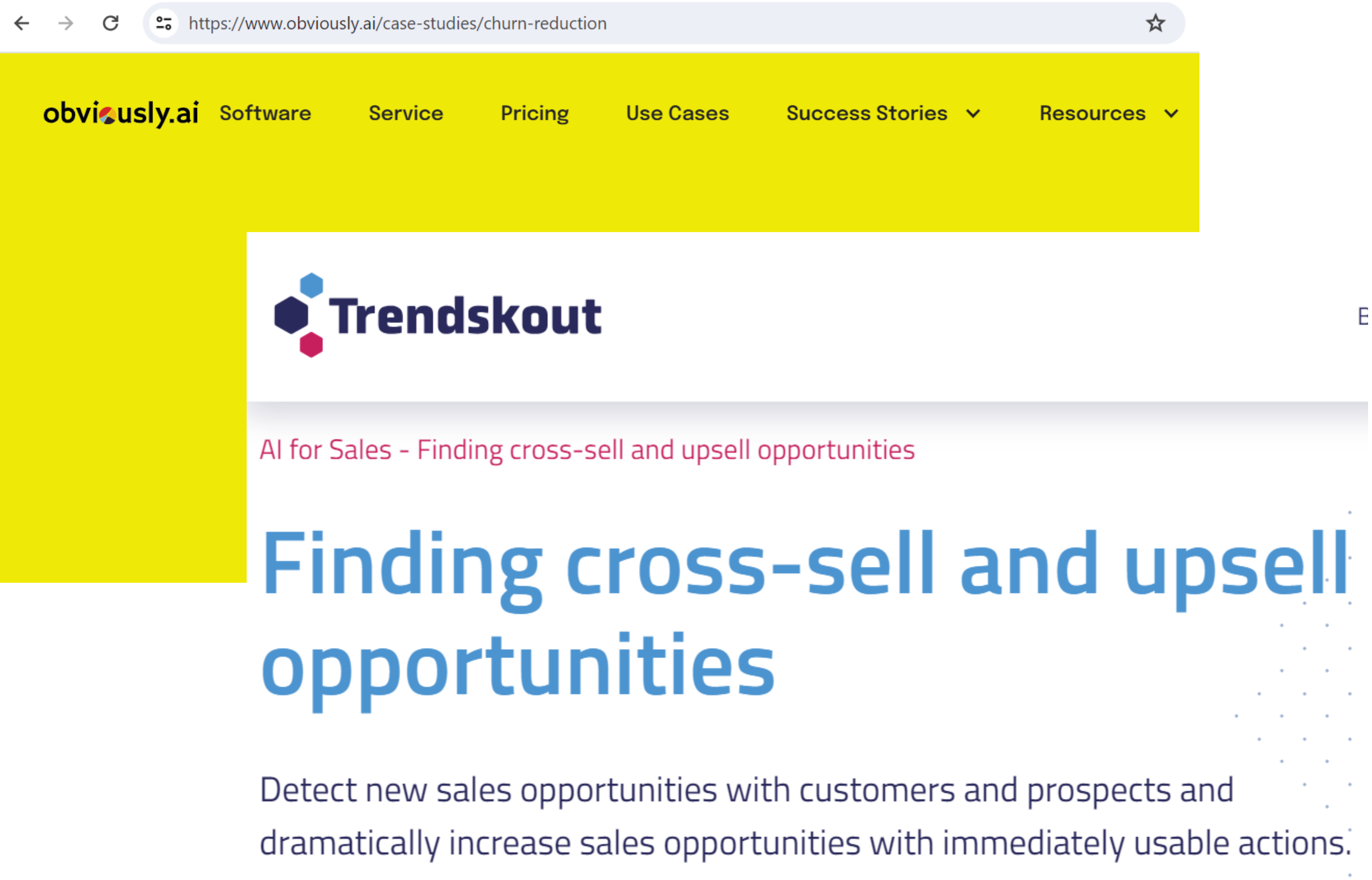
Hello! How can I assist you today?

No Code

- No Code AI Solutions (i.e. ChatBots or Workflows)
- “No Code” for Data Analyses
- AI assisted No Code App Generation

© 2024 Percona

Prediction



AI for Sales - Finding cross-sell and upsell opportunities

Finding cross-sell and upsell opportunities

Detect new sales opportunities with customers and prospects and dramatically increase sales opportunities with immediately usable actions.

© 2024 Percona

Developers

© 2024Percona

Generate Code

ChatGPT

GitHub Copilot

Dev tools built in support

© 2024 Percona

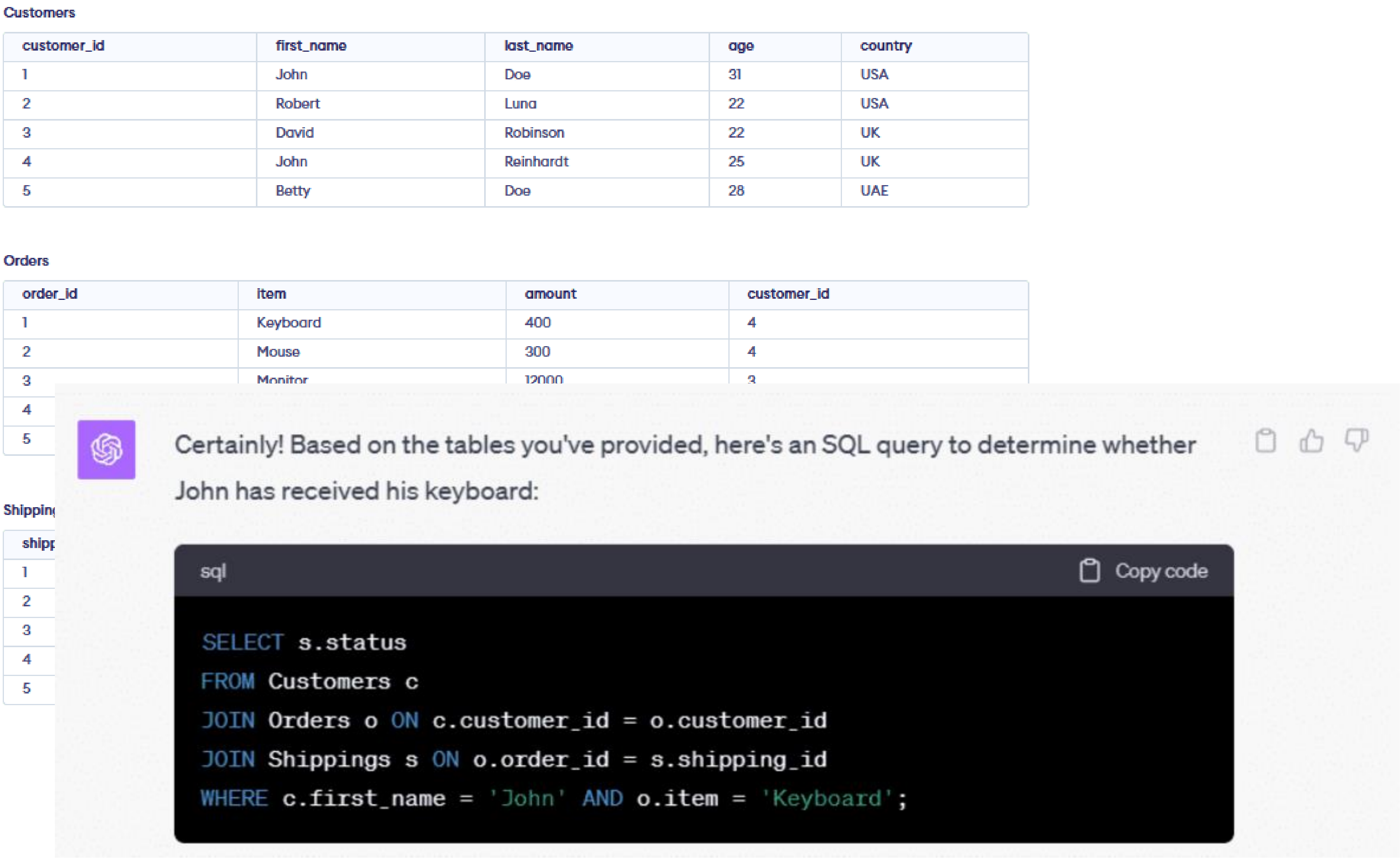
Generate SQL

NL (Natural Language) to SQL

Text to SQL

© 2024 Percona

Even Can use Images



Customers

customer_id	first_name	last_name	age	country
1	John	Doe	31	USA
2	Robert	Luna	22	USA
3	David	Robinson	22	UK
4	John	Reinhardt	25	UK
5	Betty	Doe	28	UAE

Orders

order_id	item	amount	customer_id
1	Keyboard	400	4
2	Mouse	300	4
3	Monitor	12000	3
4			
5			

Shippin

shipp
1
2
3
4
5

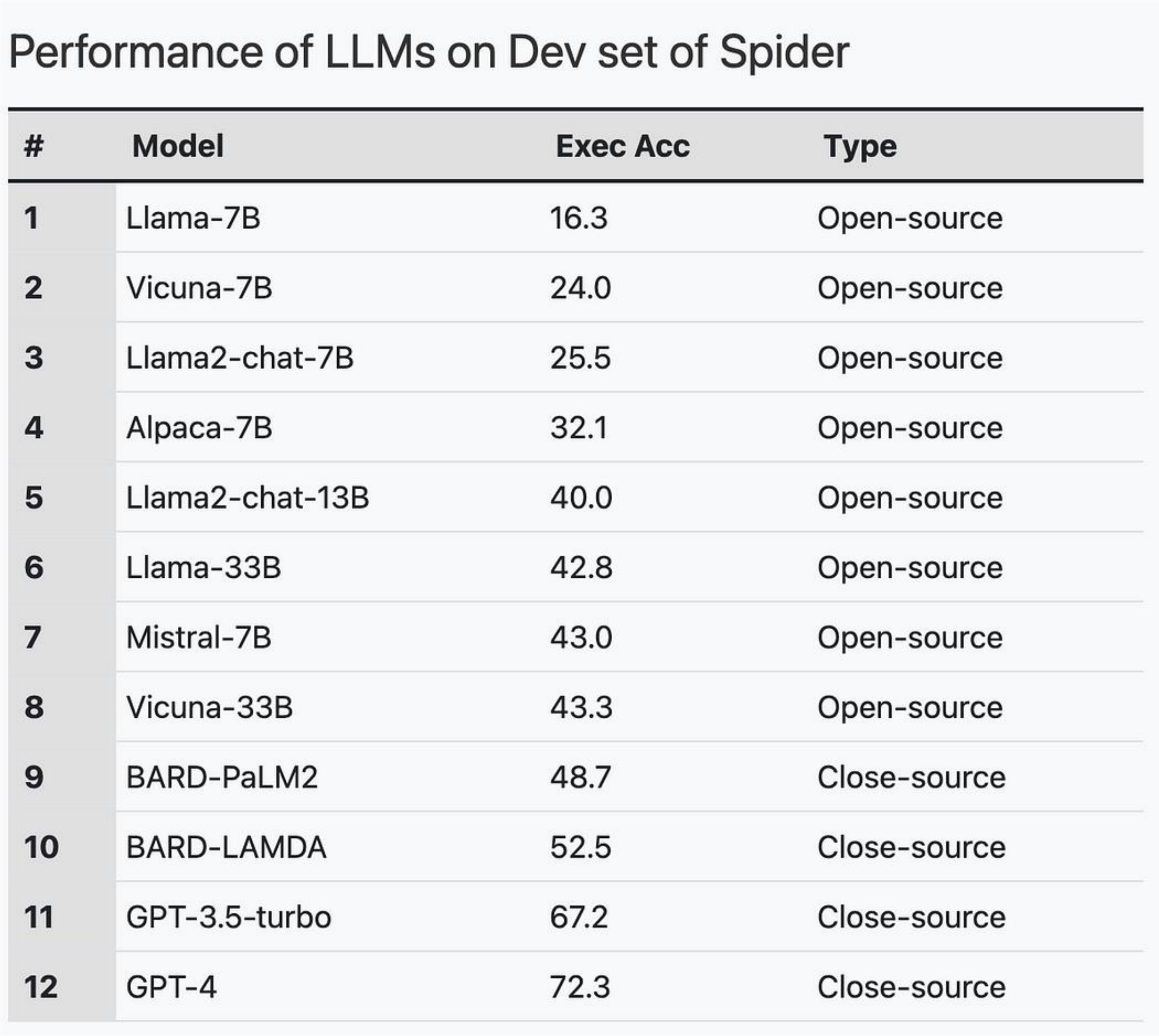
Certainly! Based on the tables you've provided, here's an SQL query to determine whether John has received his keyboard:

```sql
SELECT s.status
FROM Customers c
JOIN Orders o ON c.customer_id = o.customer_id
JOIN Shippings s ON o.order_id = s.shipping_id
WHERE c.first_name = 'John' AND o.item = 'Keyboard';
```

https://www.kdnuggets.com/5-ways-you-can-use-chatgpt-vision-for-data-analysis

© 2024 Percona

Evolving Field...



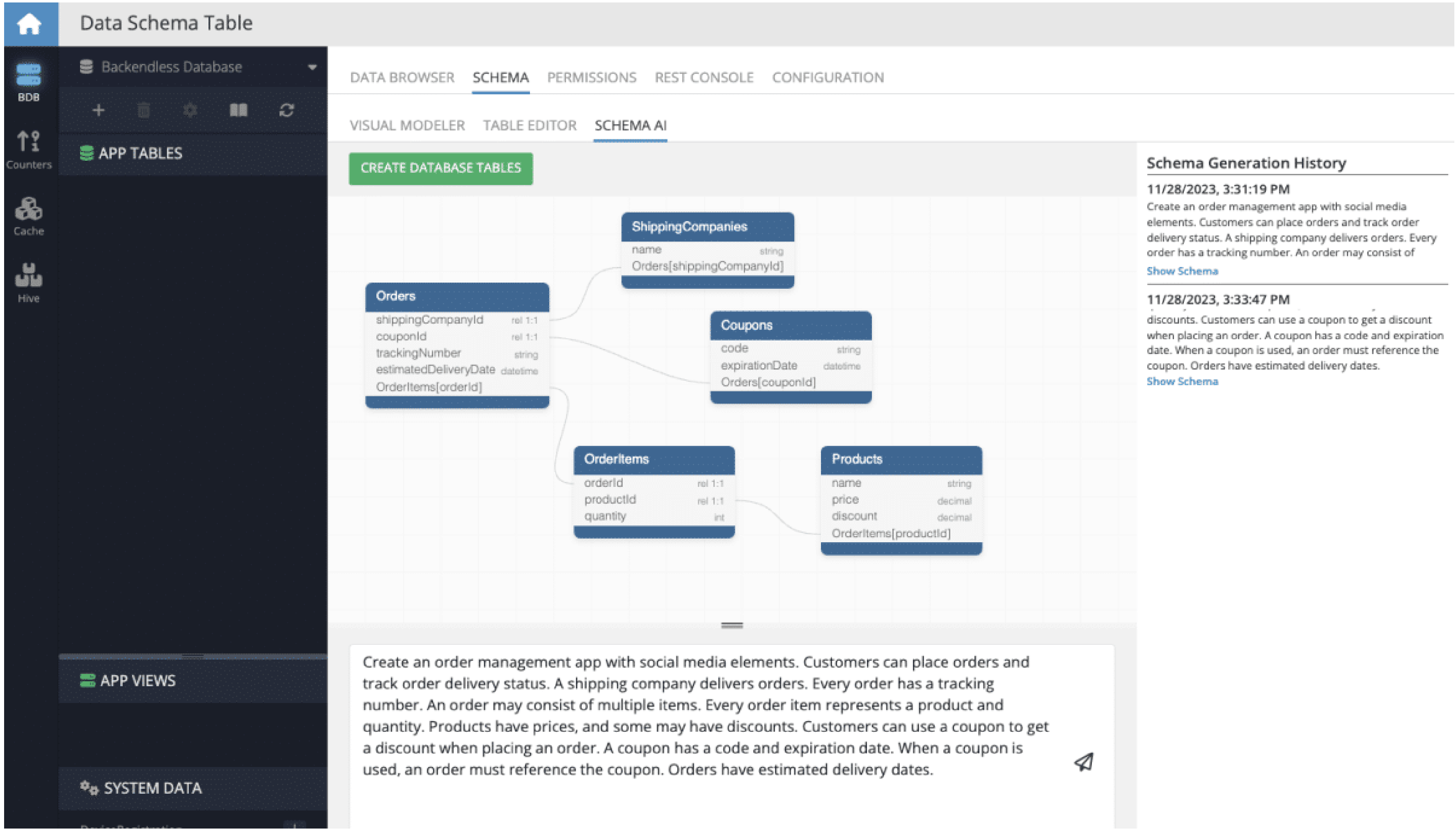
Performance of LLMs on Dev set of Spider

#	Model	Exec Acc	Type
1	Llama-7B	16.3	Open-source
2	Vicuna-7B	24.0	Open-source
3	Llama2-chat-7B	25.5	Open-source
4	Alpaca-7B	32.1	Open-source
5	Llama2-chat-13B	40.0	Open-source
6	Llama-33B	42.8	Open-source
7	Mistral-7B	43.0	Open-source
8	Vicuna-33B	43.3	Open-source
9	BARD-PaLM2	48.7	Close-source
10	BARD-LAMDA	52.5	Close-source
11	GPT-3.5-turbo	67.2	Close-source
12	GPT-4	72.3	Close-source

https://medium.com/dataherald/comparing-open-source-llms-for-nl-to-sql-53e6de42aee8

© 2024 Percona

Schema Design



ShippingCompanies	
name	string
Orders[shippingCompanyId]	

Orders	
shippingCompanyId	rel 1:1
couponId	rel 1:1
trackingNumber	string
estimatedDeliveryDate	datetime
OrderItems[orderId]	

Coupons	
code	string
expirationDate	datetime
Orders[couponId]	

OrderItems	
orderId	rel 1:1
productId	rel 1:1
quantity	int

Products	
name	string
price	decimal
discount	decimal
OrderItems[productId]	

Create an order management app with social media elements. Customers can place orders and track order delivery status. A shipping company delivers orders. Every order has a tracking number. An order may consist of multiple items. Every order item represents a product and quantity. Products have prices, and some may have discounts. Customers can use a coupon to get a discount when placing an order. A coupon has a code and expiration date. When a coupon is used, an order must reference the coupon. Orders have estimated delivery dates.

Schema Generation History

11/28/2023, 3:31:19 PM

Create an order management app with social media elements. Customers can place orders and track order delivery status. A shipping company delivers orders. Every order has a tracking number. An order may consist of

Show Schema

11/28/2023, 3:33:47 PM

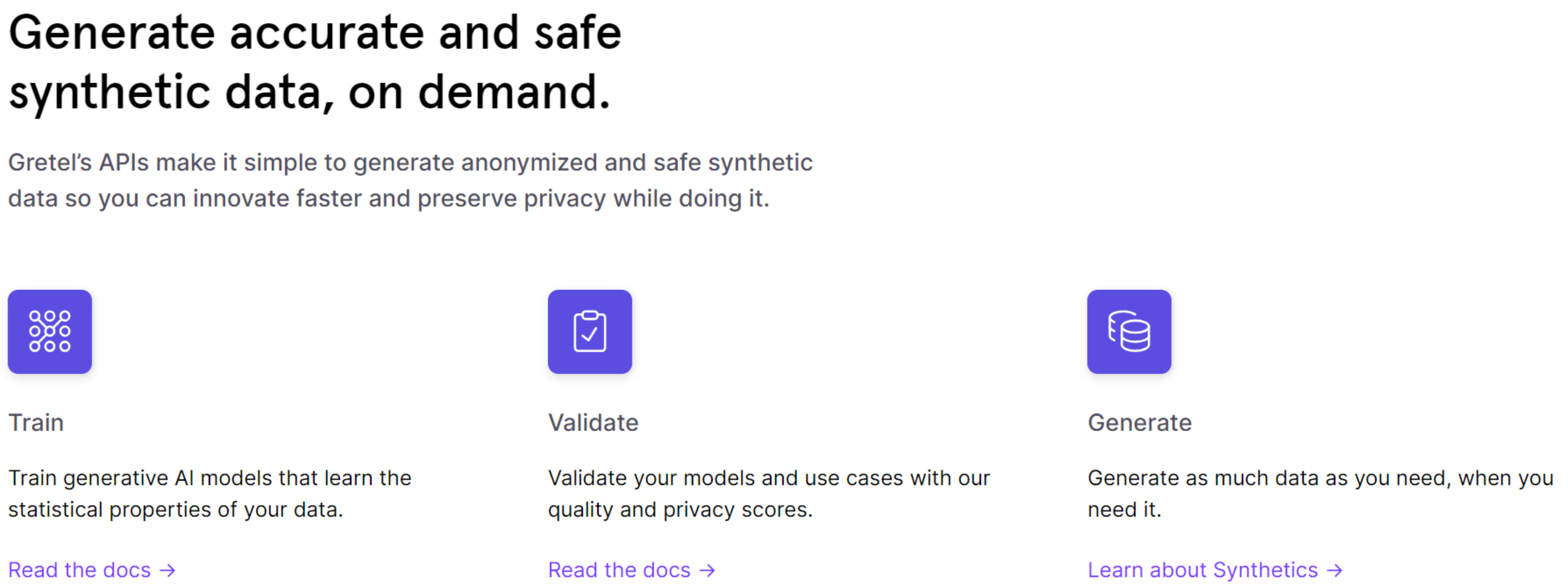
discounts. Customers can use a coupon to get a discount when placing an order. A coupon has a code and expiration date. When a coupon is used, an order must reference the coupon. Orders have estimated delivery dates.

Show Schema

https://backendless.com/introducing-schema-ai-revolutionizing-database-design-with-ai-powered-simplicity/

© 2024 Percona

PERCONA



Generate accurate and safe synthetic data, on demand.

Gretel's APIs make it simple to generate anonymized and safe synthetic data so you can innovate faster and preserve privacy while doing it.

Train

Train generative AI models that learn the statistical properties of your data.

Read the docs →

Validate

Validate your models and use cases with our quality and privacy scores.

Read the docs →

Generate

Generate as much data as you need, when you need it.

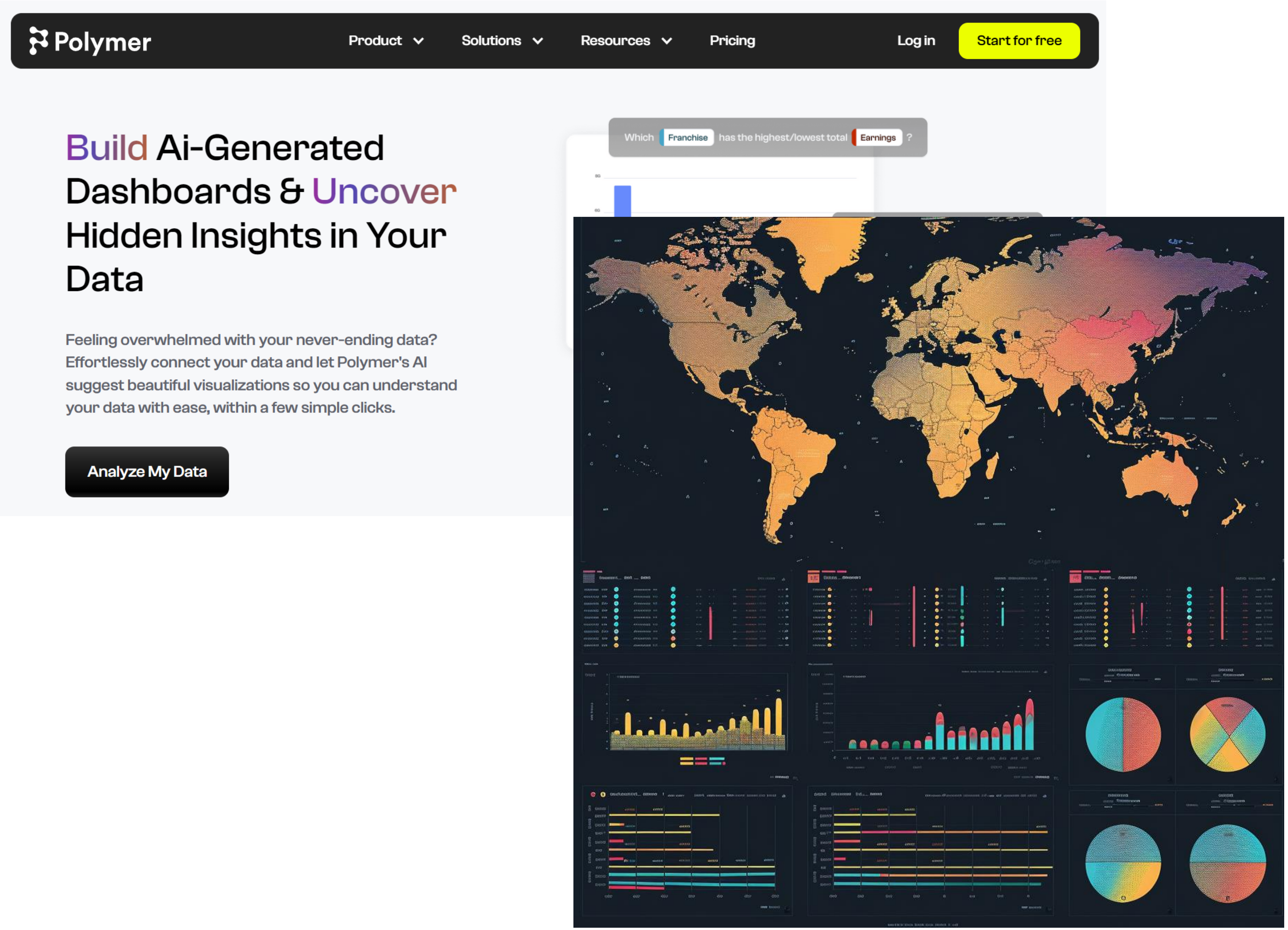
Learn about Synthetics →

https://gretel.ai/

Synthetic Data

© 2024 Percona

AI Dashboards



Build Ai-Generated Dashboards & Uncover Hidden Insights in Your Data

Feeling overwhelmed with your never-ending data? Effortlessly connect your data and let Polymer's AI suggest beautiful visualizations so you can understand your data with ease, within a few simple clicks.

Which Franchise has the highest/lowest total Earnings ?

© 2024 Percona

Vector Search

Technology Supporting building applications which use AI

© 2024 Percona

What's up with Vectors

- **AI Algorithms do not work with words and Images inside**
- **They work with Vectors**
- **Objects encoded to Vector form called "Embeddings"**
- **Finding Similar Vectors is key task**
- **Traditional Database Algorithms do not handle it well**

https://manticoresearch.com/blog/vector-search-in-databases/

Vector search support in databases

Opensource vector dbs

Milvus	2019
Vespa	2020
Weaviate	2021
Qdrant	2022

Opensource dbs and search engines

PostgreSQL	2021
Lucene	2021
Opensearch	2022
Redis	2022
SOLR	2022
Cassandra	**2023**
Typesense	**2023**
Clickhouse	**2023**
Manticore Search	**2023**
Meilisearch	**2023**
MariaDB	**In progress**
MySQL	Not yet

Non-open source dbs

Elasticsearch	2019
Oracle	**2023**
MongoDB	**2023**

Clouds

Pinecone	2019
Amazon Elasticsearch / Opensearch	2020
Google Cloud Platform	2021
Alibaba Cloud AnalyticDB	**2023**
Azure	**2023**
Amazon DocumentDB	**2023**
Cloudflare Vectorize	**2023**

© 2024 Percona

Operations

© 2024Percona

Security

Automatically Detecting data worth sanitizing

Classical Intrusion Detection

Many more

© 2024 Percona

Performance

- **Ottertune**
- **DbTune**
- **Releem**
- **EverSQL**
- **Built In**

© 2024 Percona

FinOps

Better Performance – Lower Costs

Using Most Cost Effective Infrastructure (Cloud)

Better cost Allocation

© 2024 Percona

General Operations

AiOPS

High Availability/Self-Healing

Auto Scaling

Incident Resolution

© 2024 Percona

Coroot

Observability Solution, Aspiring to revolutionize automated root cause analysis

© 2024 Percona

Concerns and Problems

© 2024Percona

Intellectual Property

WHO OWNS INTELLECTUAL PROPERTY CREATED BY AI?

CAN AI INFRINGE TRAINING DATA OWNERS COPYRIGHT?

WHAT IF TRAINING DATA "RE-CREATED" BY AI?

© 2024 Percona

PERCONA

Security and Privacy

- Best AI tends to be Cloud Based
- How is data shared with AI really used?
- Enterprise AI Offerings
- Open Source AI and Datasets
- “Small” Language Models you can run locally

© 2024 Percona

PERCONA

Lock-In

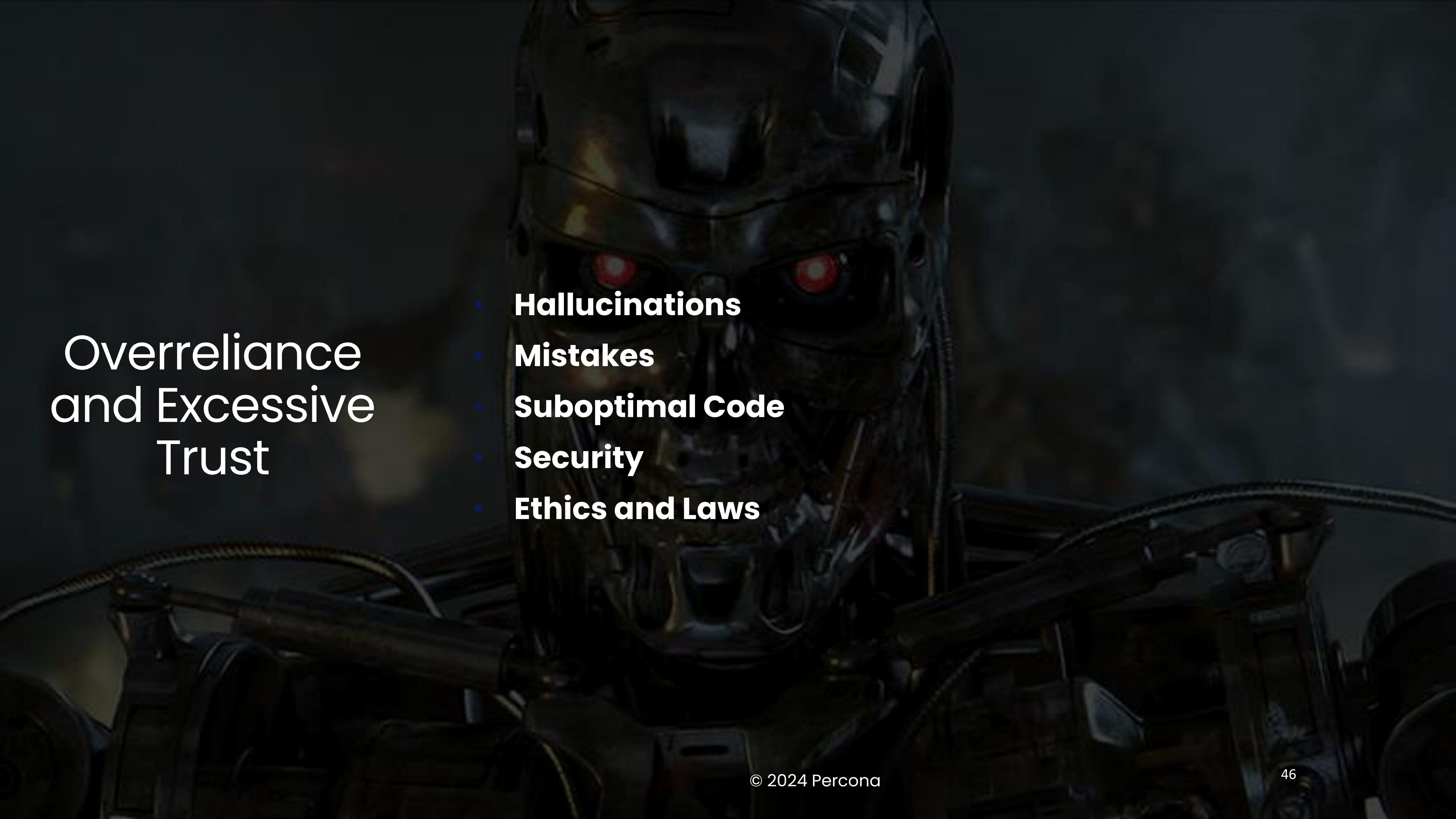
OpenAI is not really Open

“Open Source AI” is often misused term

More Open Solutions are advancing Quickly

“Commodity AI” is likely to be good enough for most tasks

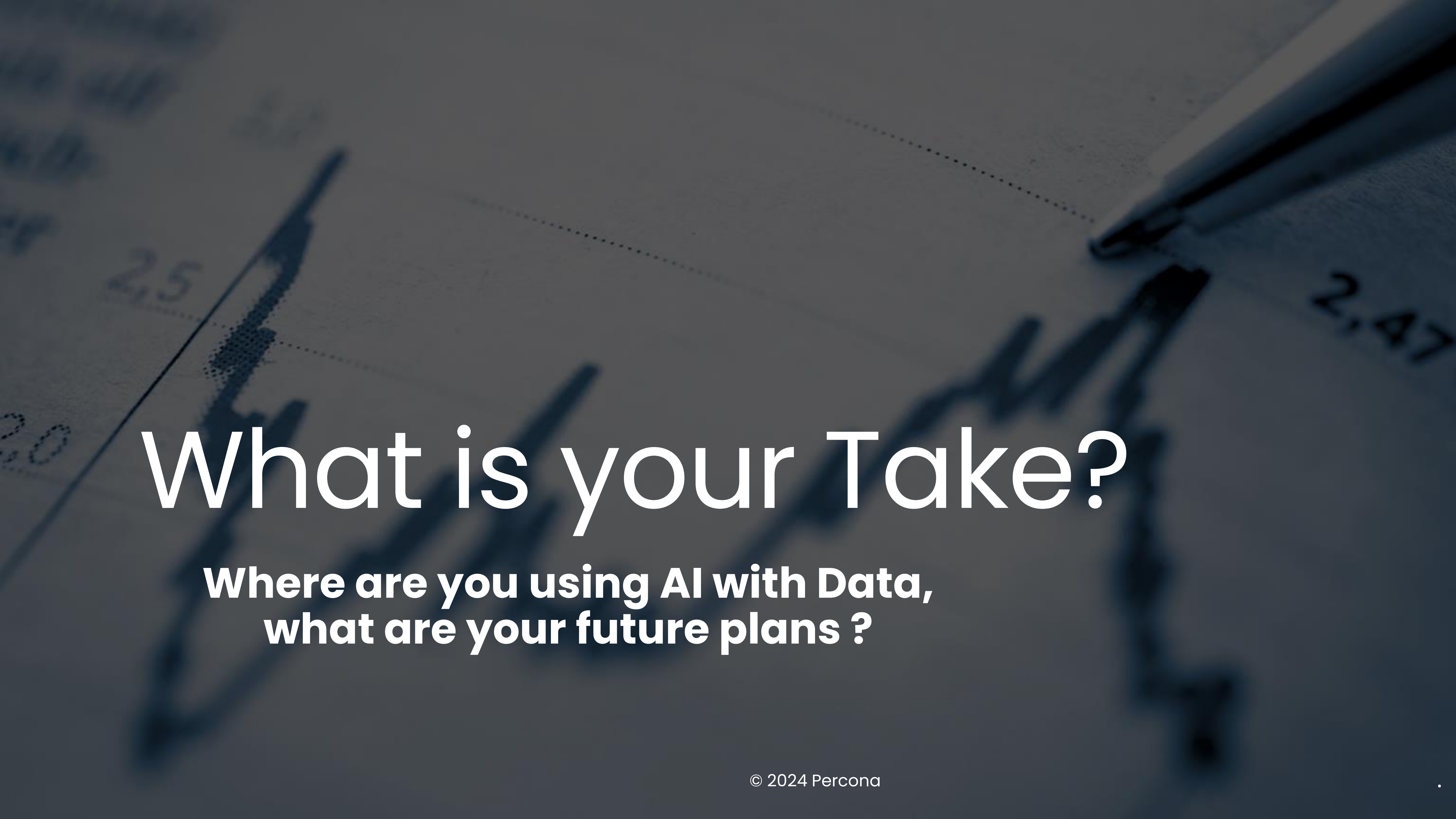
© 2024 Percona



Overreliance and Excessive Trust

- **Hallucinations**
- **Mistakes**
- **Suboptimal Code**
- **Security**
- **Ethics and Laws**

© 2024 Percona



What is your Take?

**Where are you using AI with Data,
what are you future plans ?**

© 2024 Percona

Thank you, Let’s Connect!

https://www.linkedin.com/in/peterzaitsev/
https://twitter.com/PeterZaitsev
http://www.peterzaitsev.com