

How (not?) to use Large Language Models

Pablo Mosteiro

Utrecht University: Welfare, Participation and Citizenship in a Digital World

2025-09-25



Universiteit Utrecht

Table of Contents

1 What is an LLM

2 Ethics

3 Alternatives

4 Application

What is an LLM



Universiteit Utrecht

How are they trained?

- What do you think is the next word that I will...?



Universiteit Utrecht

How are they trained?

- What do you think is the next word that I will WRITE?



How are they trained?

- What do you think is the next word that I will WRITE?
- What do you think is the next word that I will SAY?



How are they trained?

- What do you think is the next word that I will WRITE?
- What do you think is the next word that I will SAY?
- How did you guess?



LLMs as coding agents

```
a = [2, 4, 5]
```



Universiteit Utrecht

LLMs as coding agents

```
a = [2, 4, 5]  
# Compute the mean of the list
```



LLMs as coding agents

```
a = [2, 4, 5]  
# Compute the mean of the list  
Can the LLM auto-complete?
```



LLMs as bug fixers

```
# Create a dictionary with elements in the name of a person
name_parts = {"given name": "Consolacion", "surname": "Valenzuela"}
# Add a middle name
name_parts.append("middle name": "Celestina")
```

LLMs as bug fixers

```
# Create a dictionary with elements in the name of a person
name_parts = {"given name": "Consolacion", "surname": "Valenzuela"}
# Add a middle name
name_parts.append("middle name": "Celestina")
```

```
File "<stdin>", line 1
    name_parts.append("middle name": "Celestina")
               ^
SyntaxError: invalid syntax
```

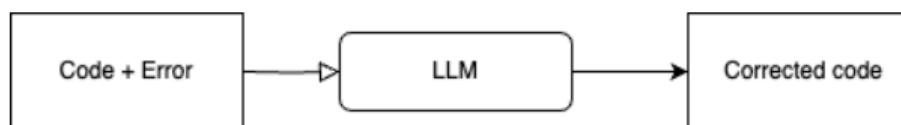
>>>

LLMs as bug fixers

```
# Create a dictionary with elements in the name of a person
name_parts = {"given name": "Consolacion", "surname": "Valenzuela"}
# Add a middle name
name_parts.append("middle name": "Celestina")
```

```
File "<stdin>", line 1
    name_parts.append("middle name": "Celestina")
                           ^
SyntaxError: invalid syntax
```

```
>>>
```



Read more: <https://github.com/microsoft/generative-ai-for-beginners>

How (or why) **not** to use LLMs



Universiteit Utrecht

Training data



World ▾ Business ▾ Markets ▾ Sustainability ▾ Legal ▾ Breakingviews ▾ Technology ▾ Investigations More ▾

My News



Sign In

Subscribe

OpenAI must face part of Intercept lawsuit over AI training

By Blake Brittain

February 20, 2025 9:45 PM GMT+1 · Updated 4 days ago



OpenAI logo is seen in this illustration taken May 20, 2024. REUTERS/Dado Ruvic/Illustration Purchase Licensing Rights [\[link\]](#)

Training data

“Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar”

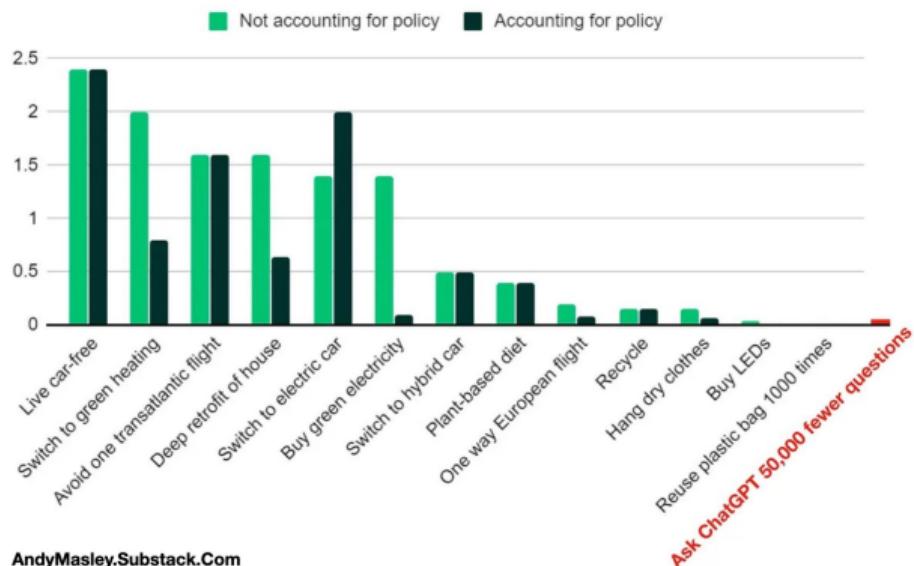
OpenAI: Achiam, J. et al (2024). *GPT-4 Technical Report*.

<https://arxiv.org/abs/2303.08774>



Environment

Figure 4. Tonnes of CO₂ avoided by selected personal lifestyle decisions accounting for government policy



AndyMasley.Substack.Com
Original climate graph taken from Founders Pledge

1

¹Source (January 2025):

<https://open.substack.com/pub/andymasley/p/individual-ai-use-is-not-bad-for>

Labor conditions

Latest Local News • Live Shows ...

CBS NEWS

"When it gets to the day before payday, they close the account and say that you violated a policy," Kanyugi said.

Employees say they have no recourse or even a way to complain.

The company told 60 Minutes that any work done "in line with our community guidelines was paid out." In March, as workers started complaining publicly, Remotasks abruptly shut down in Kenya, locking all workers out of their accounts.

The mental toll of AI training

Workers say some of the projects for Meta and OpenAI also caused them mental harm. Wambalo was assigned to train AI to recognize and weed out pornography, hate speech and excessive violence from social media. He had to sift through the worst of the worst content online for hours on end.

"I looked at people being slaughtered," Wambalo said. "People engaging in sexual activity with animals. People abusing children physically, sexually. People committing suicide."

Berhane Gebrekidan thought she'd been hired for a translation job, but she said what she ended up doing was reviewing content featuring dismembered bodies and drone attack victims.

2

²Source (November 2024):

<https://www.cbsnews.com/news/ai-work-kenya-exploitation-60-minutes/>

Effect on cognition

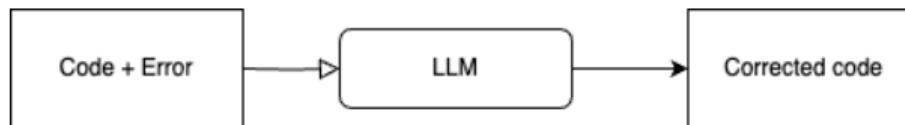
Results

Cultural modulation of neural mechanisms asserts that technological interactions significantly influence brain development and cognitive abilities. The use of artificial intelligence chatbots like ChatGPT for cognitive offloading may lead to underemployment of specific cognitive faculties, inhibiting their full maturation. This phenomenon is particularly relevant in the context of executive functions, where reliance on artificial intelligence for problem-solving can reduce cognitive effort and lead to long-term cognitive changes.¹³

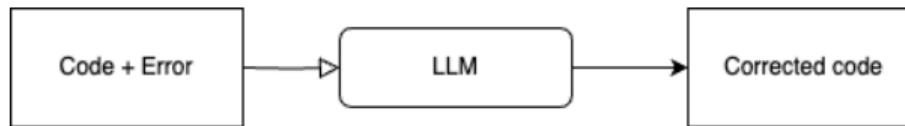
3

³Source (June 2024): Dubey et al. Redefining Cognitive Domains in the Era of ChatGPT: A Comprehensive Analysis of Artificial Intelligence's Influence and Future Implications. Med Res Arch. 2024 Jun;12(6):5383. doi: 10.18103/mra.v12i6.5383

Uploading restricted data



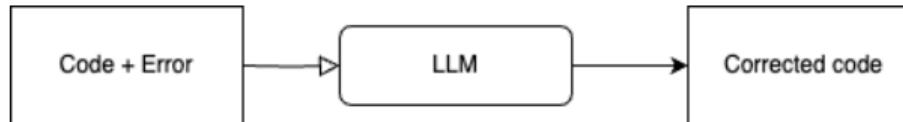
Uploading restricted data



- What if error contains sensitive information?



Uploading restricted data



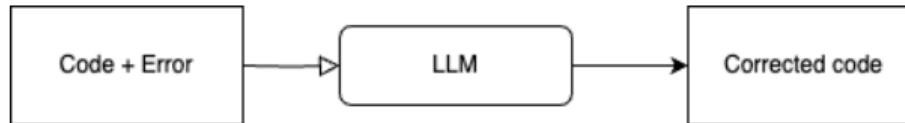
- What if error contains sensitive information?

```
import pandas as pd

def print_selected_municipalities():
    df = pd.read_csv('data/dutch_municipalities.csv', sep=';')
    for income, munic, prov in df[['avg_household_income_2012', 'municipality', 'province']].values:
        if int(income) > 40000:
            print(munic, ':', prov)
```



Uploading restricted data



- What if error contains sensitive information?

```
import pandas as pd

def print_selected_municipalities():
    df = pd.read_csv('data/dutch_municipalities.csv', sep=';')
    for income, munic, prov in df[['avg_household_income_2012', 'municipality', 'province']].values:
        if int(income) > 40000:
            print(munic, ':', prov)

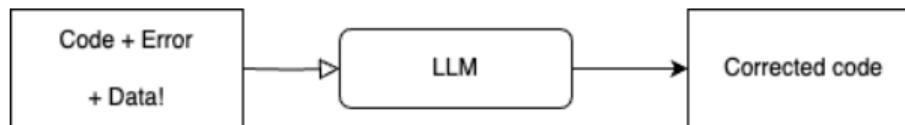
ValueError
Cell In[9], line 1
----> 1 print_selected_municipalities()

Cell In[8], line 6, in print_selected_municipalities()
    4 df = pd.read_csv('data/dutch_municipalities.csv', sep=';')
    5 for income, munic, prov in df[['avg_household_income_2012', 'municipality', 'province']].values:
----> 6     if int(income) > 40000:
    7         print(munic, ':', prov)

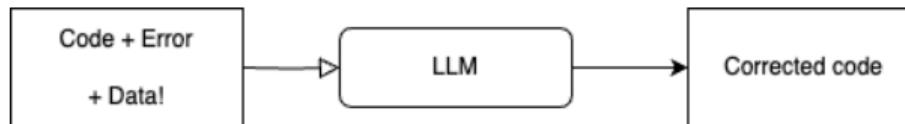
ValueError: cannot convert float NaN to integer
```



Uploading restricted data



Uploading restricted data



- May be illegal!



Conclusions

- Training probably used texts without permission



Conclusions

- Training probably used texts without permission
- Training relied on questionable labor conditions



Conclusions

- Training probably used texts without permission
- Training relied on questionable labor conditions
- Overuse of LLMs can cause cognitive deterioration



Conclusions

- Training probably used texts without permission
- Training relied on questionable labor conditions
- Overuse of LLMs can cause cognitive deterioration
- Uploading your data maybe illegal / against regulations



Conclusions

- Training probably used texts without permission
- Training relied on questionable labor conditions
- Overuse of LLMs can cause cognitive deterioration
- Uploading your data maybe illegal / against regulations
- And yet...



Convenience



iStock

Credit: takayib

Convenience



Universiteit Utrecht

What is an LLM
oooooo

Ethics
oooooooooooo

Alternatives
●oooooooo

Application
oooooooooooo

Alternatives



Universiteit Utrecht

HuggingChat

The screenshot shows the HuggingChat web application interface. At the top, there's a navigation bar with a back button, forward button, refresh button, and a URL bar containing "huggingface.co/chat/". Below the URL bar is a header with a yellow speech bubble icon and the text "HuggingChat". The main content area has a dark background.

Header:

- HuggingChat v0.9.4**
- DeepSeek R1 is now available!**
- Try it out!**

Current Model: meta-llama/Llama-3.3-70B-Instruct

Examples:

- Write an email from bullet list
- Code a snake game
- Assist in a task

Input Field: Ask anything

Model Information: Model: meta-llama/Llama-3.3-70B-Instruct - Generated content may be inaccurate or false.

Sidebar (Left Side):

- Login
- Theme
- Models (11)
- Assistants
- Tools (New)
- Settings
- About & Privacy

Switch models

X

Models

- meta-llama/Llama-3.3-70B-Instruct Active
- Qwen/Qwen2.5-72B-Instruct
- CohereForAI/c4ai-command-r-plus-08-2024
- deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
- nvidia/Llama-3.1-Nemotron-70B-Instruct-HF
- Qwen/QwQ-32B-Preview
- Qwen/Qwen2.5-Coder-32B-Instruct
- meta-llama/Llama-3.2-11B-Vision-Instruct
- NousResearch/Hermes-3-Llama-3.1-8B
- mistralai/Mistral-Nemo-Instruct-2407
- microsoft/Phi-3.5-mini-instruct

Assistants

My Assistants

meta-llama/Llama-3.3-70B-Instruct

Ideal for everyday use. A fast and extremely capable model matching closed source models' capabilities. Now with the latest Llama 3.3 weights!

[Model page](#) [Model website](#) [API Playground](#) [Copy direct link to model](#)

[New chat](#)

System Prompt

Unfortunately, ...



We are cooking something new, please stay tuned...

 Discuss on the Hub

[Privacy Policy](#)

Unfortunately, ...



We are cooking something new, please stay tuned...

 Discuss on the Hub

Privacy Policy

- BUT...

Using an LLM locally (no server)

The screenshot shows the LM Studio 0.2.27 application window. At the top, there's a navigation bar with links to lmstudio.ai, Twitter, GitHub, Discord, Terms of Use, and Export App. Logos. Below the navigation bar is a search bar with placeholder text "Search for models by keyword or paste any HuggingFace repo URL ...". A sidebar on the left lists various features: Welcome to LM Studio!, Release Notes (v0.2.27), Search (Search and download compatible model files), AI Chat (Chat with local LLMs fully offline), Multi Model (Load and prompt multiple local LLMs simultaneously), Local Server (Run an OpenAI-like HTTP server on localhost), My Models (Manage your downloaded models), and Join LM Studio's Discord Server to discuss models, prompts, workflows and more. The main content area displays two model cards: "Meta AI" (Llama 3.1 8B Instruct) and "Microsoft Research" (Phi 3 mini 4K Instruct). Both cards show file size (4.92 GB and 2.39 GB respectively), performance metrics (Small & Fast), and download links. At the bottom, there's a footer with links to Google DeepMind, gemma2, and Stability AI, along with a Model Downloads section showing 0 downloads completed.

Versatile platform

- OpenRouter.ai



Universiteit Utrecht

Versatile platform

- OpenRouter.ai
- But before we do that...



Universiteit Utrecht

Versatile platform

- OpenRouter.ai
- But before we do that...
- Choose your poison carefully / Know what you choose



Versatile platform

- OpenRouter.ai
- But before we do that...
- Choose your poison carefully / Know what you choose
- Do not share private data



Versatile platform

- OpenRouter.ai
- But before we do that...
- Choose your poison carefully / Know what you choose
- Do not share private data
- Consider if you really need this



A practical application: Sexism detection



Universiteit Utrecht

What is an LLM
ooooo

Ethics
oooooooooooo

Alternatives
ooooooo

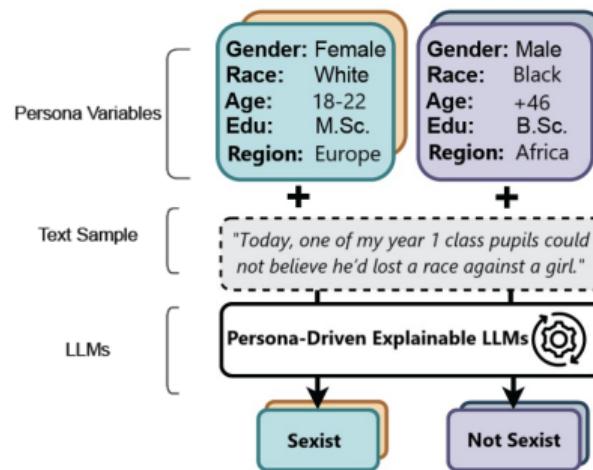
Application
oo●ooooo

Problem

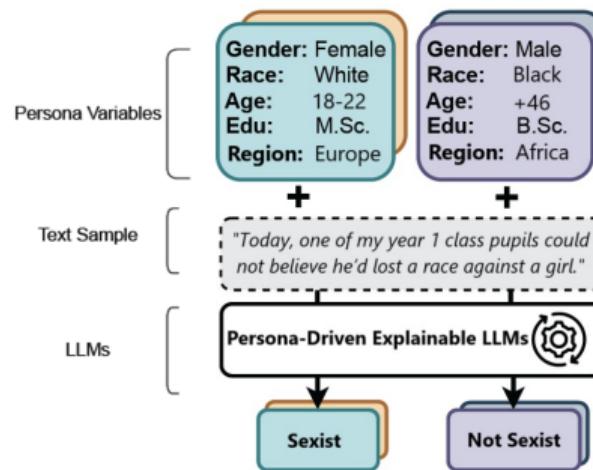


Universiteit Utrecht

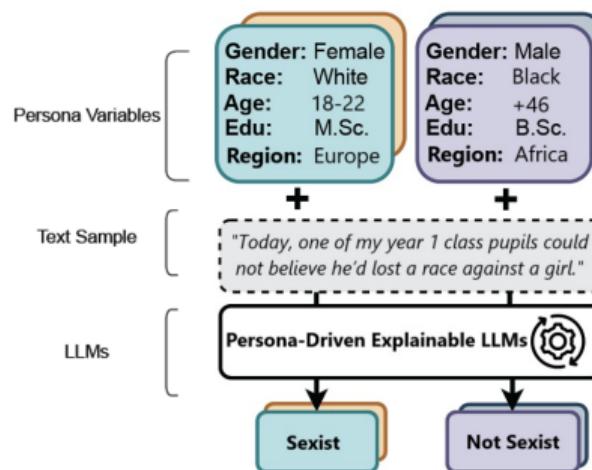
Problem



Problem



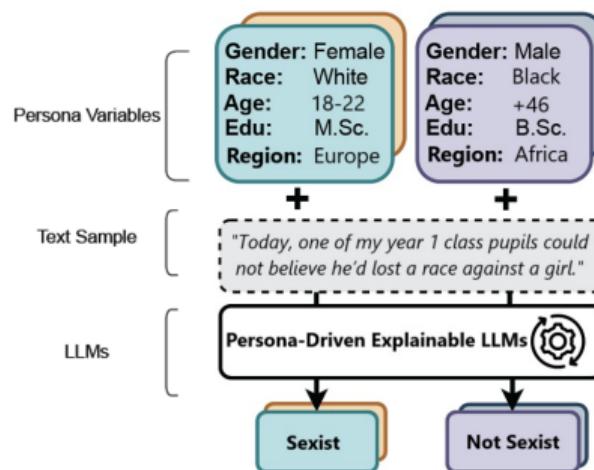
Problem



- Reliable annotations are key to building strong NLP models



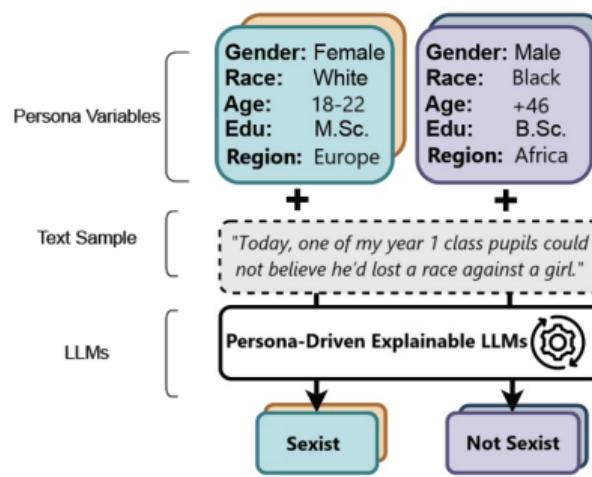
Problem



- Reliable annotations are key to building strong NLP models
- Some levels of disagreement are inevitable, particularly in subjective tasks



Problem



- Reliable annotations are key to building strong NLP models
- Some levels of disagreement are inevitable, particularly in subjective tasks
- **This study: role of the annotator's demographics features and text content in labeling decisions**



The paper

Assessing the Reliability of Annotations

In the Context of LLMs Predictions and Explanations

Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro Romero, Massimo Poesio, Ayoub Bagheri, Anastasia Giachanou

In A. Faleńska, C. Basta, M. Costa-jussà, K. Stańczak, & D. Nozza (Eds), *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 92–104). Association for Computational Linguistics.

<https://doi.org/10.18653/v1/2025.gebnlp-1.9>



Universiteit Utrecht

The paper

Assessing the Reliability of Annotations

In the Context of LLMs Predictions and Explanations

Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro Romero, Massimo Poesio, Ayoub Bagheri, Anastasia Giachanou



Data

- We used data from the EXIST 2024 challenge — the sexism detection tasks.

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → Europe, America, Africa, Asia, and the Middle East.



Data

- We used data from the EXIST 2024 challenge — the sexism detection tasks.
- We focused on Task 1—classifying tweets as sexist or not.

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → Europe, America, Africa, Asia, and the Middle East.



Data

- We used data from the EXIST 2024 challenge — the sexism detection tasks.
- We focused on Task 1—classifying tweets as sexist or not.
- Tweets in both English and Spanish

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → Europe, America, Africa, Asia, and the Middle East.



Universiteit Utrecht

Data

- We used data from the EXIST 2024 challenge — the sexism detection tasks.
- We focused on Task 1—classifying tweets as sexist or not.
- Tweets in both English and Spanish
- Each tweet in the dataset was annotated by six individuals.

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → Europe, America, Africa, Asia, and the Middle East.



Data

- We used data from the EXIST 2024 challenge — the sexism detection tasks.
- We focused on Task 1—classifying tweets as sexist or not.
- Tweets in both English and Spanish
- Each tweet in the dataset was annotated by six individuals.
- The annotators' demographic features include:

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → Europe, America, Africa, Asia, and the Middle East.



Our objectives

- Goal 1: Analyze the impact of demographic factors on annotation in the sexism detection task.



Our objectives

- Goal 1: Analyze the impact of demographic factors on annotation in the sexism detection task.
- Goal 2: Evaluate the potential of GenAI models to replace human annotators.



What is an LLM
oooooo

Ethics
oooooooooooo

Alternatives
ooooooo

Application
oooooooo●○

Terminology and prerequisites



Universiteit Utrecht

What is an LLM
oooooo

Ethics
oooooooooooo

Alternatives
ooooooo

Application
oooooooo●○

Terminology and prerequisites



Universiteit Utrecht

Terminology and prerequisites

- SHAP → highlighted (important) terms



Terminology and prerequisites

- SHAP → highlighted (important) terms
- Four *personas*:



Terminology and prerequisites

- SHAP → highlighted (important) terms
- Four *personas*:

Demographics -> Highlighting	Yes	No
Yes	GenPXAI	GenXAI
No	GenP	GenAI



WARNING

