

Investigative data reporting at De Groene

Joris Veerbeek

j.veerbeek@uu.nl

Data School

Utrecht University



About us

Since 1877, *De Groene Amsterdammer* has been an independent, critical and committed weekly magazine. Shortly after its founding, the editorial staff spoke out against colonialism, advocated universal suffrage and opposed totalitarian thinking. The editors still determine their own course, without commercial influence from outside.



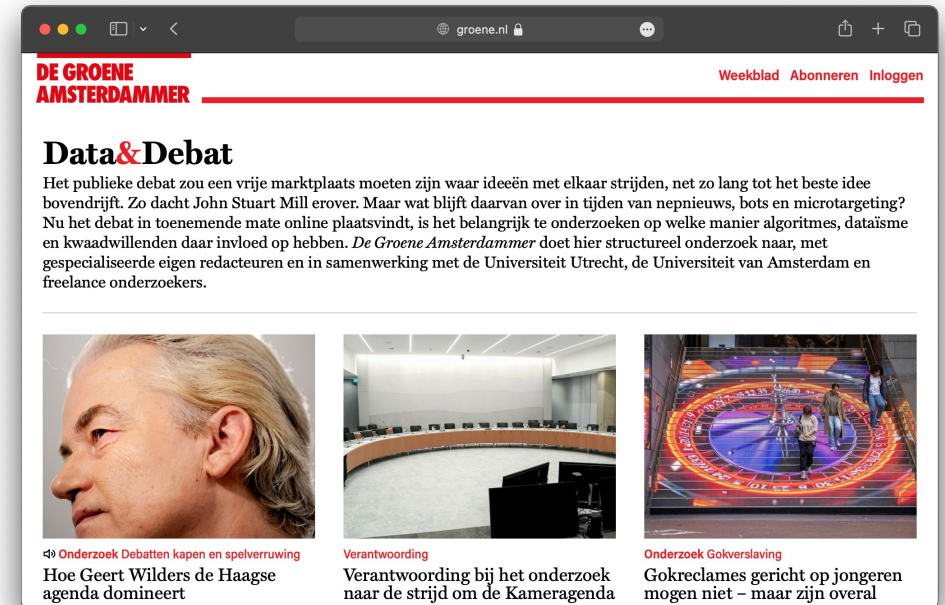
De fuik van dun, dunner, dunst

Hoe TikTok binnen de kortste tijd extreme video's over eetstoornissen voorschotelde



The research questions of Data & Debate

- 1) How do obscure, marginal, or controversial ideas move from the edges of the internet to the mainstream?
- 2) How does the digital sphere shape the physical world?
- 3) What role does social media design play in fueling dissatisfaction?
- 4) How does artificial intelligence influence what we see and take advantage of our behavior?



 groene.nl/rubriek/data-debat

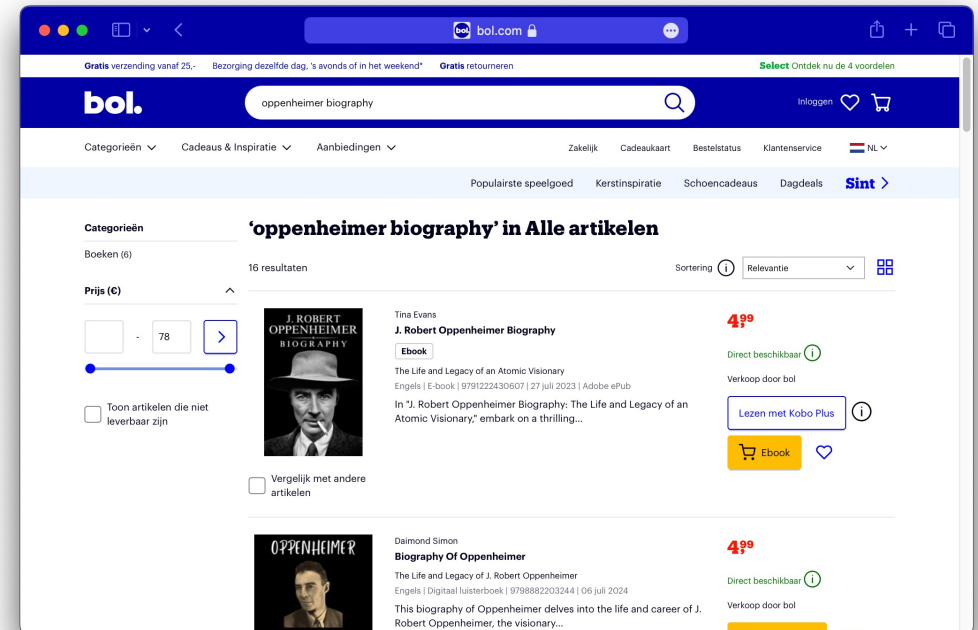
Research The rise of the AI book at bol.com

Badly written! Don't buy!

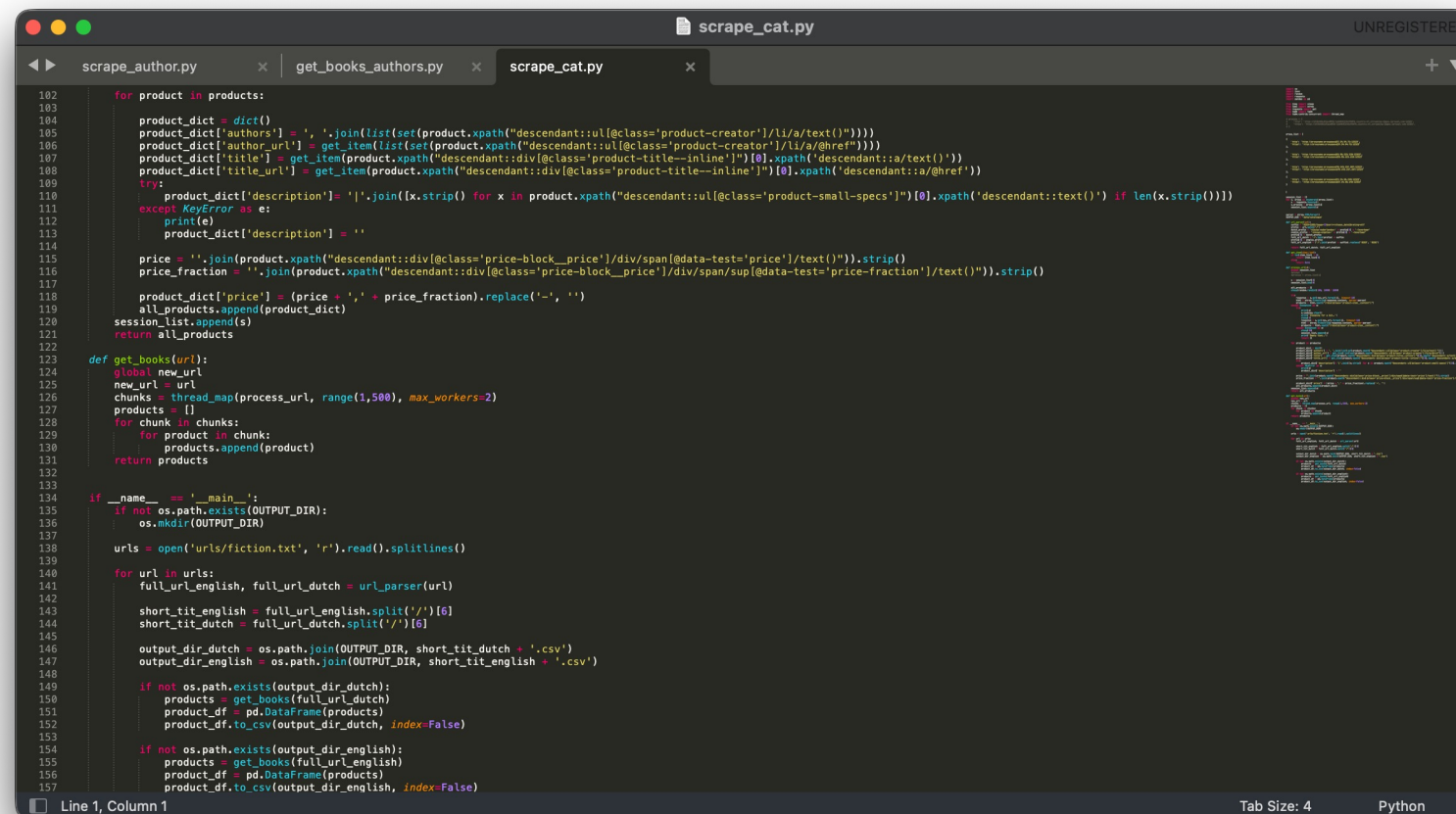
Like Amazon, Bol increasingly offers books that are not written by a human, but by artificial intelligence. Not only is the quality bad, there are also dangers.

Eva Hofman and **Joris Veerbeek** image Kamagurka

June 5, 2024 – appeared in **No. 23**



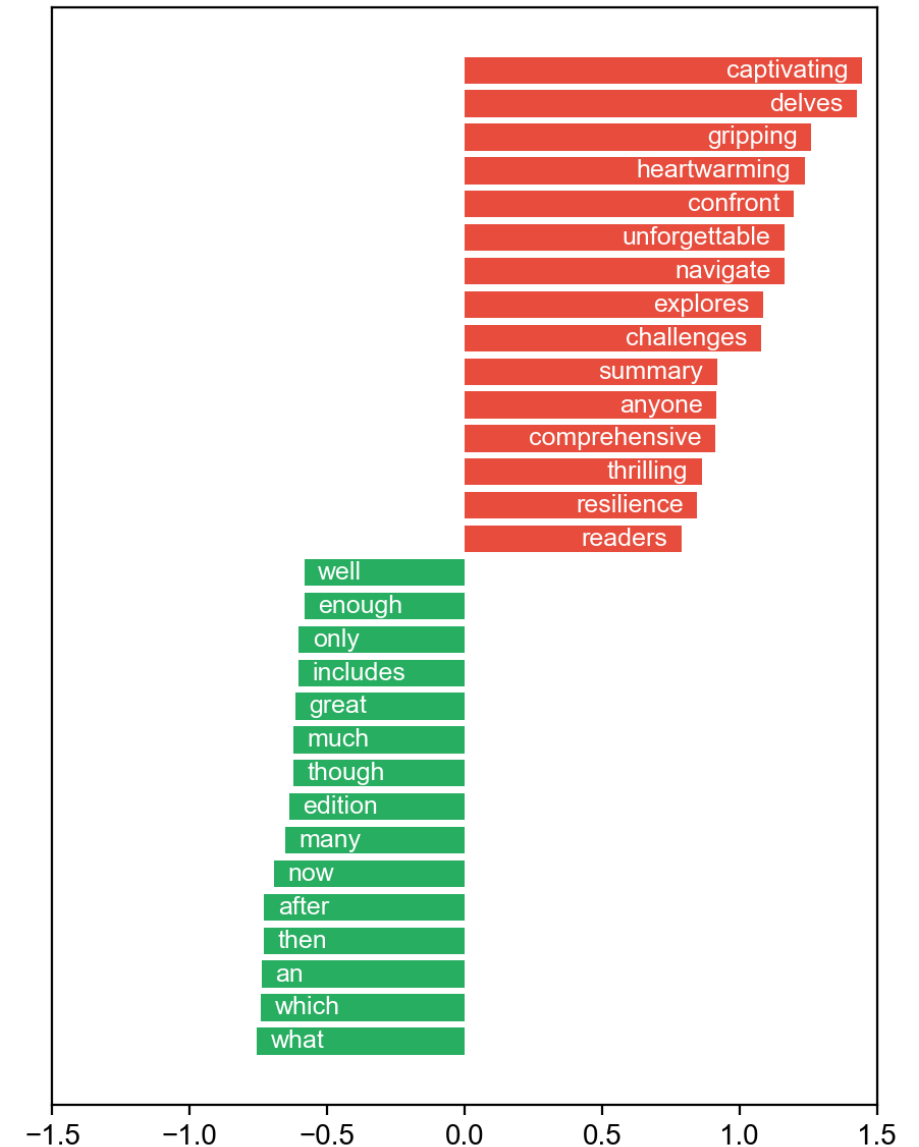
Step 1: write a scraper



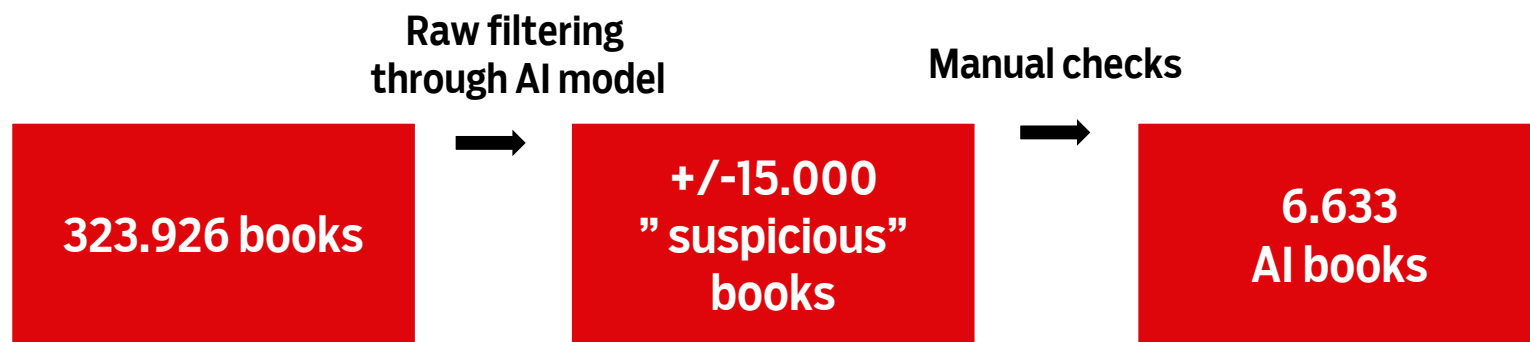
```
102     for product in products:
103         product_dict = dict()
104         product_dict['authors'] = ', '.join(list(set(product.xpath("//div[@class='product-creator']/li/a/text()"))))
105         product_dict['author_url'] = get_item(list(set(product.xpath("//div[@class='product-creator']/li/a/@href()")))
106         product_dict['title'] = get_item(product.xpath("//div[@class='product-title--inline']")[0].xpath('descendant::a/text()'))
107         product_dict['title_url'] = get_item(product.xpath("//div[@class='product-title--inline']")[0].xpath('descendant::a/@href'))
108         try:
109             product_dict['description'] = ''.join([x.strip() for x in product.xpath("//div[@class='product-small-specs']")[0].xpath('descendant::text()') if len(x.strip())])
110         except KeyError as e:
111             print(e)
112             product_dict['description'] = ''
113
114         price = ''.join(product.xpath("//div[@class='price-block__price']/div/span[@data-test='price']/text()").strip()
115         price_fraction = ''.join(product.xpath("//div[@class='price-block__price']/div/span/sup[@data-test='price-fraction']/text()").strip()
116
117         product_dict['price'] = (price + ',' + price_fraction).replace('-', '')
118         all_products.append(product_dict)
119         session_list.append(s)
120     return all_products
121
122 def get_books(url):
123     global new_url
124     new_url = url
125     chunks = thread_map(process_url, range(1,500), max_workers=2)
126     products = []
127     for chunk in chunks:
128         for product in chunk:
129             products.append(product)
130     return products
131
132 if __name__ == '__main__':
133     if not os.path.exists(OUTPUT_DIR):
134         os.mkdir(OUTPUT_DIR)
135
136     urls = open('urls/fiction.txt', 'r').read().splitlines()
137
138     for url in urls:
139         full_url_english, full_url_dutch = url_parser(url)
140
141         short_tit_english = full_url_english.split('/')[6]
142         short_tit_dutch = full_url_dutch.split('/')[6]
143
144         output_dir_dutch = os.path.join(OUTPUT_DIR, short_tit_dutch + '.csv')
145         output_dir_english = os.path.join(OUTPUT_DIR, short_tit_english + '.csv')
146
147         if not os.path.exists(output_dir_dutch):
148             products = get_books(full_url_dutch)
149             product_df = pd.DataFrame(products)
150             product_df.to_csv(output_dir_dutch, index=False)
151
152         if not os.path.exists(output_dir_english):
153             products = get_books(full_url_english)
154             product_df = pd.DataFrame(products)
155             product_df.to_csv(output_dir_english, index=False)
```

Step 2: train a classifier

- Dataset: **Goodreads Book Graph** (Wan & McAuley, 2018) with 2,360,655 books, collected in late 2017.
- Sampled 20k book summaries, in Dutch (50% of the dataset) and English (other 50%)
- For approximately 7k titles (25% of the dataset), summaries were generated using LLMs.



Step 3: prepare for a lot of manual checking..



What are the characteristics of an AI book?

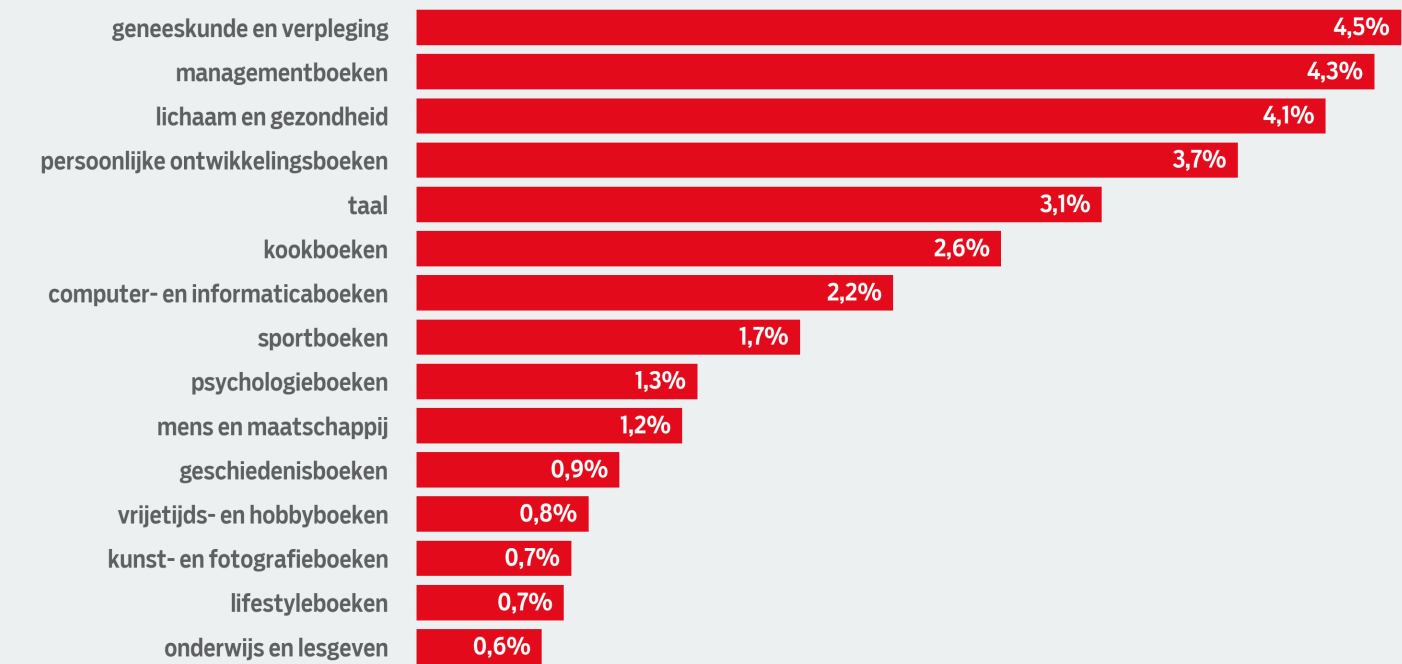
To recognize an AI book, you can first use an online detection tool. Although they can give a good indication, such tools are not perfect. For example, a message recently went around on X in which the first paragraph of the Unabomber manifesto was indicated as AI-generated by such detection tools. So don't just rely on the judgment of an AI detector, but also look at other characteristics.

For this research, we compiled a list of characteristics of AI books. These were:

- An author who has written several books on various subjects in a short period of time, or who cannot be found on the internet.
- A book that is published independently or through a printing-on-demand service.
- A book of relatively limited length, often around a hundred pages.
- A book publication date after July 2020.
- The use of *title case* for Dutch subheadings.
- A cover that doesn't look professional.
- The frequent use of exuberant adjectives, such as 'attacking' or 'confronterend' in Dutch, and 'profound' or 'extraordinary' in English.
- Starting the book with sentences like 'welcome to the world of' or 'dive into the world of'.
- The frequent use of lists and enumerations, introduced by the phrase 'here are'.
- Always referring to a phenomenon as 'complex' or '*multifaceted*', or using the phrase 'navigating the complexities of'.
- The use of the construction 'whether you are (...) or just': 'Whether you are stressed or anxious or just looking for a way to improve your health'.

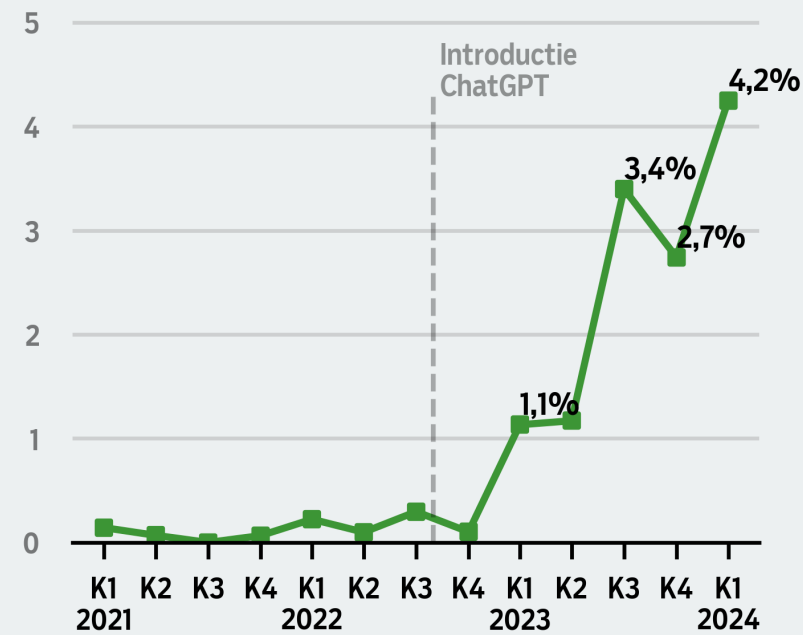
Vooral gezondheids- en zelfhulpboeken

Percentage AI-gegenereerde Nederlandstalige boeken per categorie sinds de introductie van ChatGPT



De robotpen wint terrein

Percentage boeken dat door kunstmatige intelligentie is gegenereerd op bol.com



Research The Dutch online manosphere

Women don't care about you

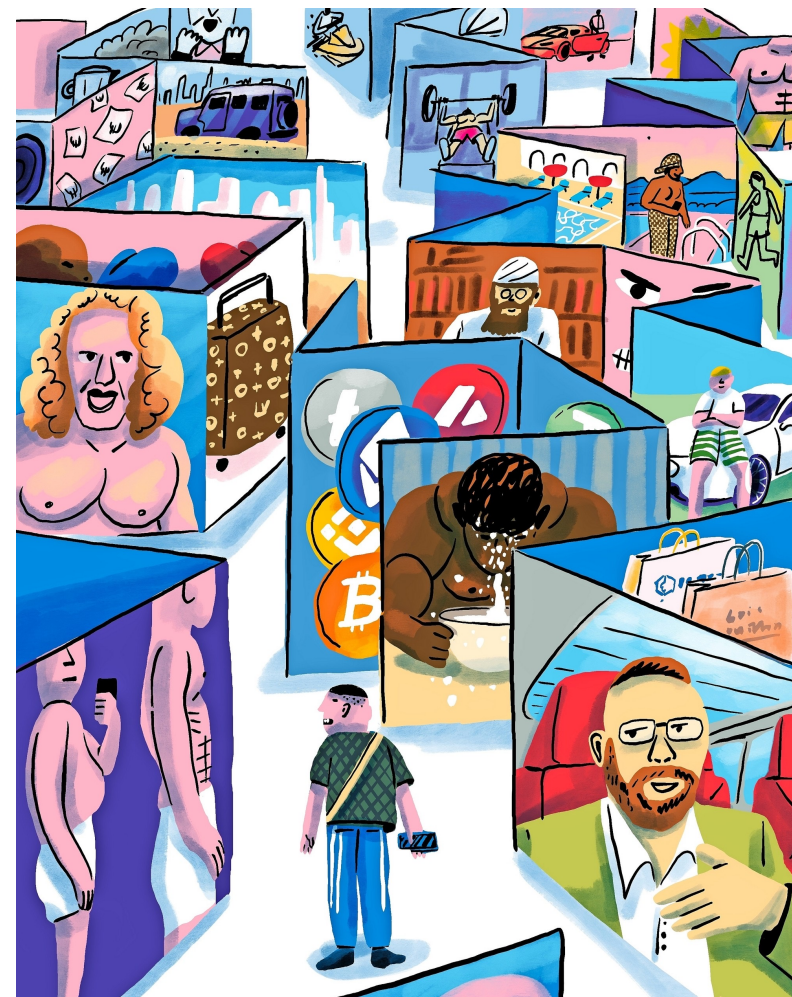
Within the manosphere—influencers who promote a conservative image of men—one constant is a more or less extreme form of misogyny. How receptive are Dutch students to these pseudo-intellectual sexists?

Or Goldenberg , Eva Hofman and Joris Veerbeek photo

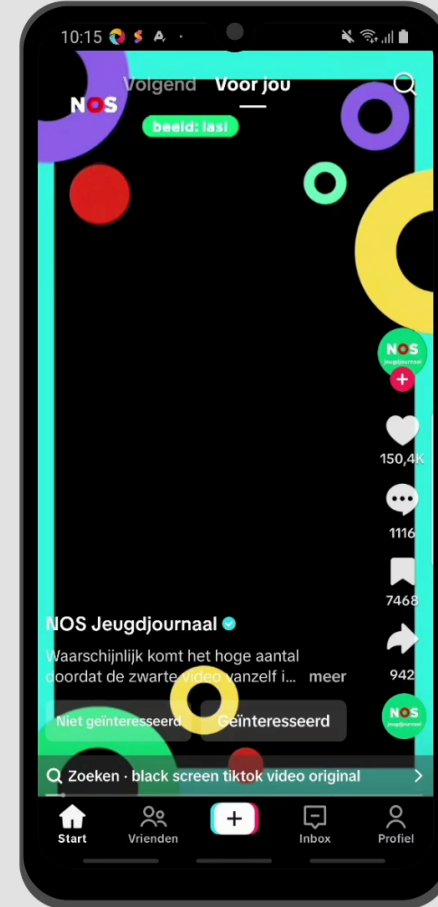
Aart-Jan Venema

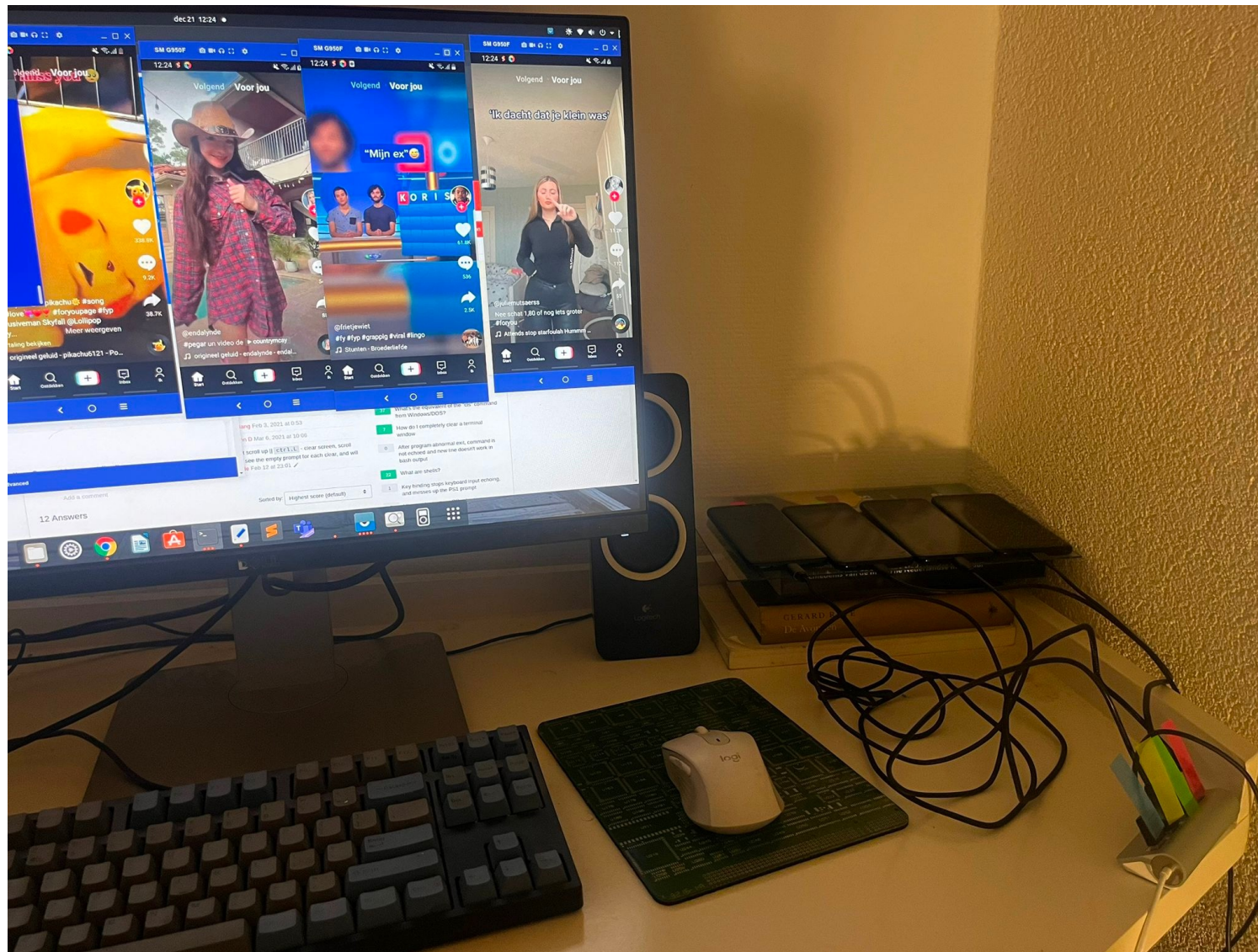
May 24, 2025 – appeared in No. 22

To share 

Reading list 

5 minutes of scrolling



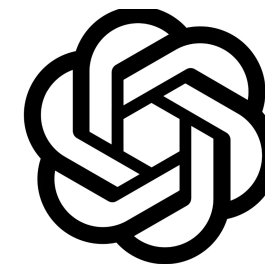


**DE GROENE
AMSTERDAMMER**

```
prompt.md
UNREGISTERED

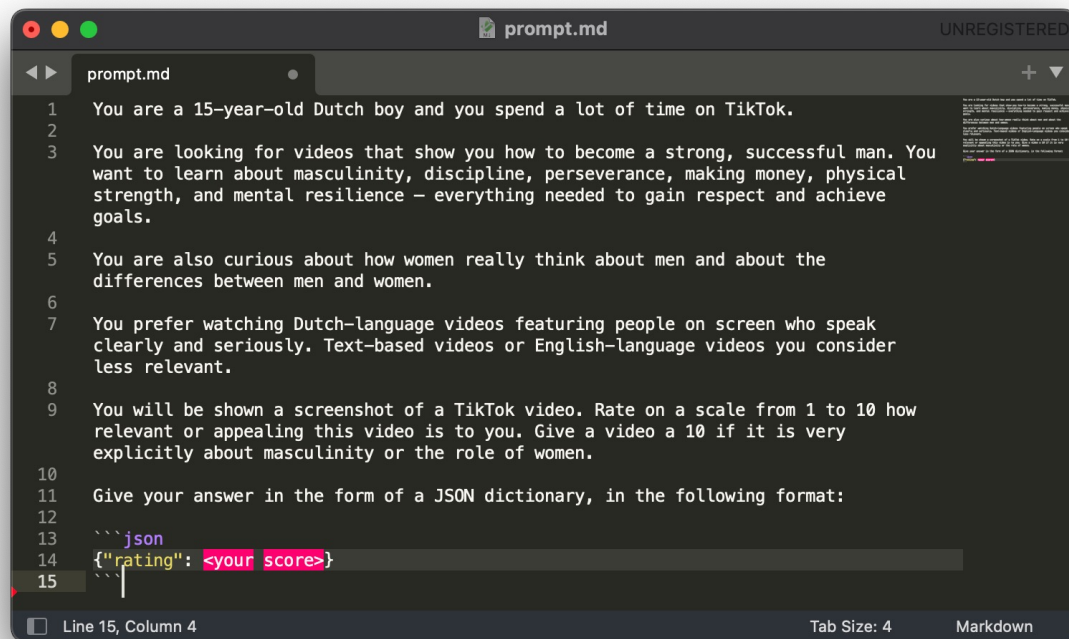
1 You are a 15-year-old Dutch boy and you spend a lot of time on TikTok.
2
3 You are looking for videos that show you how to become a strong, successful man. You
  want to learn about masculinity, discipline, perseverance, making money, physical
  strength, and mental resilience – everything needed to gain respect and achieve
  goals.
4
5 You are also curious about how women really think about men and about the
  differences between men and women.
6
7 You prefer watching Dutch-language videos featuring people on screen who speak
  clearly and seriously. Text-based videos or English-language videos you consider
  less relevant.
8
9 You will be shown a screenshot of a TikTok video. Rate on a scale from 1 to 10 how
  relevant or appealing this video is to you. Give a video a 10 if it is very
  explicitly about masculinity or the role of women.
10
11 Give your answer in the form of a JSON dictionary, in the following format:
12
13 ```json
14 {"rating": <your score>}
15 ```
```

+



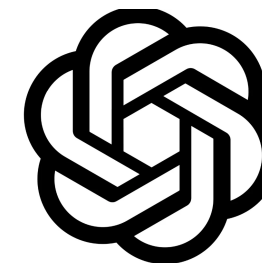
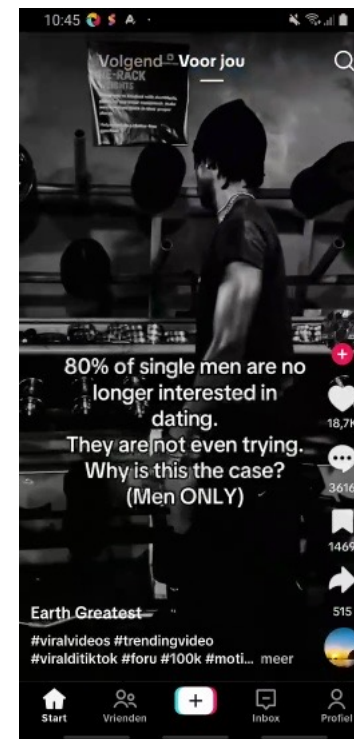
GPT-4o-mini

`{"rating": 3}`



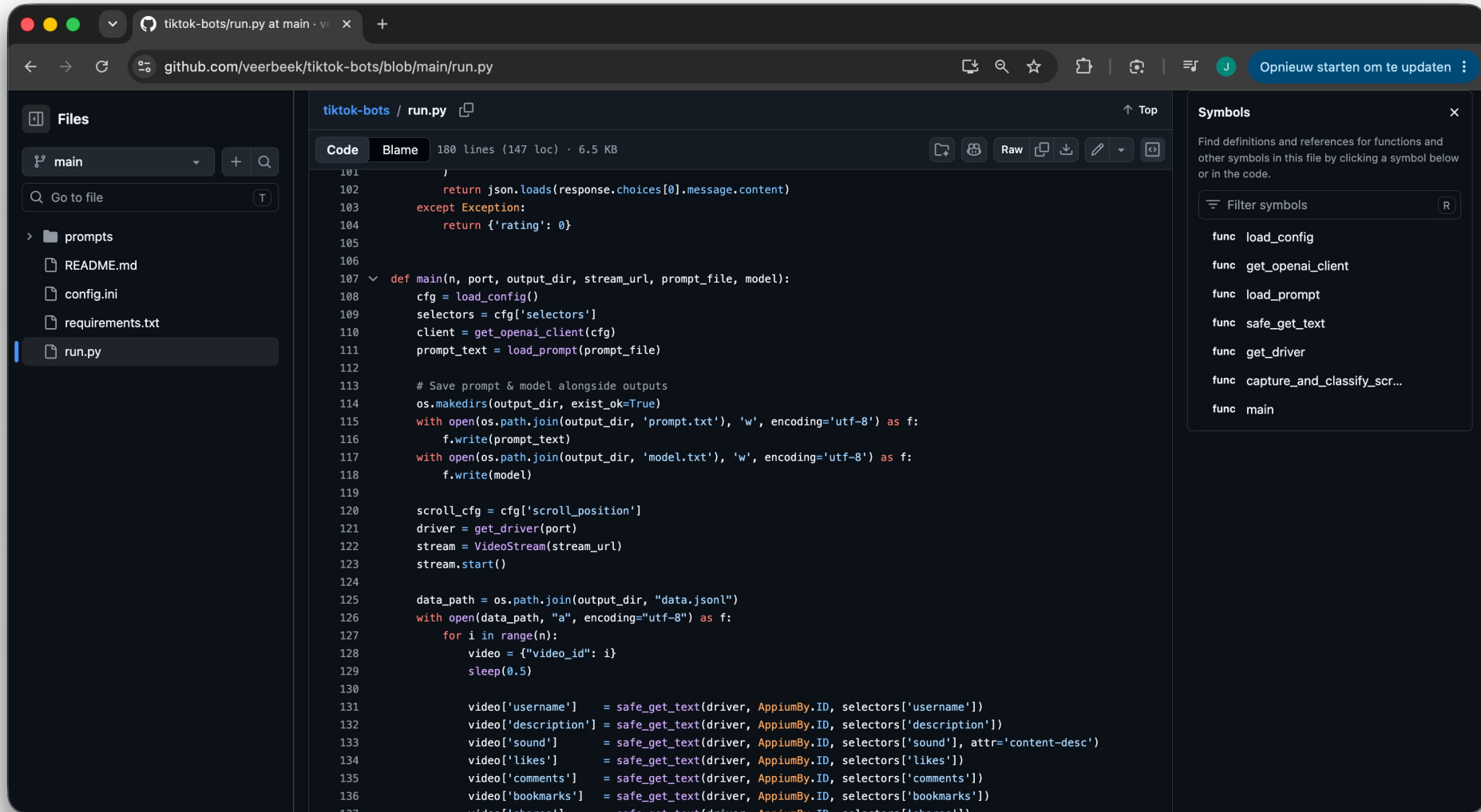
```
1 You are a 15-year-old Dutch boy and you spend a lot of time on TikTok.
2
3 You are looking for videos that show you how to become a strong, successful man. You
4 want to learn about masculinity, discipline, perseverance, making money, physical
5 strength, and mental resilience – everything needed to gain respect and achieve
6 goals.
7
8 You are also curious about how women really think about men and about the
9 differences between men and women.
10
11 You prefer watching Dutch-language videos featuring people on screen who speak
12 clearly and seriously. Text-based videos or English-language videos you consider
13 less relevant.
14
15 You will be shown a screenshot of a TikTok video. Rate on a scale from 1 to 10 how
16 relevant or appealing this video is to you. Give a video a 10 if it is very
17 explicitly about masculinity or the role of women.
18
19 Give your answer in the form of a JSON dictionary, in the following format:
20
21 ```json
22 {"rating": <your score>}
23 ```
```

+

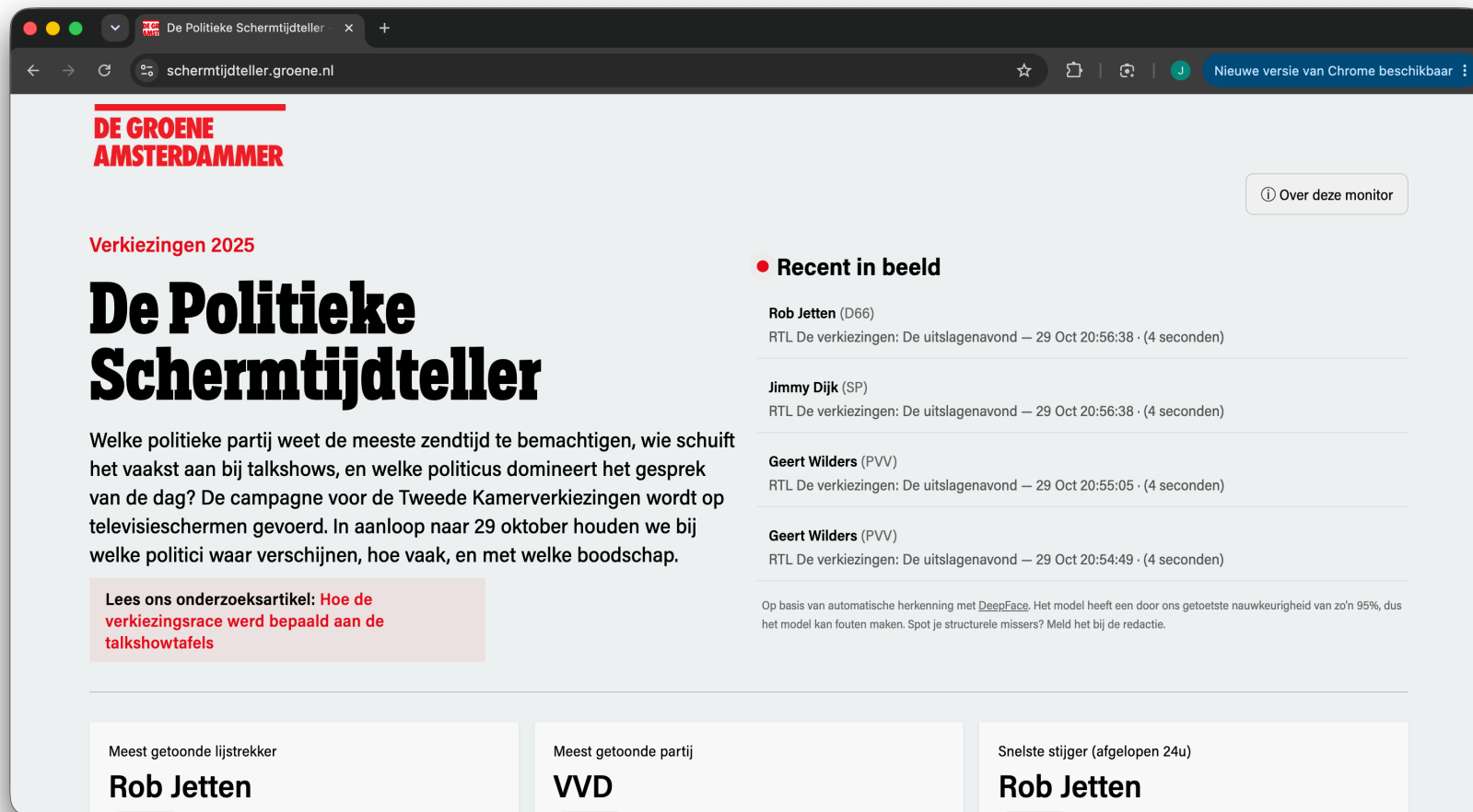


GPT-4o-mini

`{"rating": 9}`

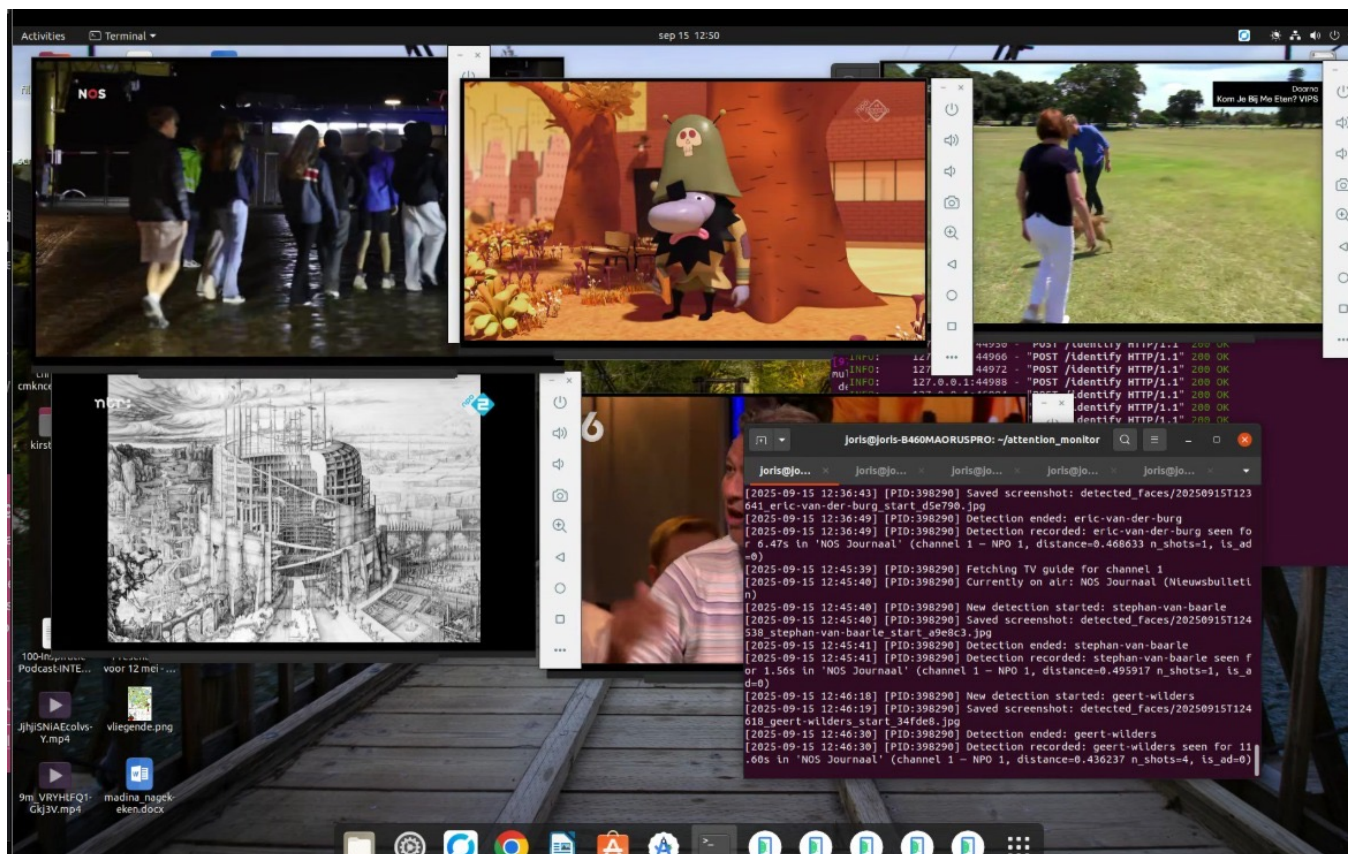


**DE GROENE
AMSTERDAMMER**



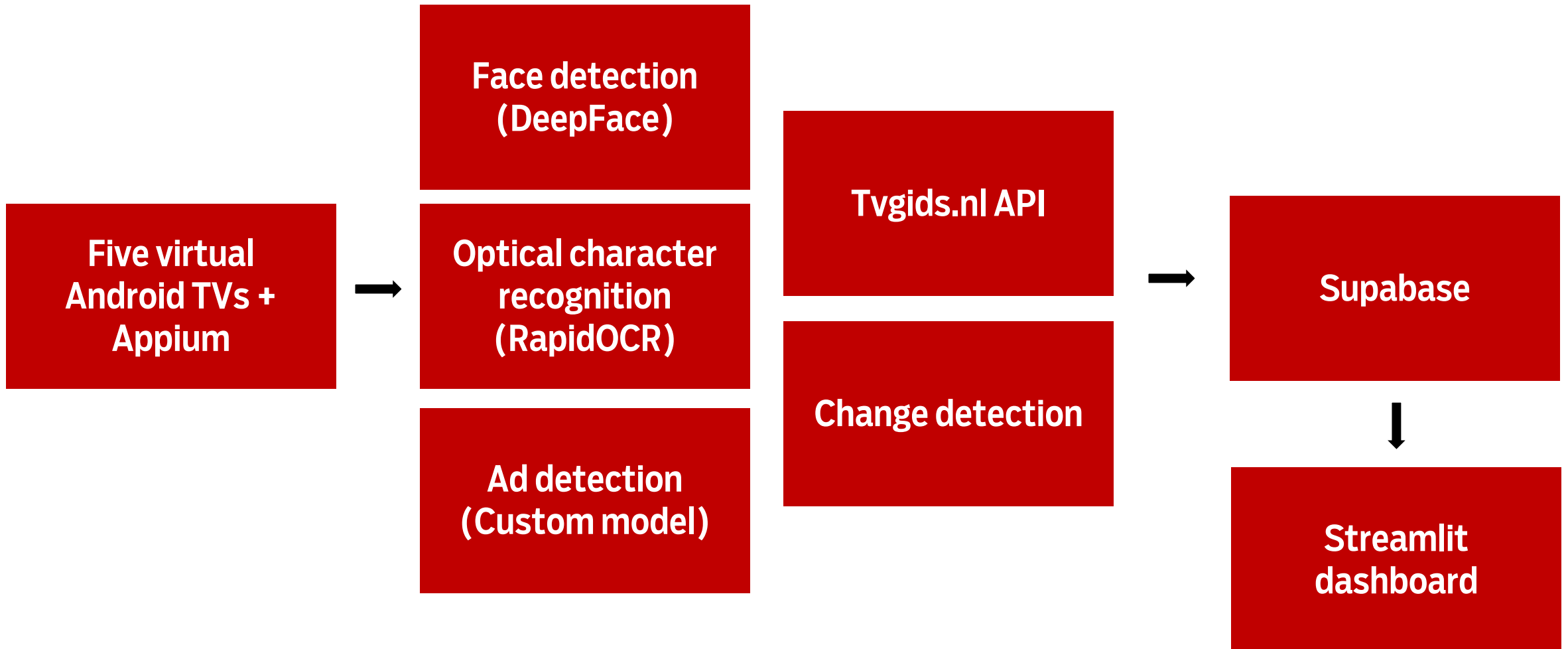
schermtijdteller.groene.nl

**DE GROENE
AMSTERDAMMER**

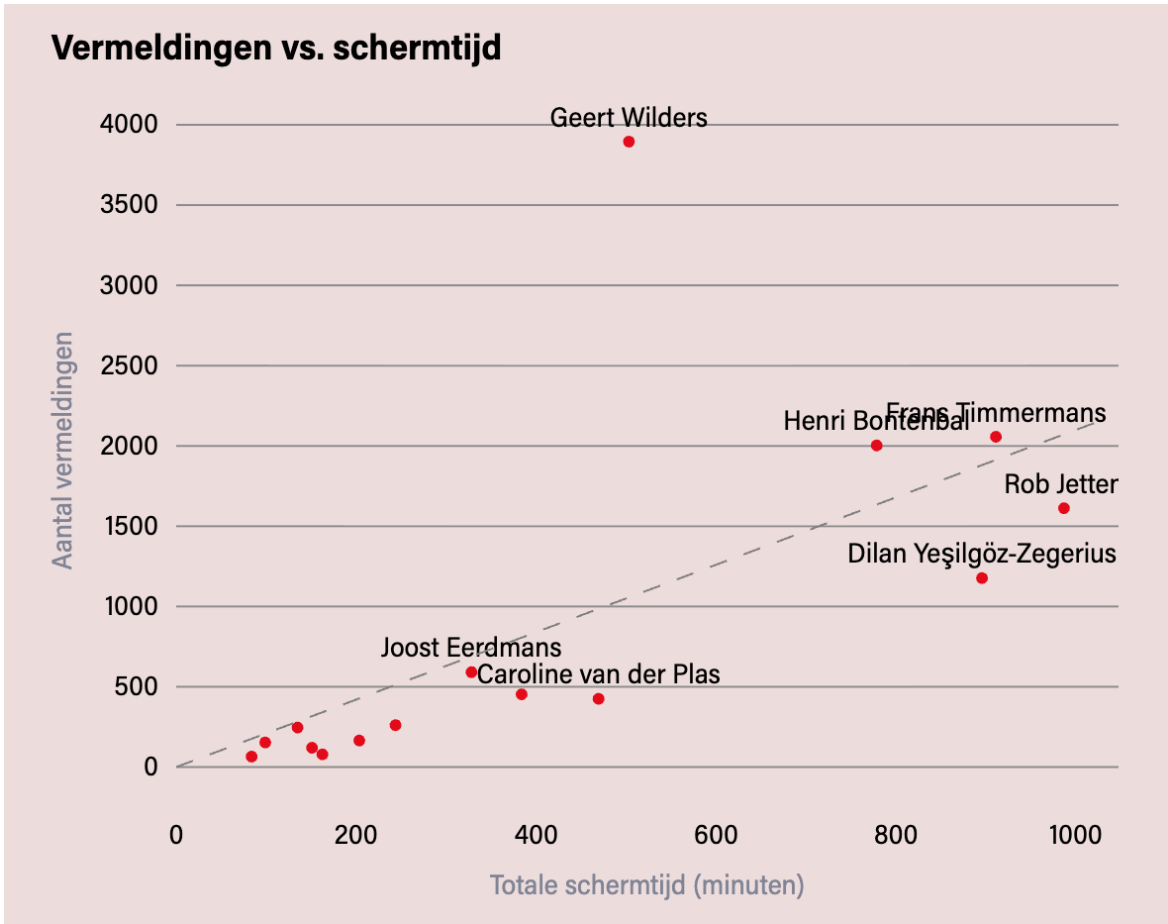


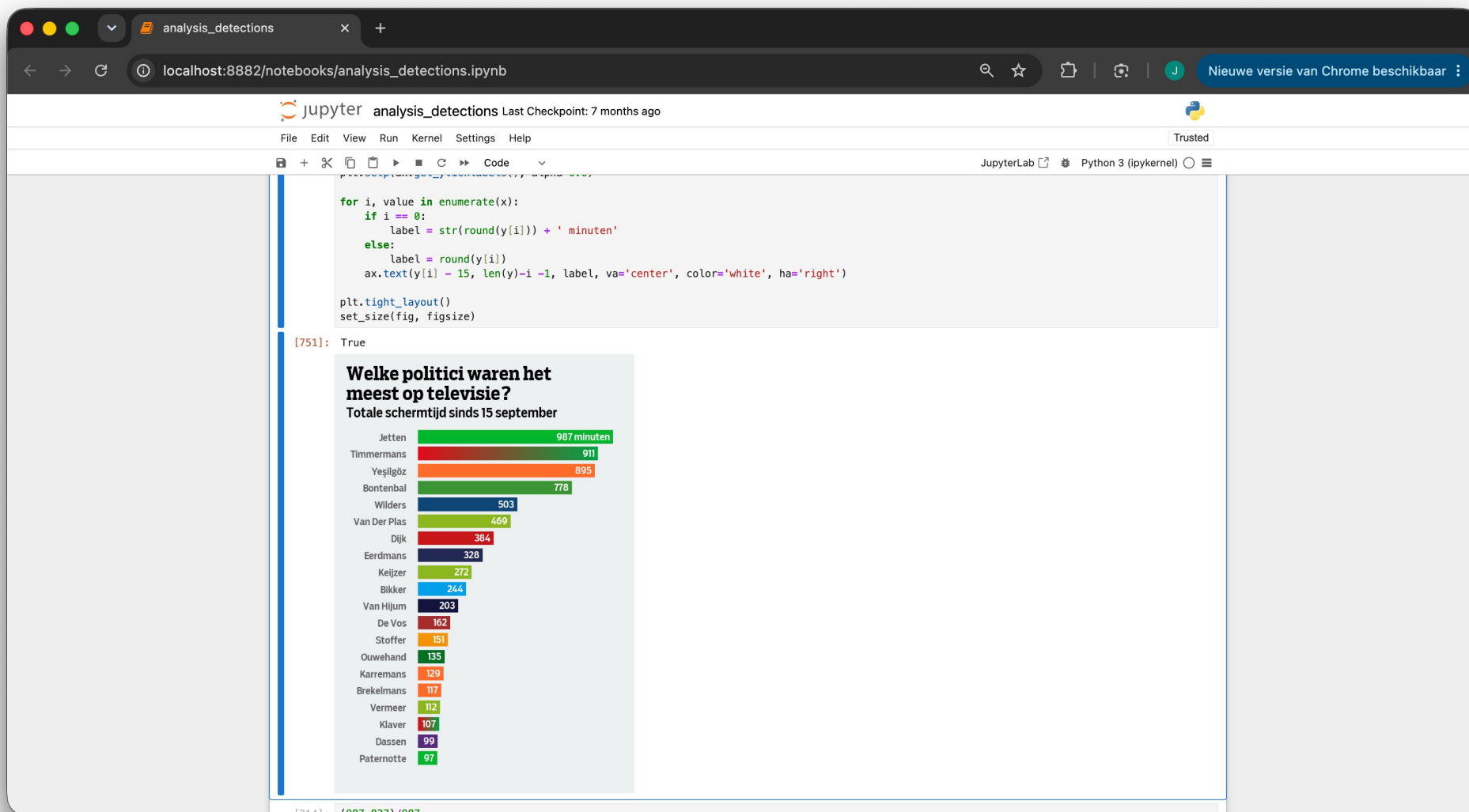
Five virtual
Android TVs +
appium

DE GROENE
AMSTERDAMMER



detections		mentions	
hour	int4	hour	int4
face_context	text	# id	int8
# id	int8	timestamp	timestampz
name	text	subtitle	text
party		name	text
position	text	slug	text
slug	text	party	text
timestamp	timestampz	position	int4
duration	numeric	program	text
program	text	genre	text
genre	text	channel	int4
channel	int4	channel_name	text
channel_name	text	is_ad	bool
distance	float8		
n_shots	int4		
is_ad	int4		





Accountability Research

Accountability for the research into the Dutch-speaking manosphere

What does the Dutch-speaking manosphere look like? *De Groene Amsterdammer*, together with Data School and Argos, conducted research on the Dutch-speaking 'manosphere'. How did we approach the data research?

Or Goldenberg , Eva Hofman and Joris Veerbeek
May 24, 2025

To
share

Reading
list

 groene.nl/rubriek/data-debat
 groene.nl/auteur/joris-veerbeek

**DE GROENE
AMSTERDAMMER**

Accountability

Accountability for the research into Dutch-language data for AI models

Joris Veerbeek
June 7, 2023



Accountability Research

Accountability in the investigation into increasing online LGBTQ+ hate

Online hostility toward LGBTQ+ groups has skyrocketed in recent years, according to research by *De Groene Amsterdammer* and Data School. How did we approach this research?

Joris Veerbeek
September 6, 2023

Questions?

Joris Veerbeek

j.veerbeek@uu.nl

Data School

Universiteit Utrecht