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

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?

Universiteit Utrecht

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?

Universiteit Utrecht

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?

Universiteit Utrecht

LLMs as coding agents

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

Universiteit Utrecht

LLMs as coding agents

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

Universiteit Utrecht

LLMs as coding agents

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

Can the LLM auto-complete?

Universiteit Utrecht

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

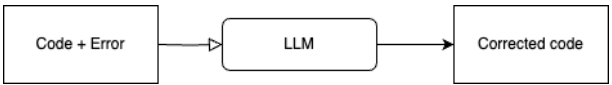
```
# Create a dictionary with elements in the name of a person
name_parts = {"given name":  "Consolacion", "surname":  "Valenzuela"}
# Add a middle name
name_parts.append("middle name":  "Celestina")
```

```
  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

Reuters 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

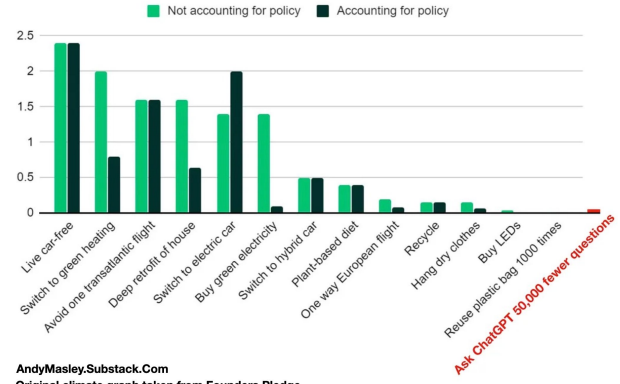
Training data

"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`

Universiteit Utrecht

Environment

Figure 4. Tonnes of CO_2 avoided by selected personal lifestyle decisions accounting for government policy



[1]

[1]Source (January 2025): https://open.substack.com/pub/andymasley/p/individual-ai-use-is-not-bad-for

Labor conditions

"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.

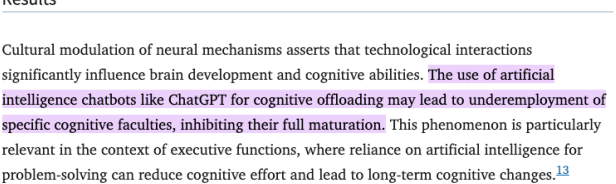
"I looked at people being slaughtered," Wambalo said. "People engaging in sexual activity with animals. People abusing children physically, sexually. People committing suicide."

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

[2] Source (November 2024): `https://www.cbsnews.com/news/ai-work-kenya-exploitation-60-minutes/`

Effect on cognition



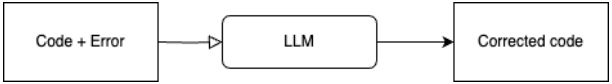
Results

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

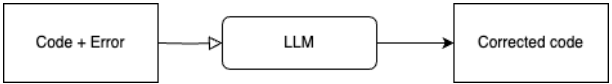
[3]Source (June 2024): Dubey et al. Redefining Cognitive Domains in the Era of ChatGPT: A Comprehensive Analysis of Artificial Intelligence's Influence and Future Implications. Med Res Arch. 2024 Jun;12(6):5383. doi: 10.18103/mra.v12i6.5383

Uploading restricted data



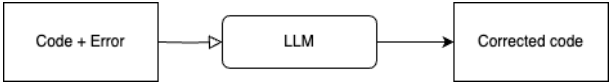
Universiteit Utrecht

Uploading restricted data



- What if error contains sensitive information?

Uploading restricted data



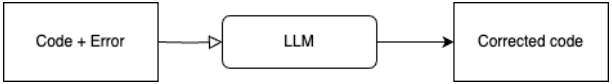
- What if error contains sensitive information?

```python
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)
```

Universiteit Utrecht

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)
```

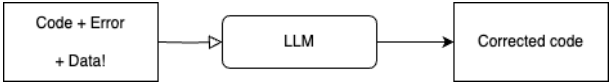
```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[9], line 1
----> 1 print_selected_municipalities()

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
```

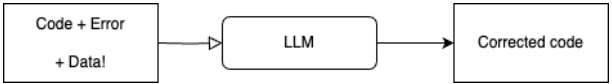
Universiteit Utrecht

Uploading restricted data



Universiteit Utrecht

Uploading restricted data



- May be illegal!

Conclusions

- Training probably used texts without permission

Universiteit Utrecht

Conclusions

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

Universiteit Utrecht

Conclusions

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

Universiteit Utrecht

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

Universiteit Utrecht

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...

Universiteit Utrecht

Convenience

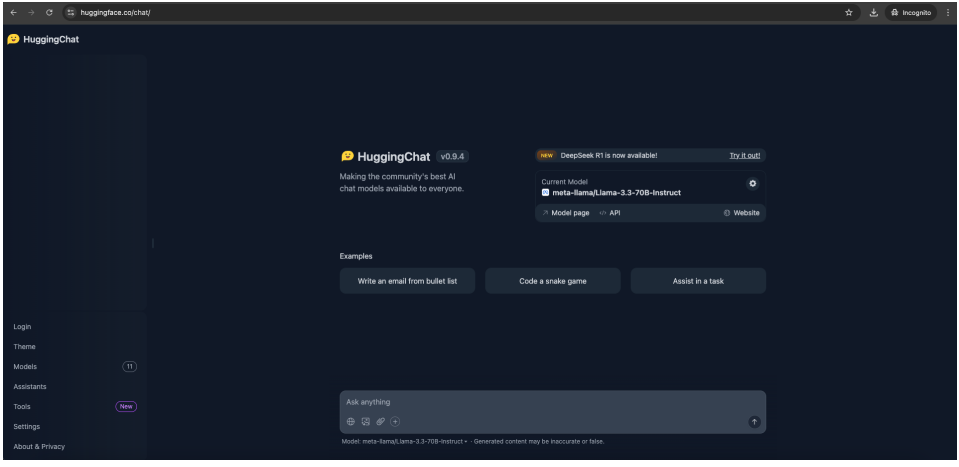
Convenience

Universiteit Utrecht

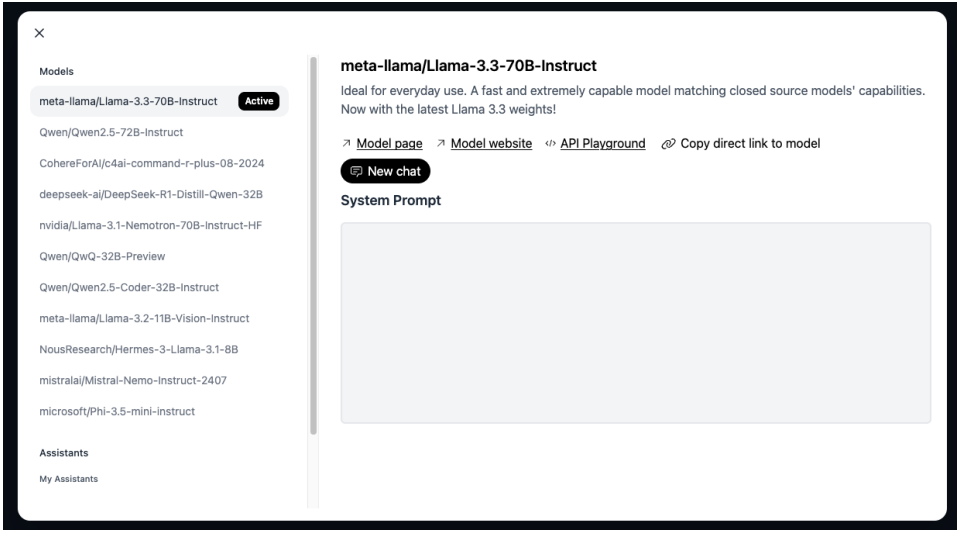
Alternatives

Universiteit Utrecht

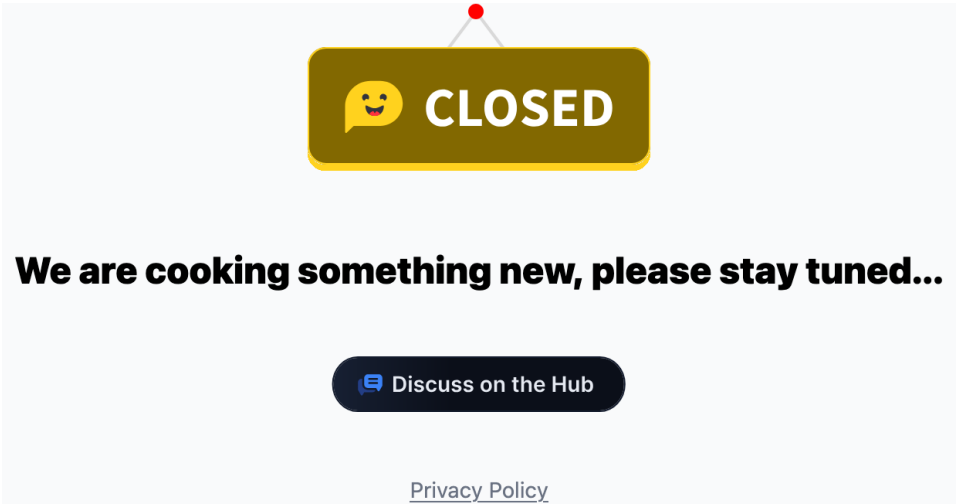
HuggingChat



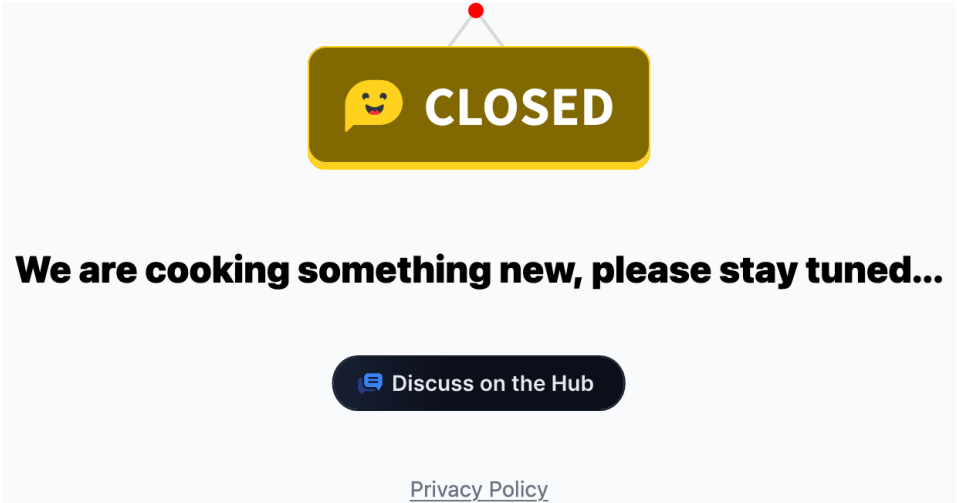
Switch models



Unfortunately, ...

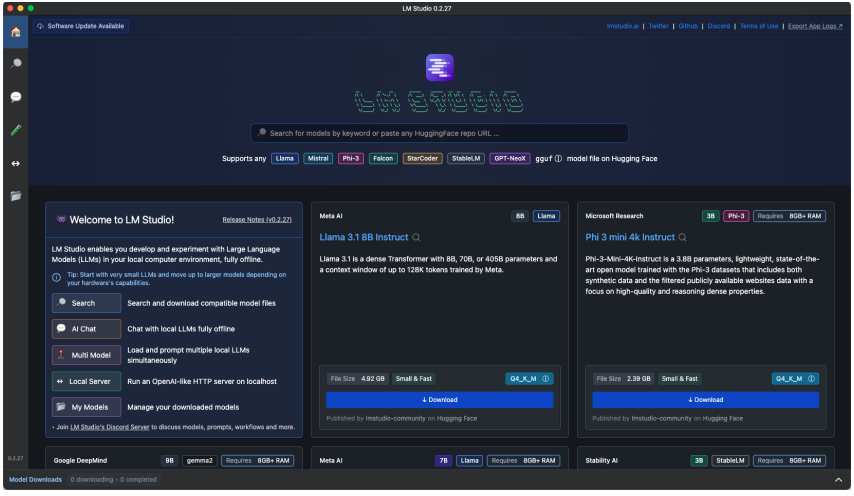


Unfortunately, ...



- BUT...

Using an LLM locally (no server)



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

Universiteit Utrecht

Versatile platform

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

Universiteit Utrecht

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

Universiteit Utrecht

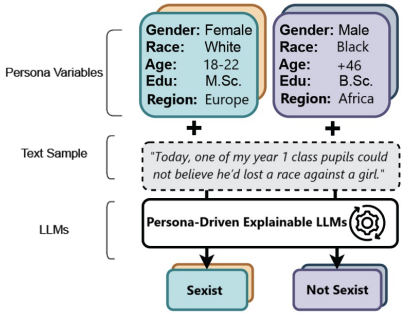
A practical application: Sexism detection

Universiteit Utrecht

Problem

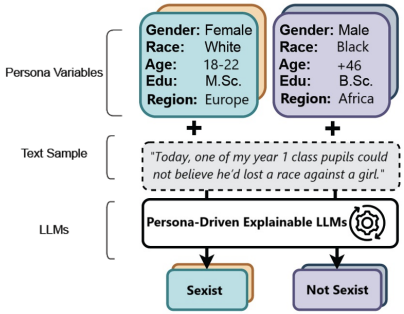
Universiteit Utrecht

Problem



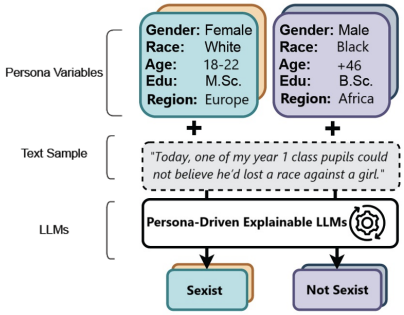
Universiteit Utrecht

Problem



Universiteit Utrecht

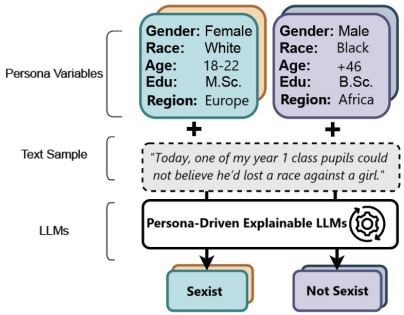
Problem



- Reliable annotations are key to building strong NLP models

Universiteit Utrecht

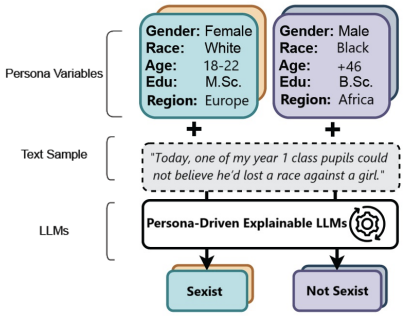
Problem



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

Universiteit Utrecht

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**

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

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.**

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.

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

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.**

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.
- 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.**

Universiteit Utrecht

Our objectives

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

Universiteit Utrecht

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.

Universiteit Utrecht

Terminology and prerequisites

Universiteit Utrecht

Terminology and prerequisites

Universiteit Utrecht

Terminology and prerequisites

- SHAP $\rightarrow$ highlighted (important) terms

Universiteit Utrecht

Terminology and prerequisites

- SHAP $\rightarrow$ highlighted (important) terms
- Four *personas*:

Universiteit Utrecht

Terminology and prerequisites

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

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

Universiteit Utrecht

WARNING