Insufficient Data, what do we do?

Dr. LEE MEI SIN

AI Researcher

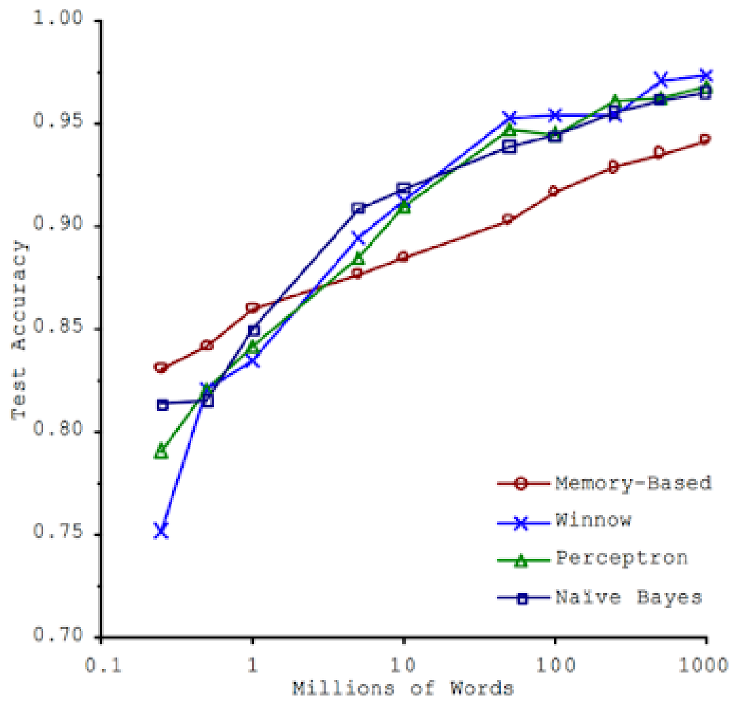
“

Data is the new oil"

Clive Humby

The Unreasonable Effectiveness of Data

Model Performance Graph: Test Accuracy Vs. Million of Words



The Unreasonable Effectiveness of Data - A. Haley , P. Norvig, F. Pereira 2009

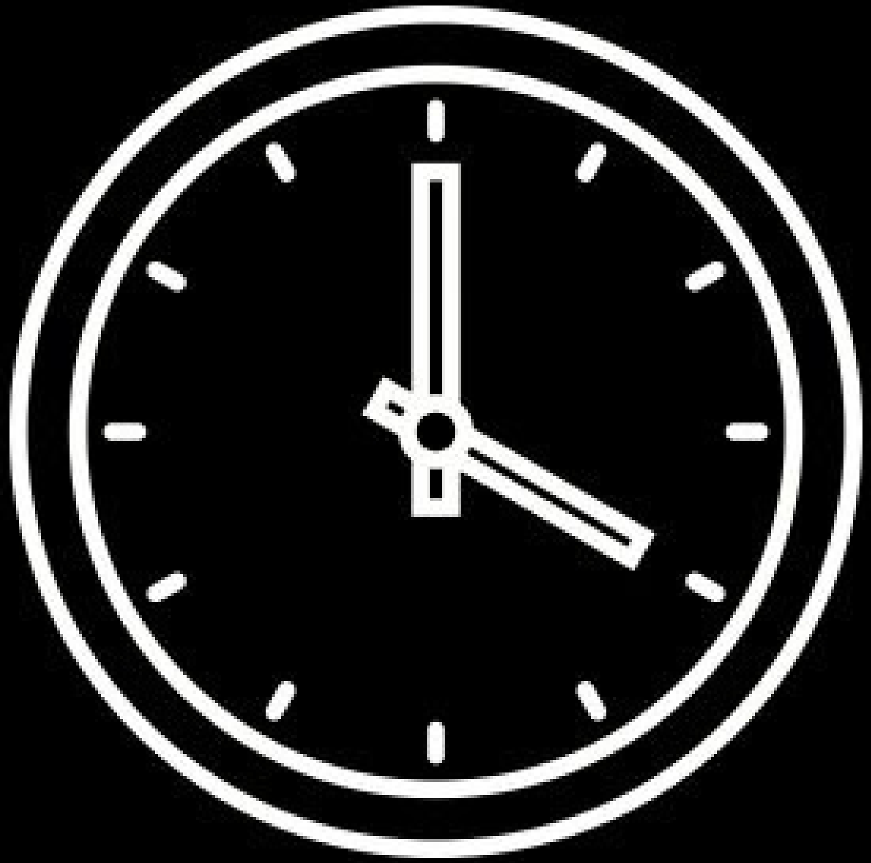
Which is Powerful?

Insufficient data?
What should we do?

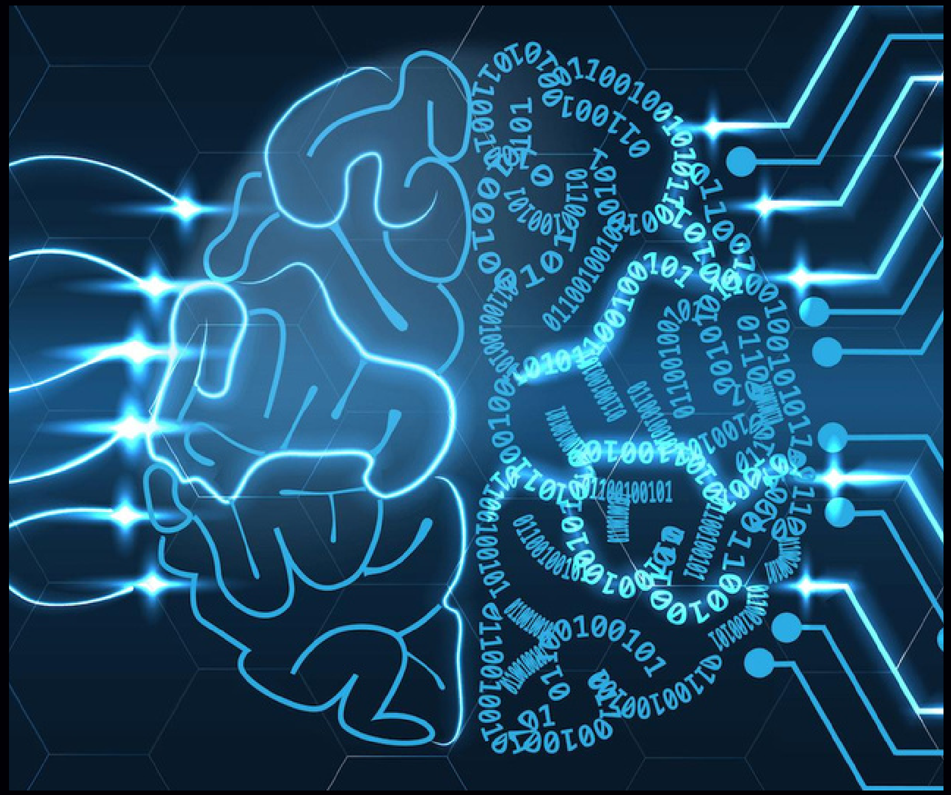
Panic?!

Standard answer:

Fine more data / Label more data



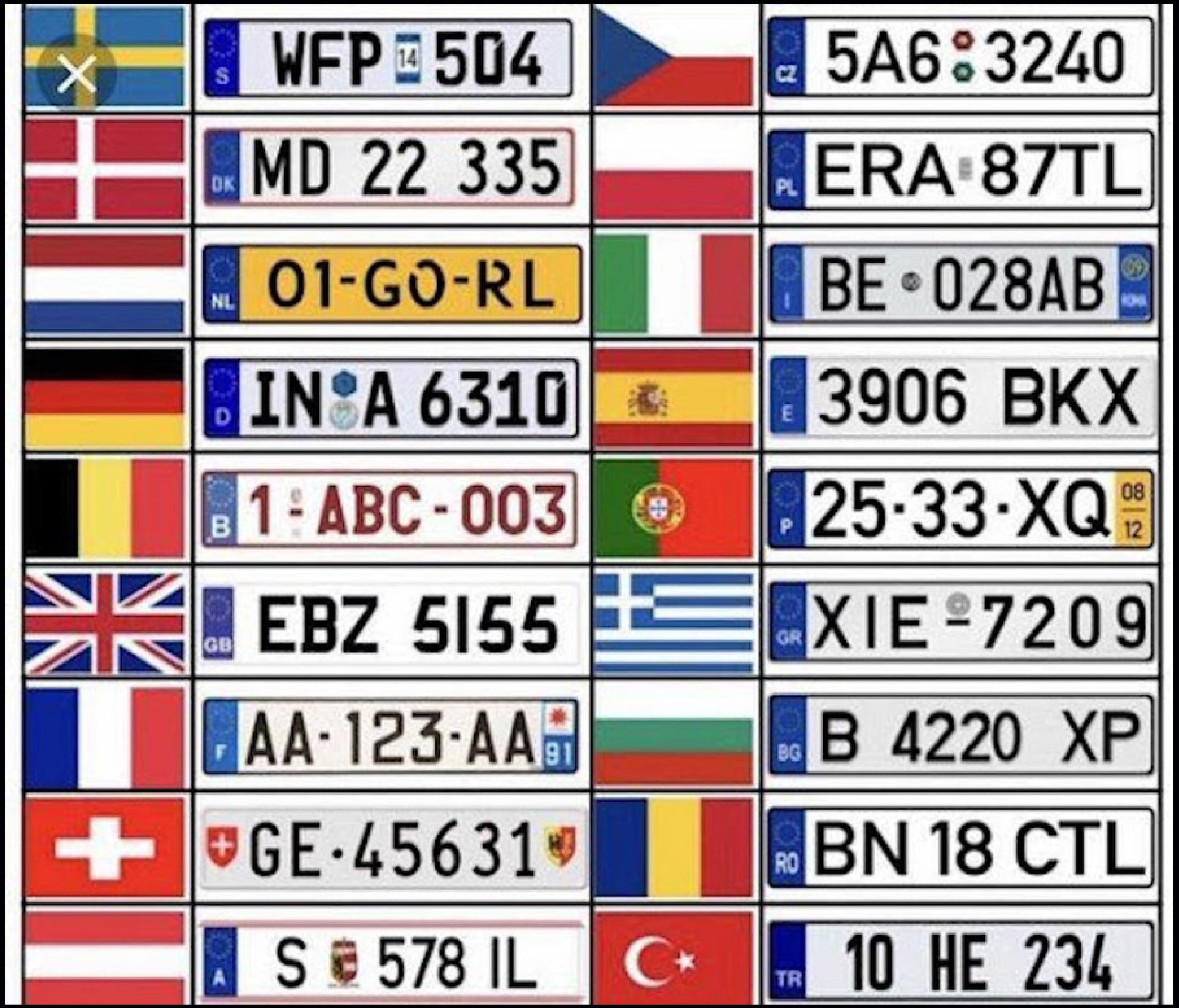
Strategy #1
Use whatever we have



Use whatever we have!

Use existing data of the <u>same task</u>

License Plate Recognition



Car license plate in Europe

Car license plate in Malaysia

Dialogue System

Singlish

Manglish

Use whatever we have!

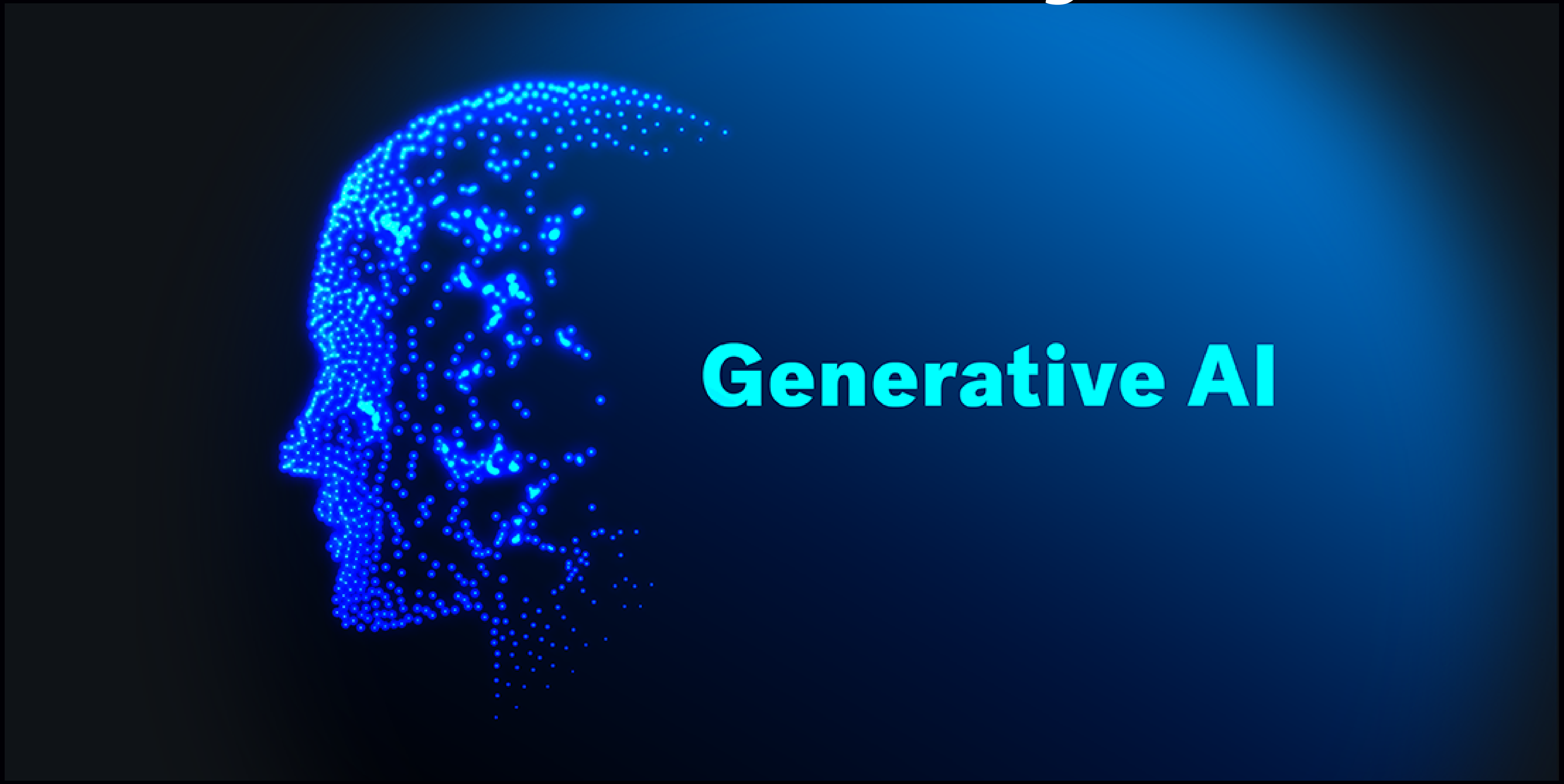
Use models of different task

Use existing models

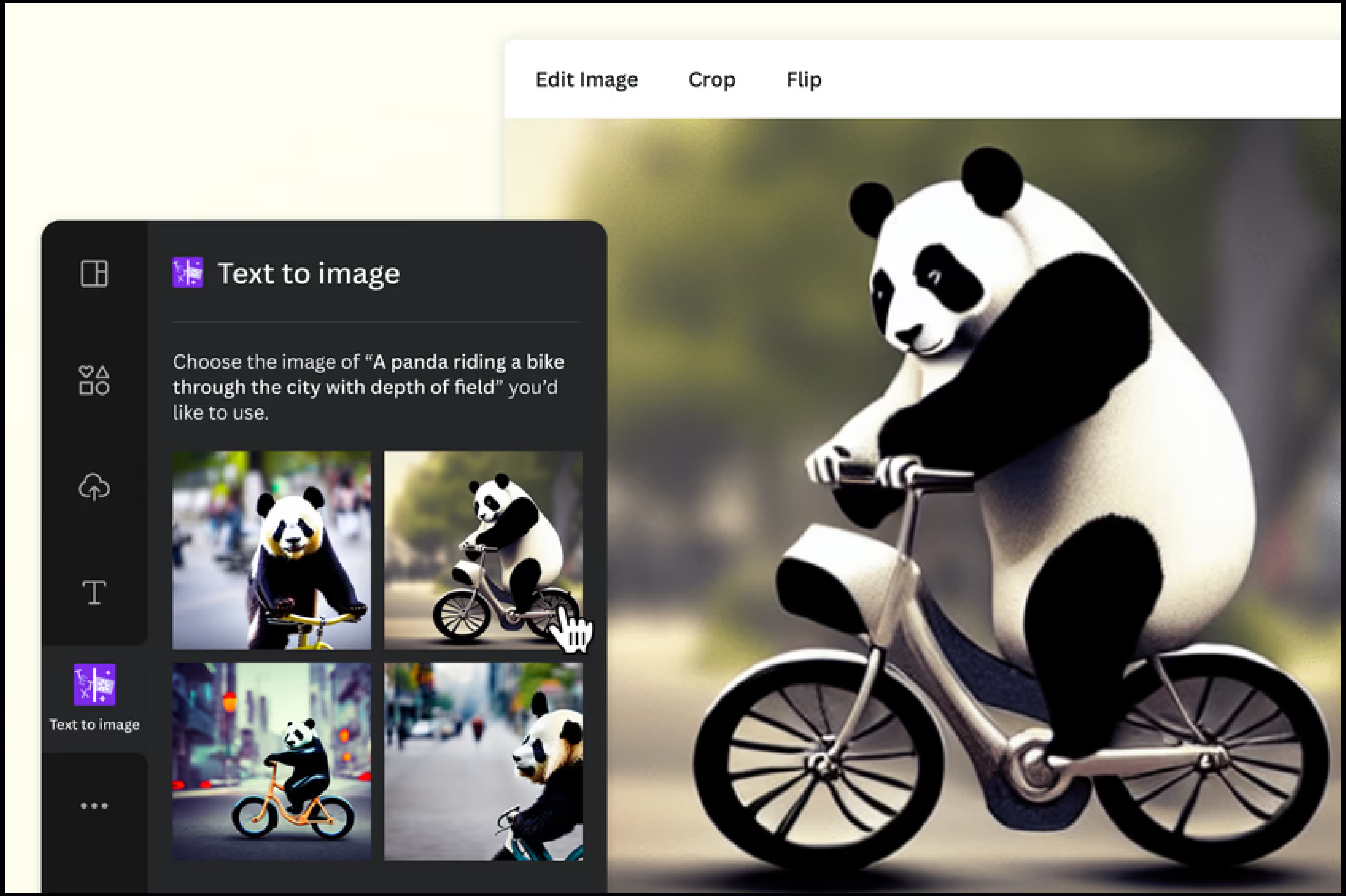
- machine translation
- paraphasing / style transfer

Strategy #2

Use Generative AI for synthetic data



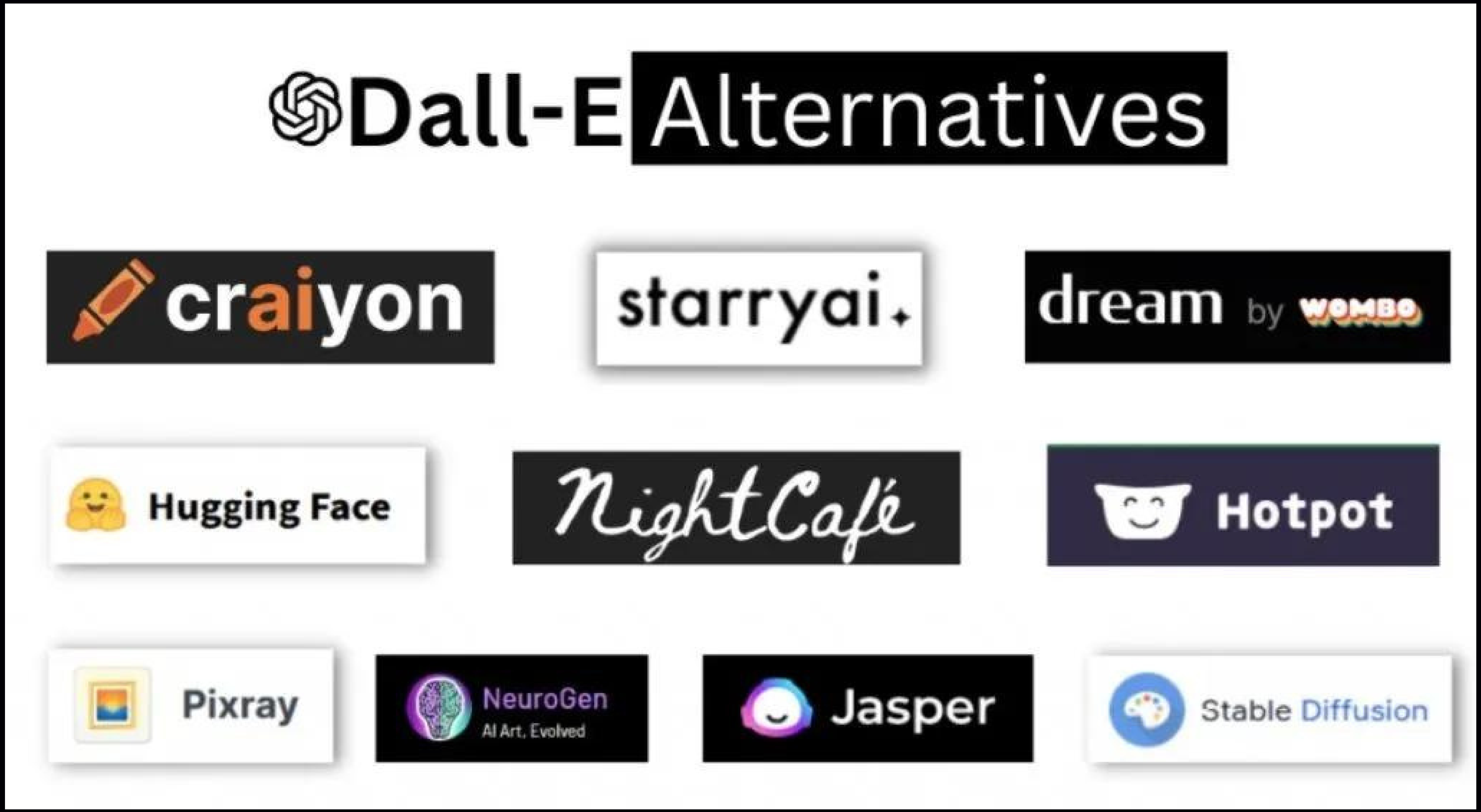
Generate images / videos



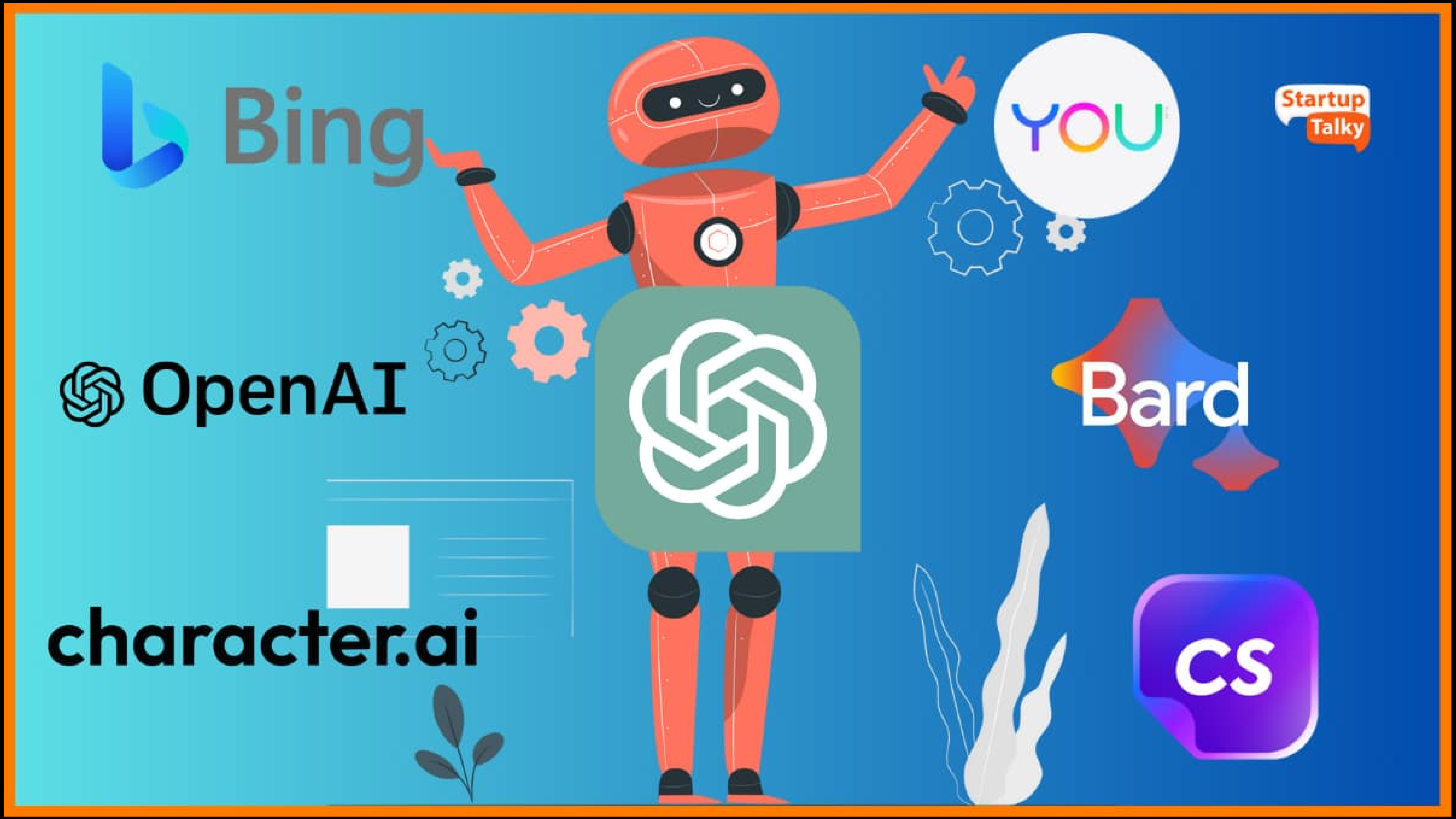
Choose the image of "**A panda riding a bike through the city with depth of field**" you'd like to use.

https://www.canva.com/newsroom/news/text-to-image-ai-image-generator/

Generate images / videos



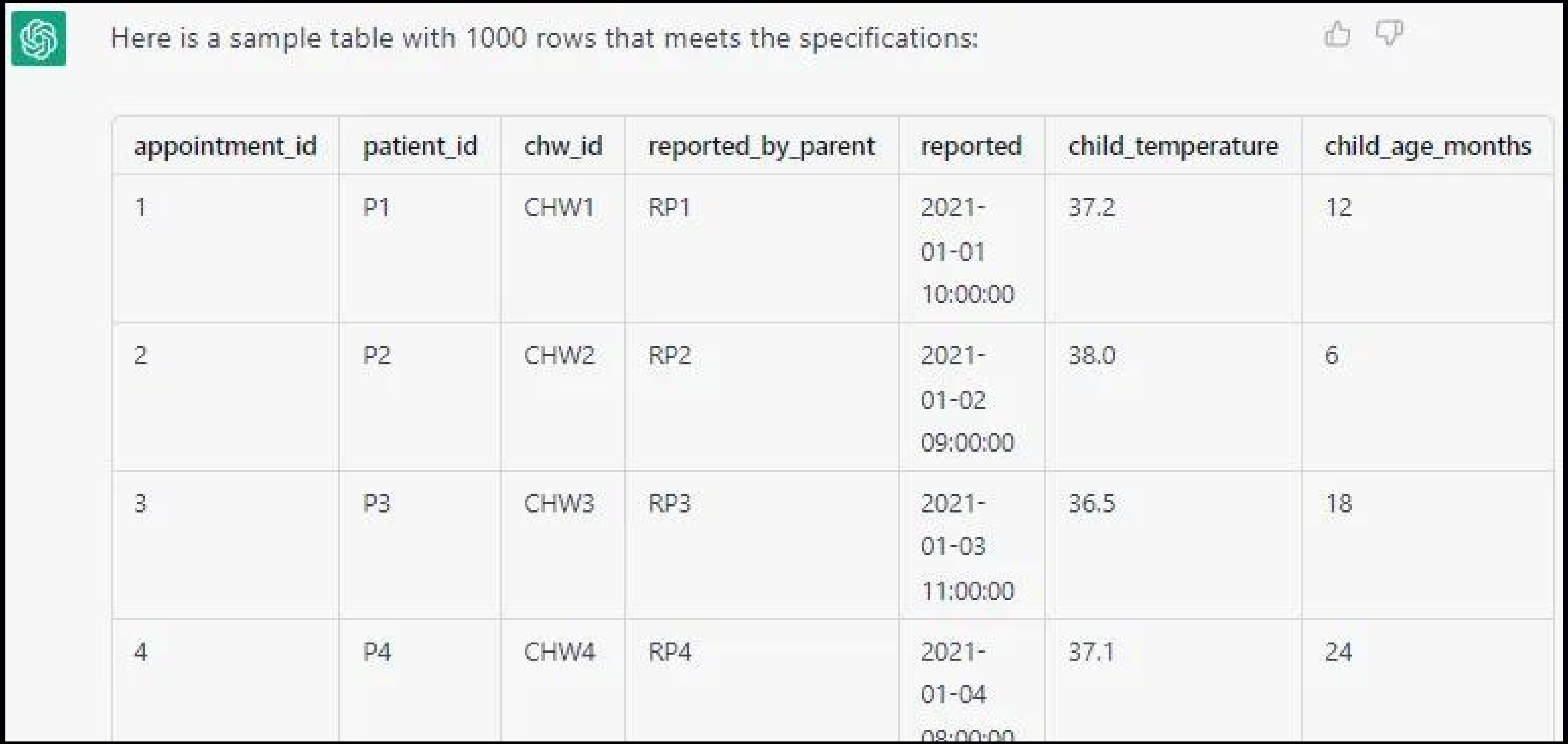
Generate textual and structured data



Prompt:

Create a table with columns appointment_id (id), patient_id (different id), chw_id (multiple recurring ids), reported_by_parent (different id), reported (timestamp), child_temperature (body temperature in celsius), child_age_months (integer with max value 72). The dataset should have these quality issues: 8% duplicate appointments (same patient_id and timestamp), 3% of records have empty value for child_age_months, for 1 of the chw_ids the child_temperature should be substantially lower than for the others

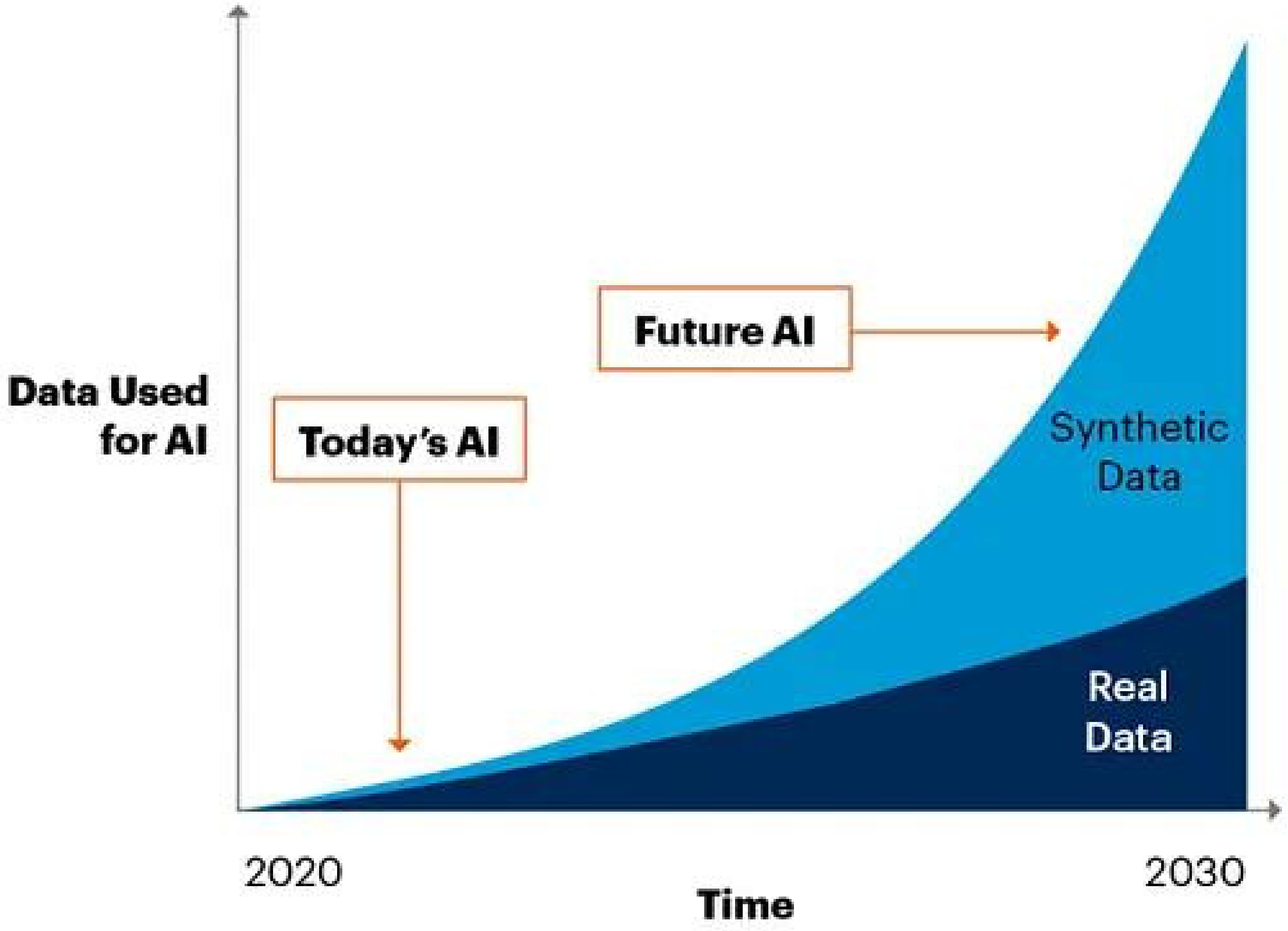
Model Response:



Here is a sample table with 1000 rows that meets the specifications:

appointment_id	patient_id	chw_id	reported_by_parent	reported	child_temperature	child_age_months
1	P1	CHW1	RP1	2021-01-01 10:00:00	37.2	12
2	P2	CHW2	RP2	2021-01-02 09:00:00	38.0	6
3	P3	CHW3	RP3	2021-01-03 11:00:00	36.5	18
4	P4	CHW4	RP4	2021-01-04 08:00:00	37.1	24

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

Source: Gartner

750175_C

Gartner

Challenges and Potential Pitfalls

- Data quality - synthetic data:
    - not reflective of actual data
    - unable to maintain correlation and dependencies between data points
- Factual inaccuracy (caused by LLM hallucinations)
- Bias

How to mitigate this?

- Proper Simulation of Real-World Conditions
- Validation against Real Data
- Maintaining Diversity
- Continuous Monitoring and Feedback
- Involve Subject Matter Experts

About the Speaker

CREATE. COLLABORATE. INNOVATE.

WOMEN
IN AI

MALAYSIA

Dr. LEE MEI SIN

- Founding member of Women in AI Malaysia
- AI Research Fellow in Monash University

✉ meisin@womeninai.co

in linkedin.com/in/meisinlee/

Q & A