AI Risks in the News:
A Study of Newspaper Reporting on Large Language Models and Artificial Intelligence

Maximilian Weber

Discourse on AI

After the release of ChatGPT

Previous research

- Media coverage for the US (1977-2018): New Economy Boom and starting from 2015 increase in discussion (Sun et al. 2020)
- Different news outlet portray AI differently (Kieslich et al. 2023)
- One study after the release of ChatGPT - focussing on headlines only (Karanouh 2023)

Research questions

- ➔ Identify topics that have been addressed negatively. What **potential risks** are associated with the widespread adoption of AI and LLMs in society?

- ➔ How do various news outlets portray the potential harms of AI and LLMs differently?

Data - Newspaper

Newspaper articles since 01.01.2020 (until Mid August 2023)

Source: Factiva

Currently updating the data - planned to 31.12.2024

Search for articles using Keywords related to Artificial Intelligence, Machine Learning and OpenAI products (e.g. ChatGPT, GPT3.5, GPT4)

Newspapers from UK and US (and Germany)

Newspapers

UK

The Daily Mirror
The Guardian
The Independent
The Sun
The Times

US

New York Times
Wall Street Journal
Washington Post

Keywords

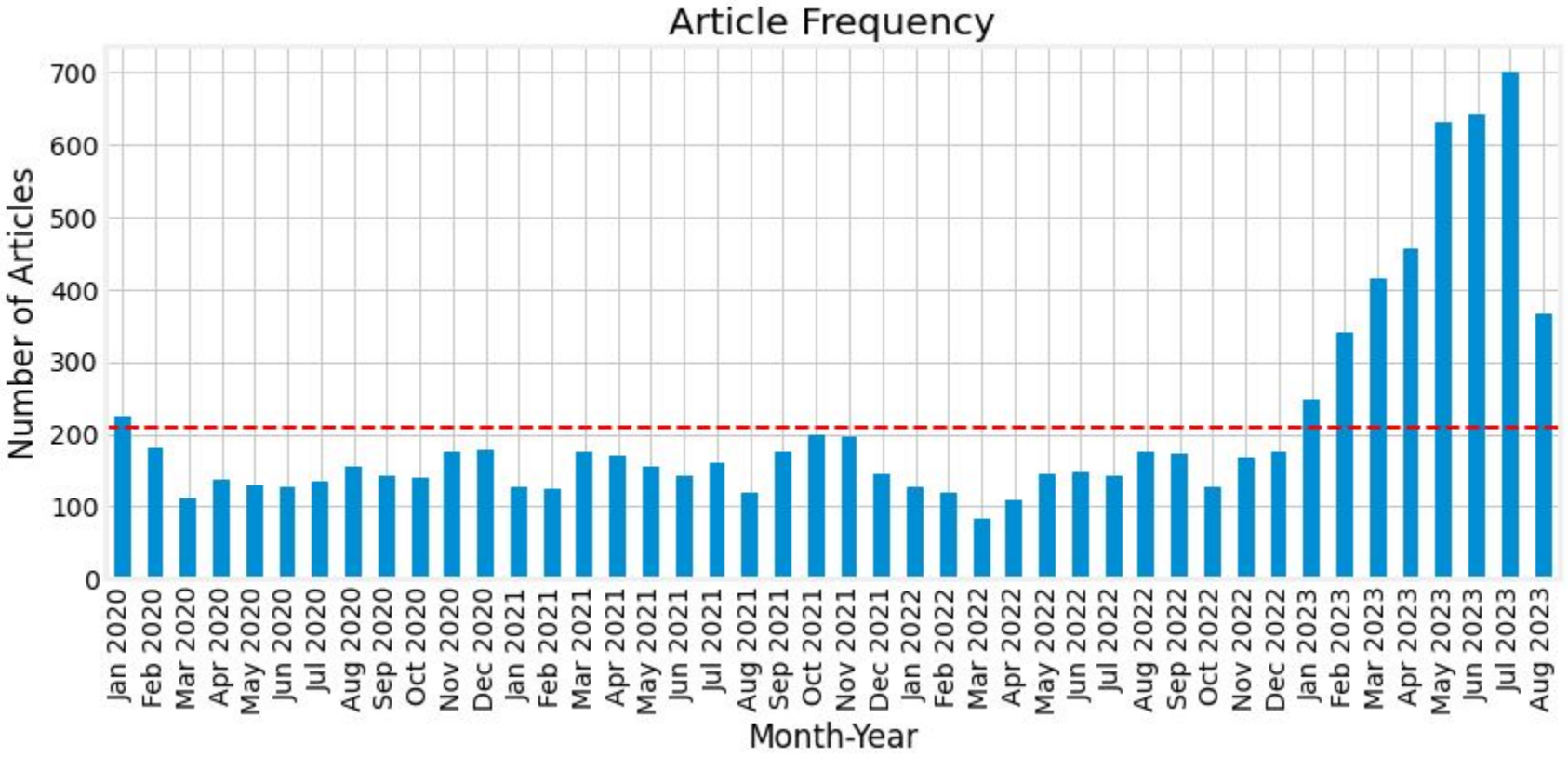
chatgpt or
chat-gpt or
chat gpt or
openai or
open ai or
gpt-4 or
gpt4 or
gpt-3 or
gpt3 or

chatbot* or
artificial* intelligence* or
generative model* or
deep learning or
machine learning or
artificial general intelligence

Top Keywords

keyword	frequency	number of articles
artificial intelligence	10329	7705
chatgpt	4770	1571
chatbot	3308	1314
openai	2653	924
machine learning	1622	1234
...	...	...

N= 9212


Article Frequency
Number of Articles
0
100
200
300
400
500
600
700
Jan 2020
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Sep 2020
Oct 2020
Nov 2020
Dec 2020
Jan 2021
Feb 2021
Mar 2021
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
Sep 2021
Oct 2021
Nov 2021
Dec 2021
Jan 2022
Feb 2022
Mar 2022
Apr 2022
May 2022
Jun 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Month-Year

Newspaper stance UK

The Daily Mirror

center-left or labor-leaning stance

The Independent

centrist political stance and is known for its balanced coverage

The Times

center-right or conservative leaning

Left - Right Newspaper Stance

The Guardian

center-left or liberal leaning

The Sun

right-wing or conservative stance

Newspaper stance US

Annotation

Article -> Paragraphs

9219 Articles

=>

30571 Paragraphs using the presented Keywords + “ AI “

Annotation Guidelines

Objective:

Determine if the given paragraph discusses the harms, risks or negative implications associated with AI models.

Instructions: If the paragraph explicitly or indirectly discusses harms, risks, or negative implications associated with AI models, label it as “yes”. Key words and phrases for this category might be: “danger”, “concern”, “mislead”, “regulate”, “biased”, “ethical issues”, “potentially harmful”, etc. Sometimes, even if AI is discussed in a positive light or neutrally, it can contain a subtle reference to a risk or harm. Be vigilant for these cases. If the paragraph does not discuss any negative implications or harms of AI models, label it as “no”.

I annotated 320 paragraphs

From each newspaper 40 random cases

Examples

Harms/Risks discussed

When people use these models, data are sent back to the developer to enable continued improvements, presenting the potential for an organization to **unintentionally share proprietary or confidential information.** OpenAI disclosed in March that it took ChatGPT temporarily offline because a bug allowed some users to see the titles from a user's chat history.

Not discussed

Technology has been developed by a British company that uses wide-angle cameras and artificial intelligence (AI) to detect pedestrians beside the road and then turn the lights red without the need to touch a button.

Guanaco and Llama 2 - both 13b

“Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters”

I use the 13 billion parameter model

https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

“The pretraining data has a cutoff of September 2022, but some tuning data is more recent, up to July 2023.”

Evaluation on 320 paragraphs

Prompt	No F1-Score	Yes F1-Score	accuracy	macro avg F1-Score	weighted avg F1-Score
Guanaco 13b					
zero shot	0.32	0.47	0.41	0.40	0.36
few shot	0.09	0.45	0.31	0.27	0.19
Llama 2 13b					
zero shot 1	0.83	0.02	0.72	0.43	0.6
zero shot 2	0.82	0.15	0.71	0.49	0.63
one shot	**0.90**	**0.72**	**0.85**	**0.81**	**0.85**
few shot 1	0.24	0.43	0.35	0.34	0.29
few shot 2	0.22	0.44	0.34	0.33	0.28

Best Approach - Llama 2

Using Llama 2 to decide if risks or harms are mentioned

One-Shot Prompt

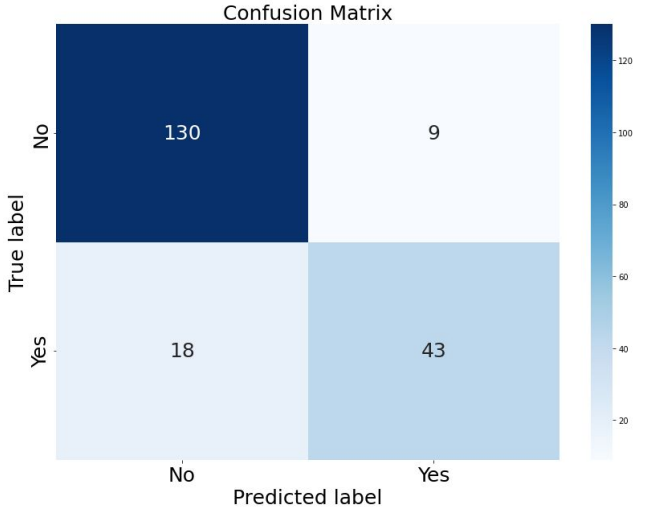
```
def analyze_text(text):
prompt = f""" [INST] <<SYS>> You are an assistant for categorizing paragraphs. <</SYS>> [/INST]
        [INST] I am given a paragraph:  -"While the potential of artificial intelligence is vast
        and exciting, there are considerable risks involved. If not used ethically, AI models can
        perpetuate harmful biases."
        Based on the content of the paragraph, determine if it discusses the potential harms or
        risks of AI models. Make sure to only return the label and nothing more. [/INST] Yes

        [INST] I am given the paragraph:"{text}"
        Does it discuss the potential harms or risks of AI models?
        Only return the label "Yes" or "No". [/INST]"""
```

Best Prompt - Result

Metric	Global	UK	US
F1-Score-Weighted	0.85	0.86	0.83
F1-Score-Macro	0.81	0.83	0.77
Yes - F1-Score	0.72	0.76	0.64

Evaluation on the 320 annotated paragraphs
From each newspaper 40 random cases

UK:



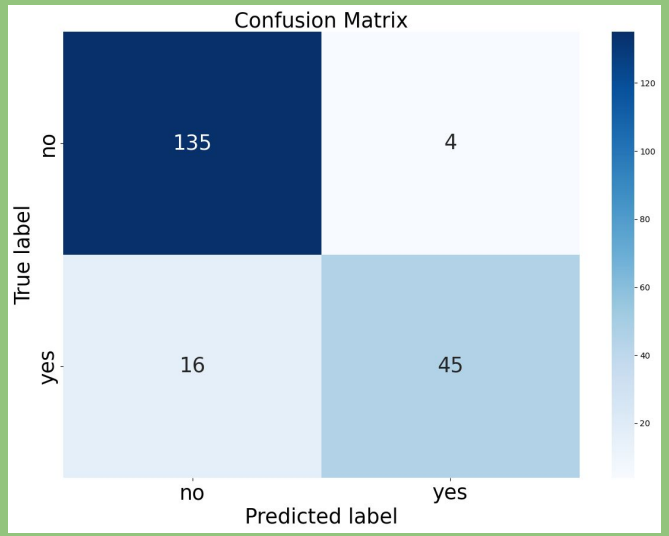
Best Prompt - Result - Llama 3.1

Meta-Llama-3.1-8B-Instruct

Metric	Global	UK	US
F1-Score-Weighted	0.89	0.90	0.88
F1-Score-Macro	0.85	0.87	0.81
Yes - F1-Score	0.78	0.82	0.71

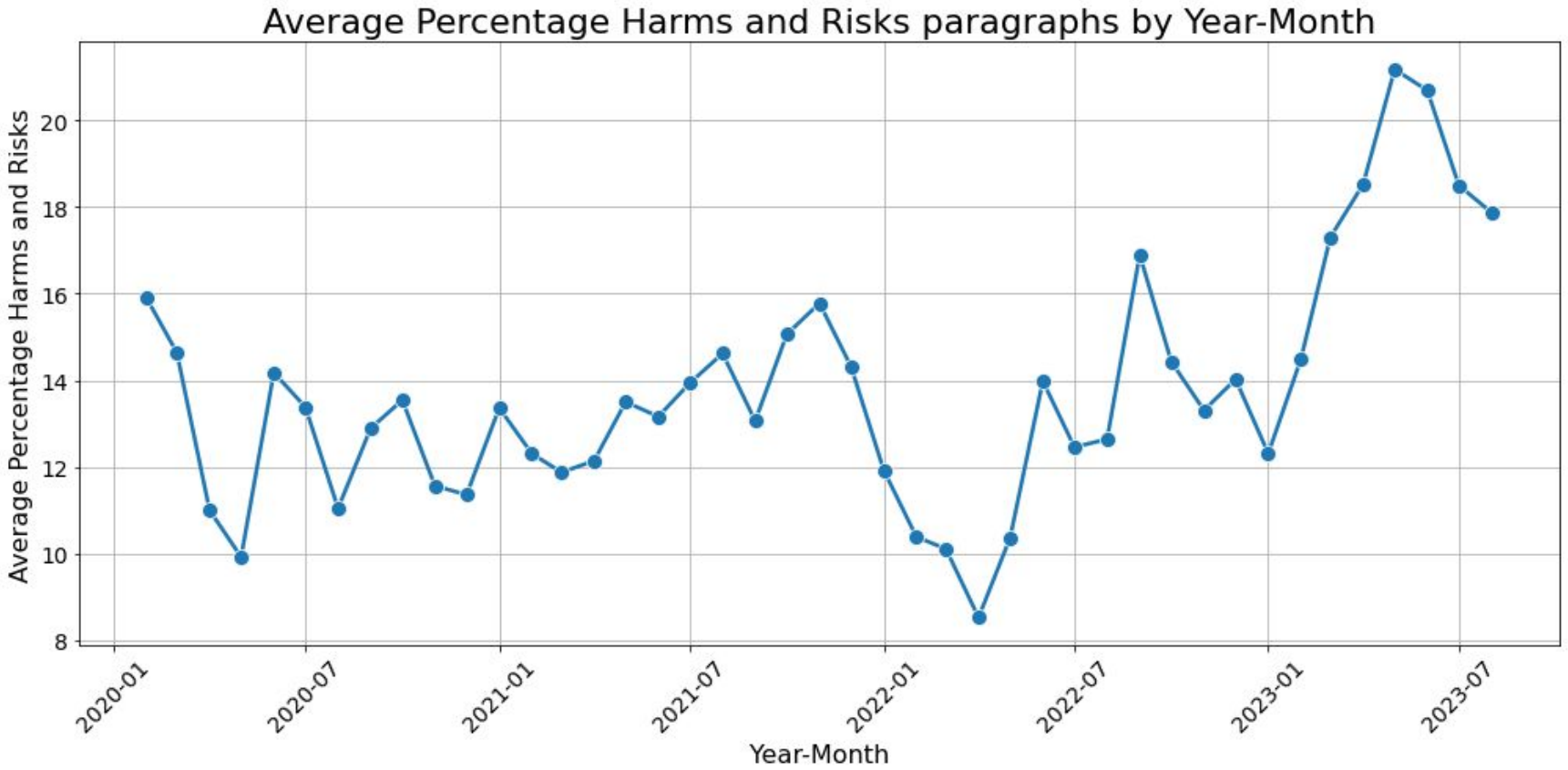
Evaluation on the 320 annotated paragraphs
From each newspaper 40 random cases

UK:



Results

Time trend


Average Percentage Harms and Risks paragraphs by Year-Month
Average Percentage Harms and Risks
20
18
16
14
12
10
8
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
Year-Month

UK

Newspaper	Total articles	Only harms	Only not harms	Percentage only harms	Percentage paragraphs harms
The Daily Mirror	304	17	248	6.42	10.58
The Guardian	2024	181	1272	12.46	**20.81**
The Independent	350	23	245	8.58	16.38
The Sun	450	42	321	11.57	17.50
The Times	2025	142	1570	8.29	13.01

Extract reasons - Llama 2

```
"""<s>[INST] <<SYS>> You are an assistant responsible
for providing reasons why a researcher determined that a
given paragraph was marked as discussing potential
harms, risks, or negative consequences of AI models
<</SYS>>    (EXAMPLE NOT SHOWN)
I am given the paragraph:

   "{text}"

   Based on the content of the paragraph, identify the
potential harms or risks of AI models that are
discussed. Give a short answer (max 100 words): [/INST]
   """
```

results in standardized output

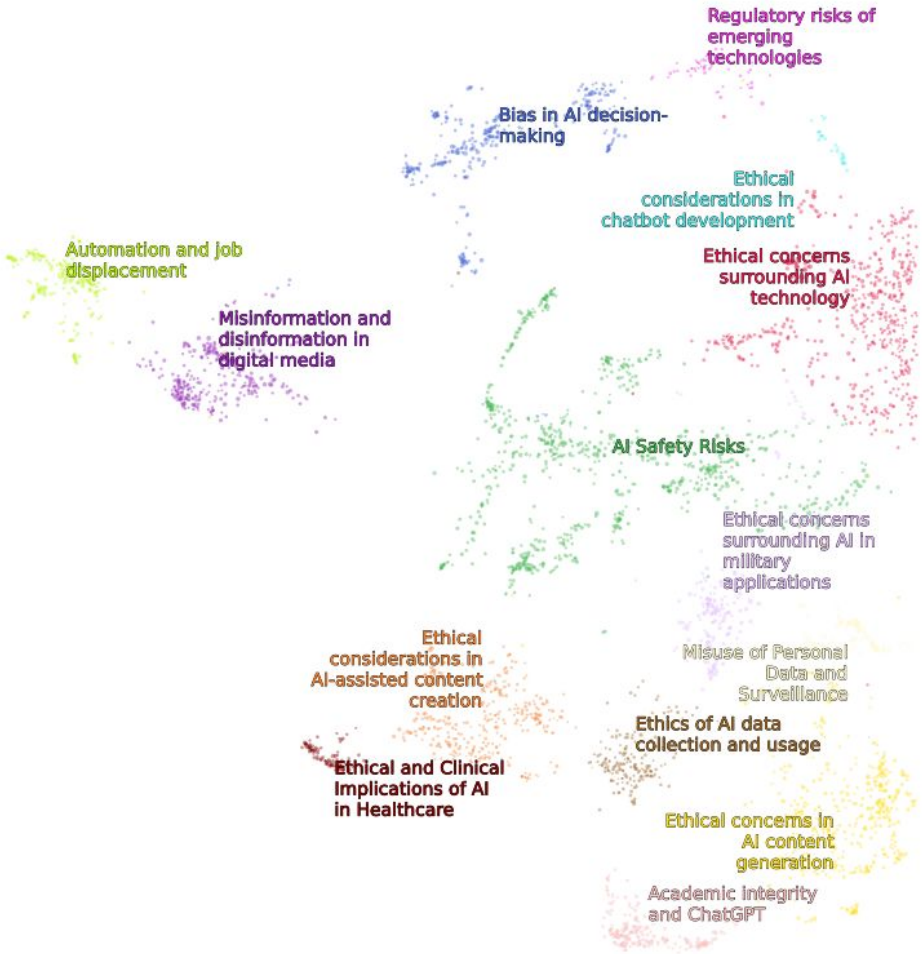
Extract reasons - Examples

Paragraph:

“Most people are unable to tell they are watching a “deepfake” video even when they are informed that the content has been digitally altered, research suggests. The term “deepfake” refers to a video where artificial intelligence and deep learning – an algorithmic method used to train computers – has been used to make a person appear to say something they have not.”

Extracted reason:

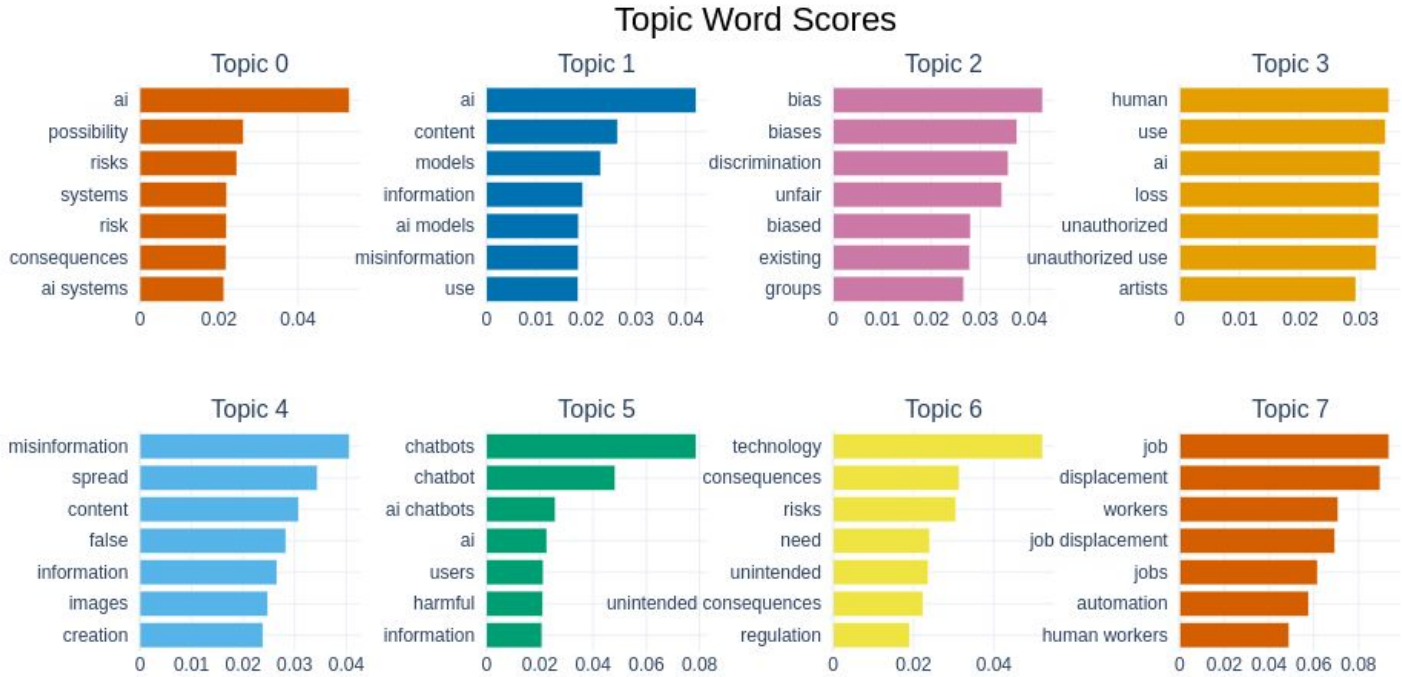
“The potential harm or risk of AI models discussed in the paragraph is the manipulation of media content through the use of deepfakes, which can deceive individuals into believing false information and potentially cause harm to individuals or society as a whole.”



Topic Model - BERTopic

'AI **Safety Risks**',
'Ethical concerns in AI **content generation**',
'Bias in AI decision-making',
'Ethical considerations in AI-assisted **content creation**',
'**Misinformation and disinformation** in digital media',
'Ethical considerations in **chatbot development**',
'**Regulatory risks** of emerging technologies',
'Automation and **job displacement**',
'**Academic integrity** and ChatGPT',
'**Privacy** Concerns with Facial Recognition Technology',
'Ethical concerns surrounding AI in **military applications**',
'Ethics of AI **data collection** and usage',
'Misuse of Personal Data and **Surveillance**',
'Ethical and Clinical Implications of **AI in Healthcare**',
'**Deepfakes** and Misinformation'

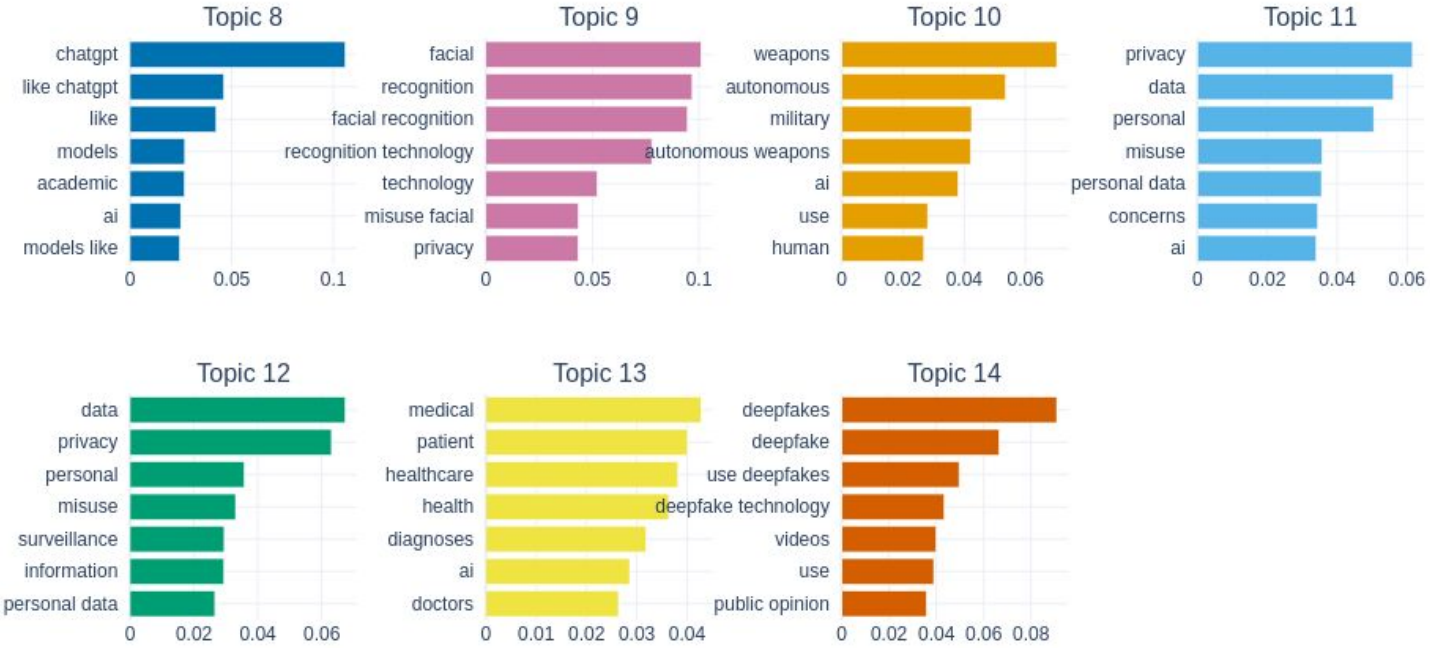
Topic Model - BERTopic



0	AI Safety Risks	'ai', 'risks', 'dangers', 'consequences', 'danger', 'risk', 'intelligence', 'safety', 'harm', 'concerns'	the dangers of unchecked AI development and the need for safeguards to prevent these risks. These risks may include the misuse of AI, the potential for AI to replace human jobs, and the risk of AI systems making biased or harmful decisions.
1	Ethical concerns in AI content generation	'misinformation', 'ai', 'disinformation', 'misuse', 'undermine', 'consequences', 'technology', 'misleading', 'accountability', 'biases'	the misuse of AI technology for political propaganda and manipulation, the creation of false or misleading images and information, and the potential for social collapse due to the widespread use of AI-generated content.
2	Bias in AI decision-making	'biases', 'bias', 'discrimination', 'discriminatory', 'biased', 'disparities', 'racism', 'unfair', 'stereotypes', 'ai'	gender and racial bias in hiring processes, and the issue of AI-driven prejudice. These biases can lead to unfair treatment of certain groups and individuals, perpetuating existing social inequalities.
3	Ethical considerations in AI-assisted content creation	'ai', 'copyright', 'writers', 'infringement', 'creativity', 'copyrighted', 'artists', 'art', 'creators', 'artistic'	the displacement of human writers due to the increasing use of AI in the industry, leading to job loss and a decline in the quality of content. Additionally, the use of AI may limit residual payments for writers and potentially lead to the exploitation of their work without proper compensation.

4	Misinformation and disinformation in digital media	'misinformation', 'disinformation', 'undermine', 'propaganda', 'conspiracy', 'media', 'consequences', 'misleading', 'inaccuracies', 'spreading'	the spread of misinformation and manipulated content through social media platforms, leading to the amplification of false claims and the potential harm to individuals and communities.
5	Ethical considerations in chatbot development	'chatbots', 'chatbot', 'bots', 'consequences', 'conversations', 'misuse', 'harmful', 'responses', 'concerns', 'harm'	the risk of inappropriate responses from the chatbot, potentially leading to harm or negative consequences for users.
6	Regulatory risks of emerging technologies	'risks', 'dangers', 'danger', 'safety', 'concerns', 'threats', 'consequences', 'technology', 'risk', 'harmful'	the need for guardrails to mitigate real risks posed by the technology, such as unintended consequences or negative impacts on society.
7	Automation and job displacement	'automation', 'unemployment', 'workforce', 'workers', 'employment', 'ai', 'robots', 'employees', 'machines', 'economic'	is the replacement of human workers by automation, leading to job loss and unemployment.

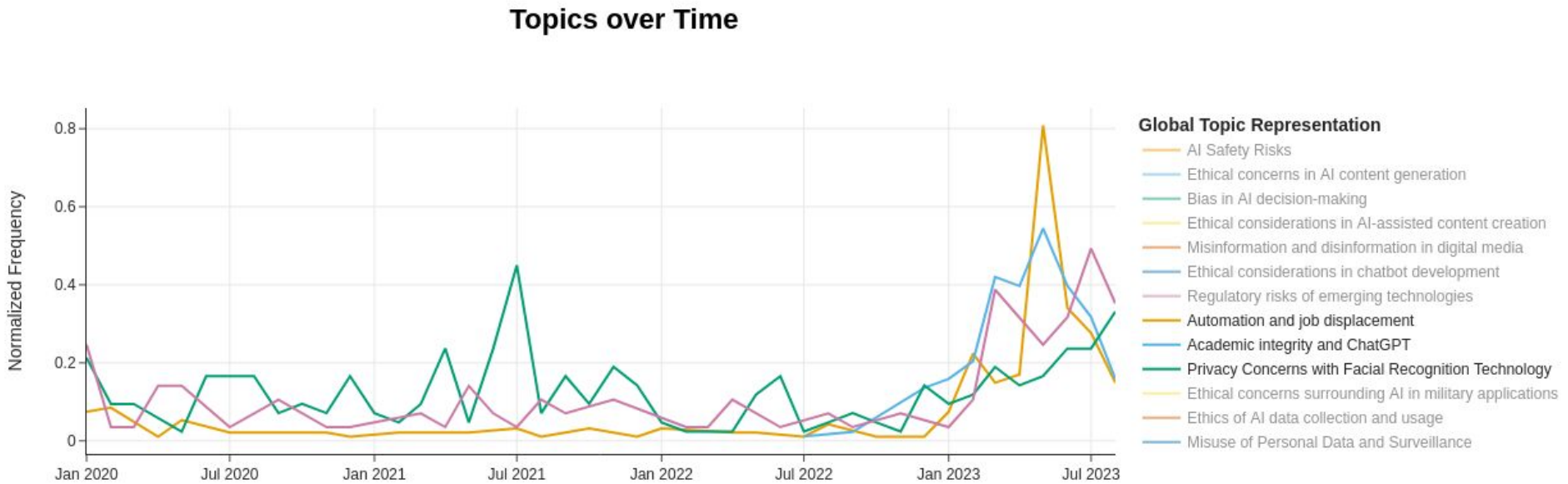
Topic Model - BERTopic



8	Academic integrity and ChatGPT	'chatgpt', 'dishonesty', 'students', 'cheating', 'chatbots', 'misuse', 'academic', 'student', 'essays', 'cheat'	is the use of ChatGPT by students to cheat on their assignments, which could lead to academic dishonesty and undermine the integrity of the education system.
9	Privacy Concerns with Facial Recognition Technology	'facial', 'biometric', 'faces', 'surveillance', 'privacy', 'face', 'misuse', 'recognition', 'concerns', 'biases'	the misuse of facial recognition technology, the collection and storage of sensitive personal data, and the potential for bias in algorithms used for facial recognition.
10	Ethical concerns surrounding AI in military applications	'ai', 'autonomous', 'weapons', 'warfare', 'technology', 'war', 'robots', 'artificial', 'military', 'intelligence'	is the misuse of technology for military purposes, specifically the development of autonomous weapons that can cause harm to human life and violate human rights.
11	Ethics of AI data collection and usage	'privacy', 'ai', 'risks', 'misuse', 'data', 'ethical', 'concerns', 'surveillance', 'security', 'infringing'	privacy concerns due to the use of personal data for training AI models, and the risk of data breaches or unauthorized access to sensitive information.

12	Misuse of Personal Data and Surveillance	'privacy', 'surveillance', 'misuse', 'threats', 'security', 'risks', 'theft', 'data', 'oversight', 'infringing'	the collection and use of personal data without consent, the linking of online activities to personal information, and the potential for increased data processing and surveillance.
13	Ethical and Clinical Implications of AI in Healthcare	'misdiagnosis', 'healthcare', 'medical', 'doctors', 'patients', 'diagnoses', 'ai', 'patient', 'concerns', 'health'	in healthcare include concerns about accuracy, bias, and the misuse of patient data. There may be a risk of AI systems replacing human clinicians, leading to job loss and decreased access to medical care for vulnerable populations. Additionally, there may be ethical concerns around the use of AI in diagnosis and treatment decisions.
14	Deepfakes and Misinformation	'deepfakes', 'deepfake', 'disinformation', 'misinformation', 'undermine', 'propaganda', 'malicious', 'deception', 'misuse', 'media'	the use of deepfakes to spread misinformation and manipulate public opinion, as well as the potential for AI-generated images to be used in political propaganda and hate speech.

Topic Model


Topics over Time
Normalized Frequency
0.8
0.6
0.4
0.2
0
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Global Topic Representation
AI Safety Risks
Ethical concerns in AI content generation
Bias in AI decision-making
Ethical considerations in AI-assisted content creation
Misinformation and disinformation in digital media
Ethical considerations in chatbot development
Regulatory risks of emerging technologies
Automation and job displacement
Academic integrity and ChatGPT
Privacy Concerns with Facial Recognition Technology
Ethical concerns surrounding AI in military applications
Ethics of AI data collection and usage
Misuse of Personal Data and Surveillance

	2 - Bias in AI decision-making	14 - Deepfakes and Misinformation	8 - Academic integrity and ChatGPT	7 - Automation and job displacement
The Daily Mirror	8.57	1.90	**5.71**	**8.57**
The Guardian	**8.92**	1.85	3.36	3.89
The Independent	7.11	0.79	3.16	2.37
The Sun	0.76	1.91	4.20	**7.25**
The Times	4.29	1.98	4.62	6.37

Conclusion

- Demonstrated an increase in AI discourse, particularly around harms and risks
- AI coverage across different media outlets
- Diverse range of AI-related topics: misinformation, job displacement, privacy, and ethics
- Necessity for context-aware evaluations when applying gllm across different data
- Next step: study with fine-tuned classification models for more nuanced analysis

Thank you

➔ weber.aca@gmail.com

Appendix

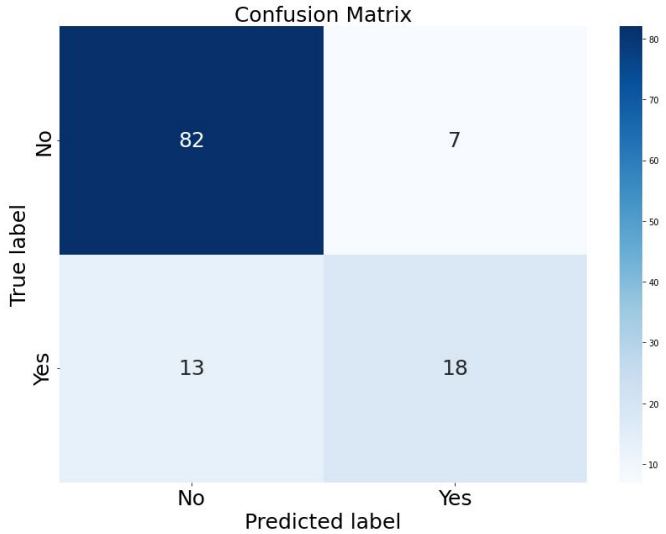
US

Newspaper	Total articles	Only harms	Only not harms	Percentage only harms	Percentage paragraphs harms
The New York Times	1514	138	1104	11.11	15.75
The Wall Street Journal	1590	70	1239	5.35	9.61
The Washington Post	962	89	631	12.36	**18.26**

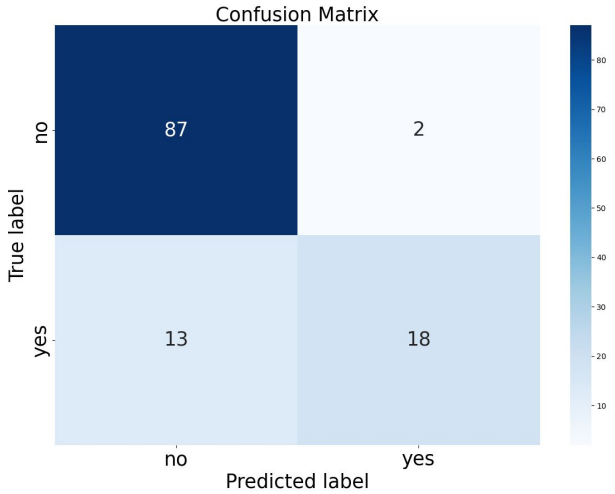
News Outlet	Country	articles	paragraphs	Date - old	Date - youngest
The Daily Mirror	UK	304	715	2020-01-02	2023-08-18
The Guardian	UK	2024	8021	2020-01-01	2023-08-22
The Independent	UK	350	1009	2020-01-03	2023-08-19
The Sun	UK	450	1140	2020-01-01	2023-08-21
The Times	UK	2025	5149	2020-01-01	2023-08-22
Chicago Tribune	US	181	771	2022-09-09	2023-08-19
The New York Times	US	1514	4379	2020-01-01	2023-08-22
The Wall Street Journal	US	1590	5805	2020-01-02	2023-08-22
The Washington Post	US	962	4353	2020-01-01	2023-08-22
Sum		13597	42958		

Best Prompt - Result

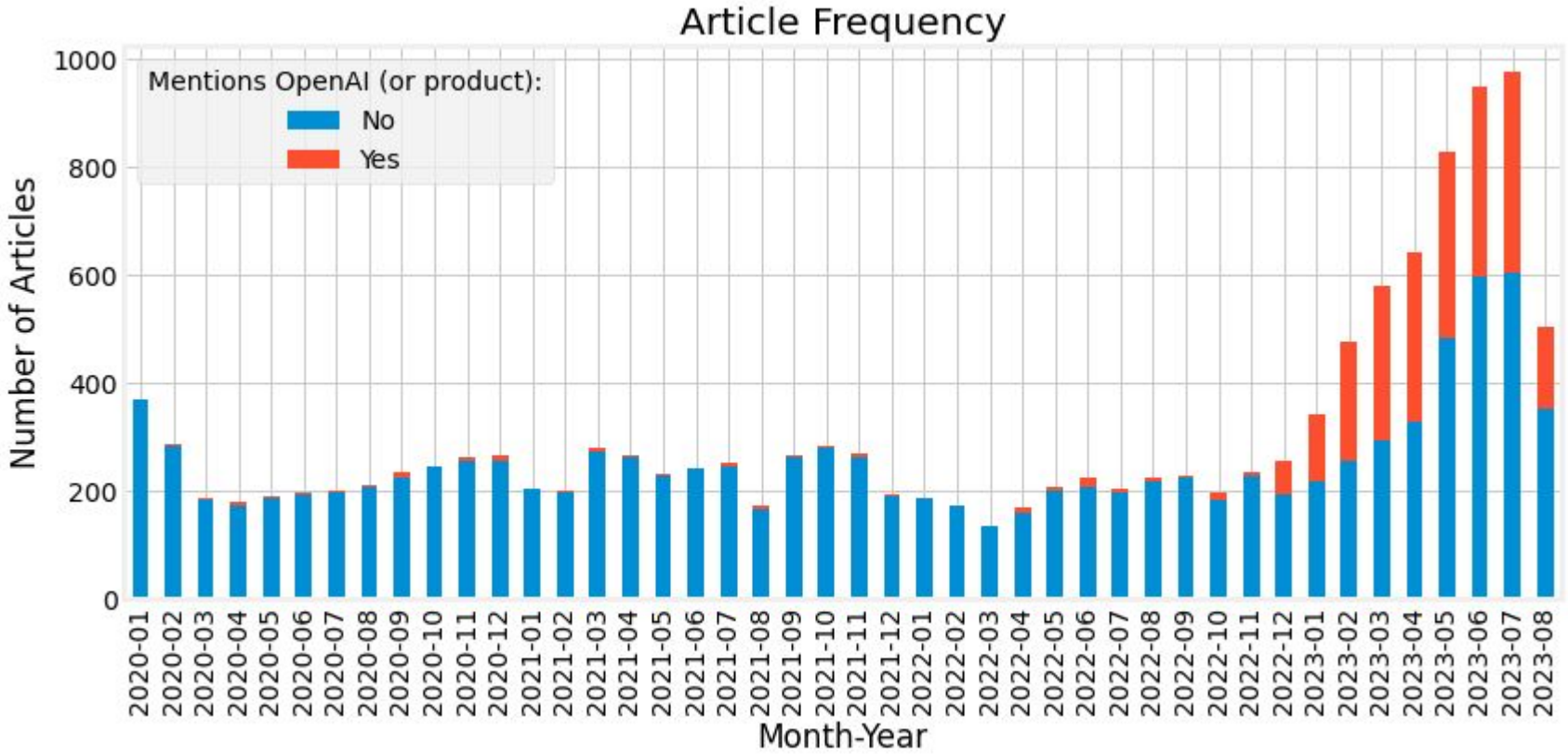
US: Llama 2 - 13b



Llama3.1 - 8b



German


Article Frequency
Mentions OpenAI (or product):
No
Yes
Number of Articles
0
200
400
600
800
1000
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11
2022-12
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
Month-Year

Germany

Newspaper	Total articles	Only harms	Only not harms	Percentage only harms	Percentage paragraphs harms
BILD	119	12	94	11.32	**13.71**
DIE ZEIT	359	11	281	3.77	8.62
Die Welt	958	28	768	3.52	8.98
Frankfurter Rundschau	588	33	463	6.65	12.08
Süddeutsche Zeitung	1665	51	1380	3.56	8.11
taz - die tageszeitung	508	26	393	6.21	11.32

Top keywords - Articles

keyword	frequency	number of articles
artificial* intelligence*	10639	7946
künstliche* intelligenz	6336	3800
chatgpt	5996	1979
chatbot*	3988	1664
openai	2984	1095

Outlet	Country	percentage harms and risks discussed
BILD	Germany	22.5
DIE ZEIT	Germany	40.0
Die Welt	Germany	15.0
Frankfurter Rundschau	Germany	37.5
Süddeutsche Zeitung	Germany	40.0
taz - die tageszeitung	Germany	52.5
The Daily Mirror	UK	27.5
The Guardian	UK	40.0
The Independent	UK	37.5
The Sun	UK	25.0
The Times	UK	22.5
The New York Times	US	32.5
The Wall Street Journal	US	12.5
The Washington Post	US	32.5
Average		32.0

News Outlet	Country	articles	paragraphs	Date - old	Date - youngest
BILD	Germany	119	265	2020-01-03	2023-07-31
DIE ZEIT	Germany	359	1163	2020-01-03	2023-08-17
Die Welt	Germany	958	3375	2020-01-02	2023-08-22
Frankfurter Rundschau	Germany	588	1345	2020-01-04	2023-08-22
Süddeutsche Zeitung	Germany	1665	4122	2020-01-03	2023-08-22
taz - die tageszeitung	Germany	508	1346	2020-01-02	2023-08-21
The Daily Mirror	UK	304	715	2020-01-02	2023-08-18
The Guardian	UK	2024	8021	2020-01-01	2023-08-22
The Independent	UK	350	1009	2020-01-03	2023-08-19
The Sun	UK	450	1140	2020-01-01	2023-08-21
The Times	UK	2025	5149	2020-01-01	2023-08-22
Chicago Tribune	US	181	771	2022-09-09	2023-08-19
The New York Times	US	1514	4379	2020-01-01	2023-08-22
The Wall Street Journal	US	1590	5805	2020-01-02	2023-08-22
The Washington Post	US	962	4353	2020-01-01	2023-08-22
Sum		13597	42958		

	precision	recall	f1-score	support
No	0.81	0.93	0.87	408
Yes	0.79	0.55	0.65	192
accuracy			0.81	600
macro avg	0.80	0.74	0.76	600
weighted avg	0.81	0.81	0.80	600