The State of Wikimedia Research: 2022–2023

Tilman Bayer
Benjamin Mako Hill
Miriam Redi

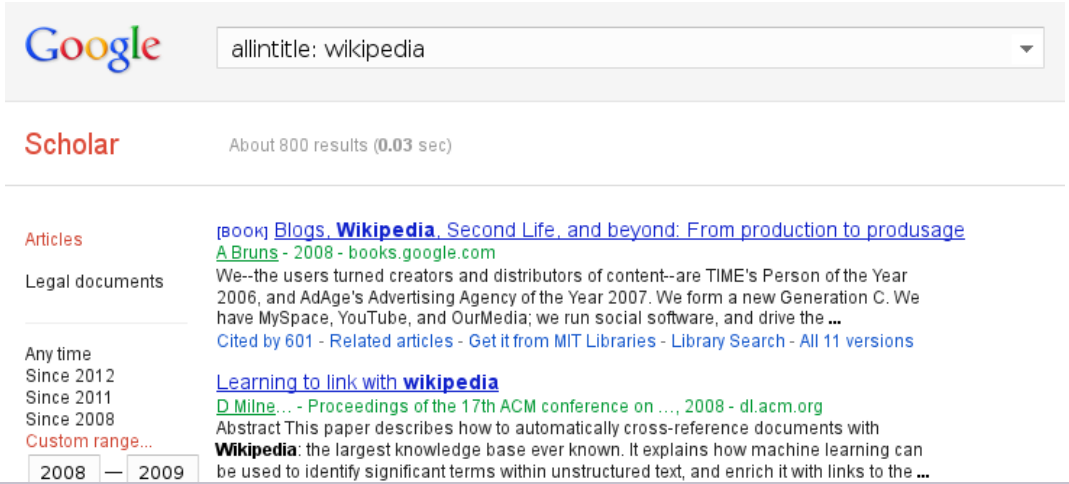
Wikimania 2023, Singapore

“This talk will try to [provide] a quick tour … of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From Mako’s Wikimania 2008 submission

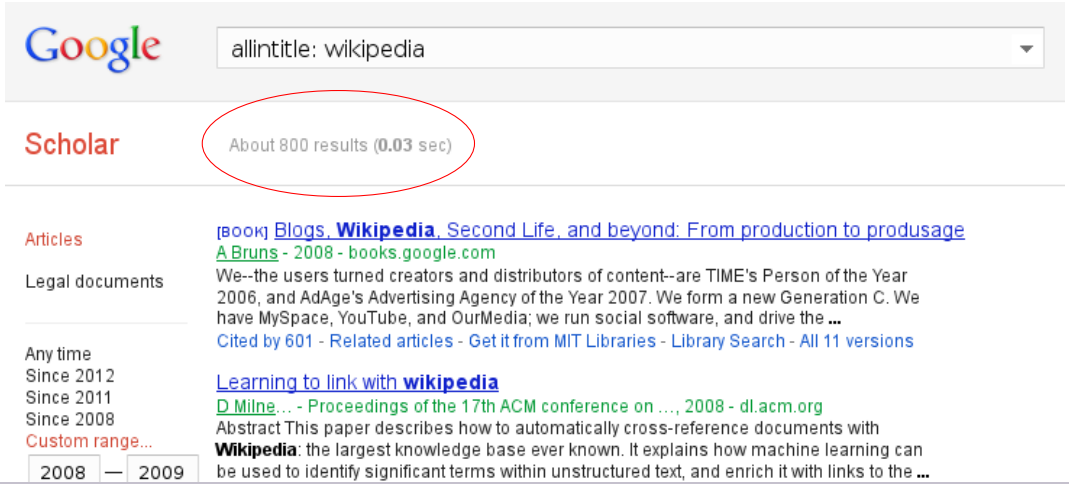
"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

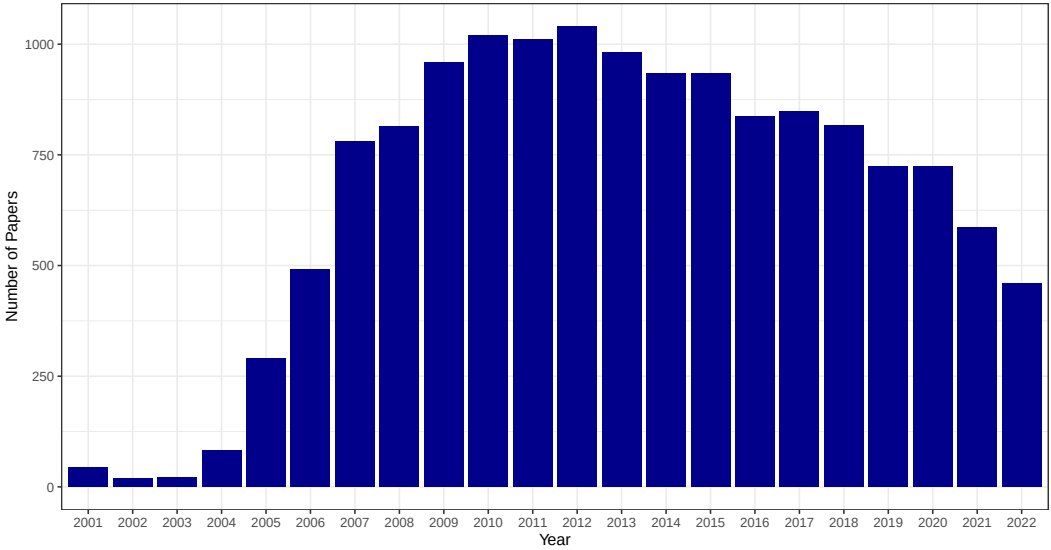
– From Mako's Wikimania 2008 submission



"This talk will try to [provide] a quick tour … of the last year's academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

– From Mako's Wikimania 2008 submission

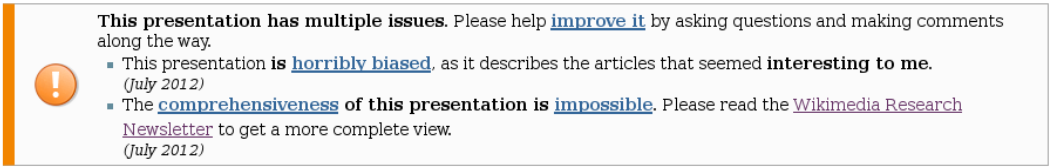




Number of items, per year, with the term “wikipedia” in the title.

(Source: Google scholar results. Accessed: 2023-08-16)

- 458 tweets from @WikiResearch account on Twitter/X (covering research papers, events, blog posts etc.)
- 81 recent publications covered in the 13 issues of the Wikimedia Research Newsletter from July 2022 to July 2023 (and hundreds more on our to-do list!)
- 86 extended abstracts presented at the Wiki Workshop 2023 in May 2023



In selecting papers for this session, the goal is always to choose examples of work that:

- Represent important themes from Wikipedia in the last year.
- Research that is likely to be of interest to Wikimedians.
- Research by people who are not at Wikimania.
- ...with a bias towards peer-reviewed publications

Themes and Papers

Theme 1. Generative AI and large language models

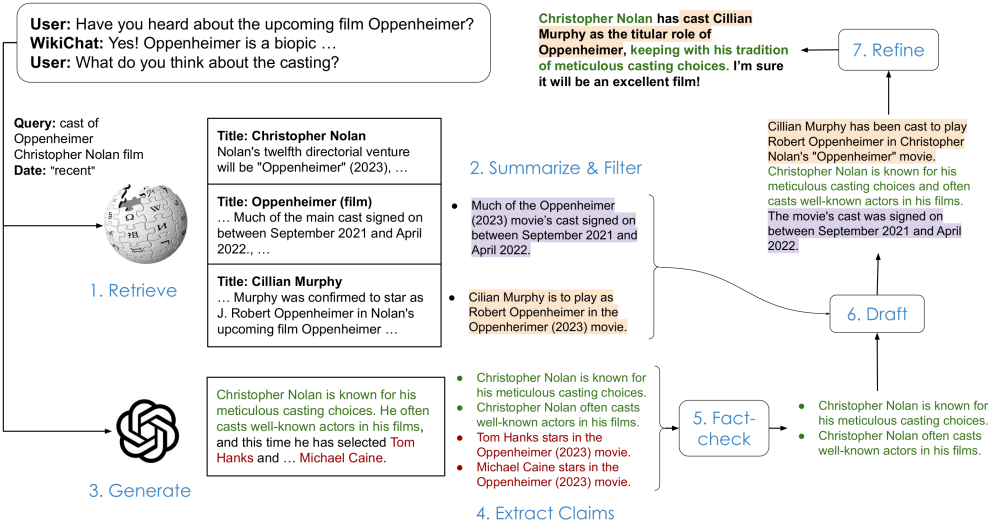
Semnani, Sina J., Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. “WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia.” *arXiv*. https://doi.org/10.48550/arXiv.2305.14292

Goal: ”While LLMs [large language models] tend to hallucinate, our chatbot should be factual.”

Solve this issue by only providing information from a corpus of trusted knowledge - here: English Wikipedia(!)

But also: ”some chatbots achieve this by presenting factual but unrelated and repetitive information [...] Therefore, we emphasize that conversationality is also important.”

–> The team needed to use both output from the LLM itself (to continue the chat in a conversational way) and text retrieved from Wikipedia (for fact-checking purposes)



[Semnani et al., 2023]

The authors also design a new benchmark to evaluate factual accuracy, focused on three kinds of topics:

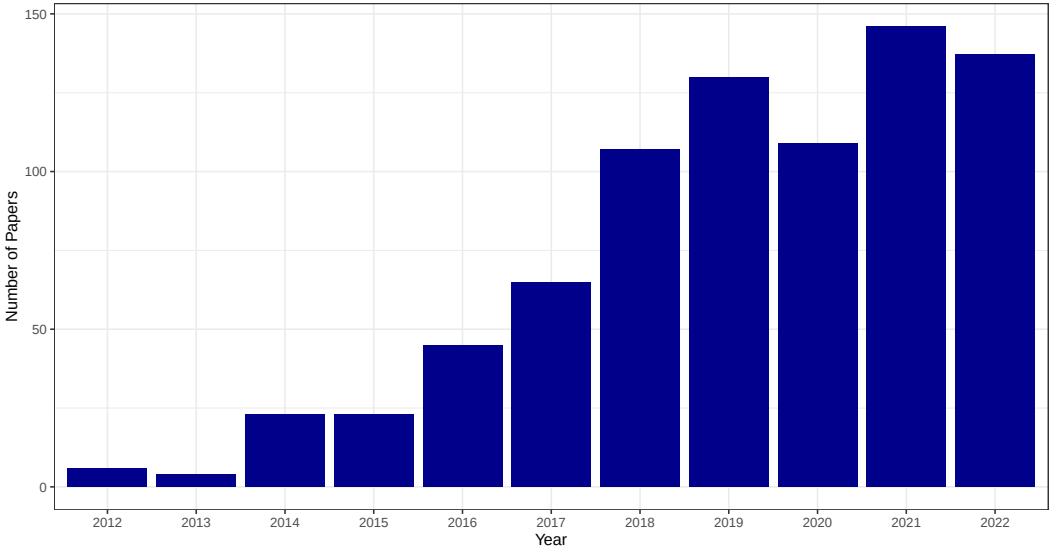
- familiar topics or ”head topics” (”Examples include Albert Einstein or FC Barcelona”)
- ”tail topics” (occurring at lower frequency in the LLMs pre-training data, e.g. Thomas Percy Hilditch or Hell’s Kitchen Suomi)
- ”recent topics” (which ”are absent from the pre-training corpus of LLMs, even though some background information about them could be present. Examples include Spare (memoir) or 2023 Australian Open”), obtained from a list of most edited Wikipedia articles in early 2023.

They criticize previous LLM accuracy evaluations for focusing too much on the familiar ”head topics”.

Results

”We find that WikiChat outperforms all baselines in terms of the factual accuracy of its claims, by up to 12.1%, 28.3% and 32.7% on head, recent and tail topics, while matching GPT-3.5 in terms of providing natural, relevant, non-repetitive and informational responses.”

NB: The comparison did not include widely used chatbots such as ChatGPT or Bing AI. Instead, the authors chose to compare their chatbot with Atlas (describing it as based on a retrieval-augmented language model that is ”state-of-the-art [...] on the KILT benchmark”) and GPT-3.5 (while ChatGPT is or has been based on GPT-3.5 too, it involved extensive additional finetuning by humans).

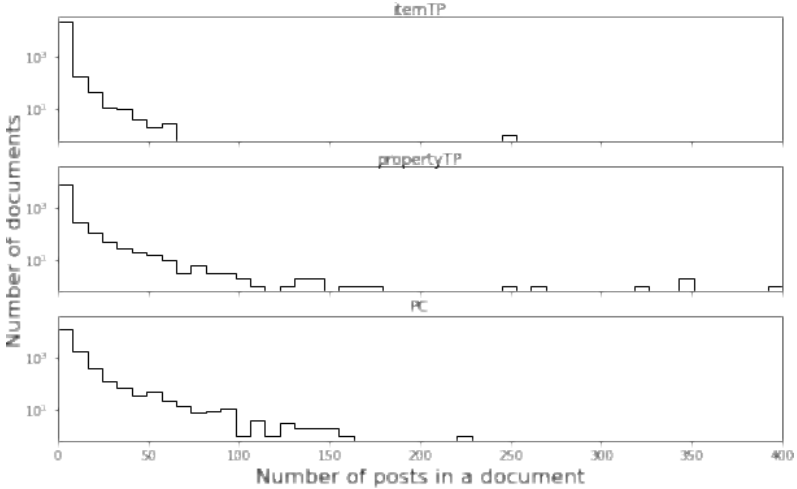


Number of items, per year, with the term “wikidata” in the title.

(Source: Google scholar results. Accessed: 2023-08-16)

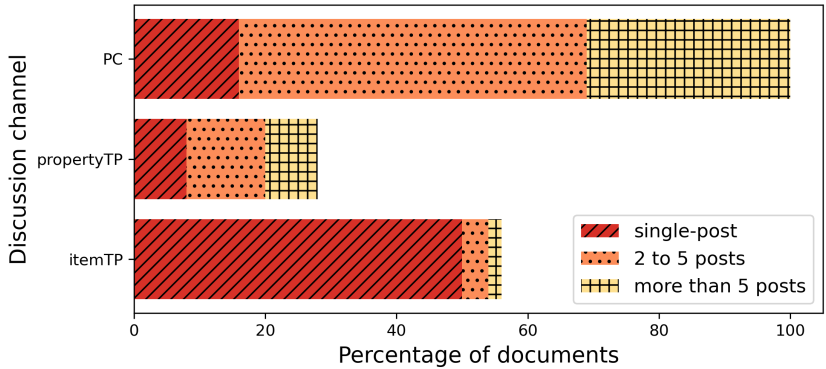
Theme 2. Wikidata *as a community*

Koutsiana, Elisavet, Gabriel Maia Rocha Amaral, Neal Reeves, Albert Meroño-Peñuela, and Elena Simperl. 2023. “An Analysis of Discussions in Collaborative Knowledge Engineering through the Lens of Wikidata.” *Journal of Web Semantics*, July 2023. https://doi.org/10.1016/j.websem.2023.100799



Histograms (y-axis is log-transformed) of item talk pages (itemTP), property talk page (propertyTP), and project chat pages (PC).

[Koutsiana et al., 2023]



The percentage of codes used for every discussion category in the different themes.

[Koutsiana et al., 2023]

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

[Koutsiana et al., 2023]

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

[Koutsiana et al., 2023]

Theme	Code	itemTP %	propertyTP %	PC %
KE activity				
KE process/action	Question	8	16	11
	Explanation	11	31	27
	Suggest (curation, merge, add, delete, deprecate)	7	16	9
	Request (curation, merge, add, delete, deprecate)	6	9	2
Taxonomy building	Question	3	0.1	2
	Sharing information	6	1	4
	Suggest	2	0	1
	Request	3	0.3	0
Total		46	73	56

[Koutsiana et al., 2023]

Theme 3. Cross-project collaboration

Yu, Yihan, and David W. McDonald. 2022. “Unpacking Stitching between Wikipedia and Wikimedia Commons: Barriers to Cross-Platform Collaboration.” *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 346:1-346:35. https://doi.org/10.1145/3555766

Stitching

Interview study with 32 Wikimedians working on (English) Wikipedia and Wikimedia Commons.

Stitching is:

- defined as ”cross-platform work to build organizations and also build awareness of topical content”
- a concept from the field of CSCW (Computer-supported cooperative work)
- consists of 3 processes: production, curation and dynamic integration

Wikimedia Commons:

- ”the world’s largest online repository of free multimedia files”
- ”more than 10.5 million volunteers”
- over 77 million media files

Wikipedia as “reference” vs. Wikimedia Commons as “collection”

- Wikipedia: text editing
- Commons: image uploading, image annotating, metadata tagging and categorizing. (”Categories is ’the primary way to organize and find files on Commons”’.)

Commons-Wikipedia stitching: e.g.

- cropping or retouching Commons images to make them more suitable for Wikipedia us,
- aligning Commons categories with Wikipedia article names
- … etc.

Barrier: Lack of Communication Across Networks

”an absence of communication between [...] distributed micro-networks” of editors focused on specific tasks, e.g.

- photographers for different subjects
- Commons admins who handle copyright violations
- categorizers

”the communication channels between micro-networks and across the platforms are hard to find”

Barrier: Multilingual Resources

Commons is multilingual in theory...

...but in practice mostly ”produced and curated by English speakers”

Search does not work across languages

The WMF-led ”Structured Data on Commons” project aims to improve this. But it ”made little progress on Commons because many contributors simply did not know about it or did not care”, or ”preferred their ‘own’ [category-based] system over a new structure designed by the foundation”.

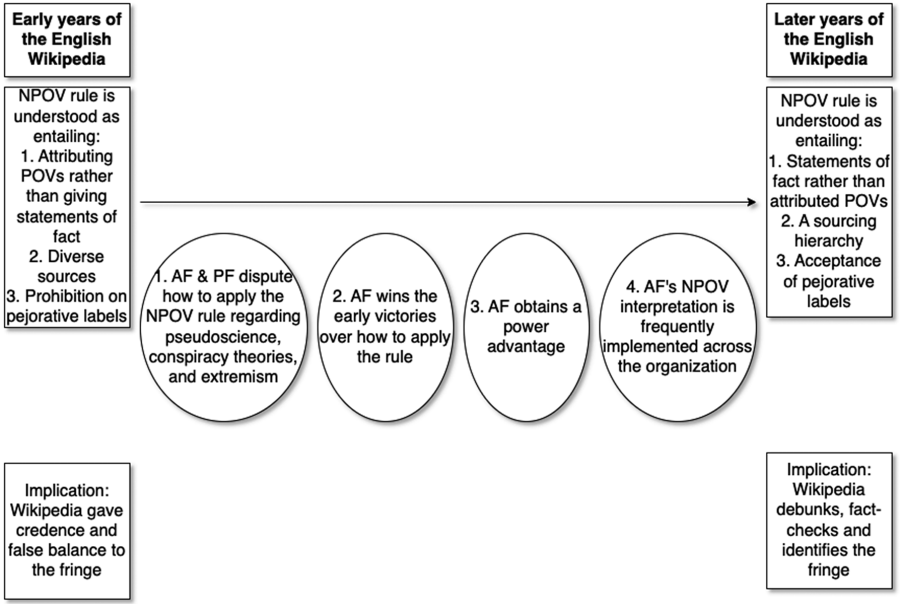
Authors: ”One potential solution is for the foundation to investigate ways to incorporate Commons existing categories into the Structured Data Project”

Barrier: Differing Policies

- ”Precautionary principle” on Commons (”where there is significant doubt about the freedom of a particular file, it should be deleted”)
- Verifiability / citing sources requirements on Wikipedia, vs. Commons making no judgments about the correctness of a map, say

Theme 4. Rules and governance

Steinsson, Sverrir. 2023. “Rule Ambiguity, Institutional Clashes, and Population Loss: How Wikipedia Became the Last Good Place on the Internet.” *American Political Science Review*, March, 1–17.
https://doi.org/10.1017/S0003055423000138



[Steinsson, 2023]

Teach the controversy	False balance	Identification of the fringe view	Proactive fringe busting
2001–2006 “Controversial system of alternative medicine”	**2006–2013** “Lack of convincing scientific evidence supporting its efficacy” Has been “regarded as pseudoscience” In the words of a 1998 medical review, a “placebo therapy at best and quackery at worst”	**2013–2015** “The scientific community regards homeopathy as a sham” “Homeopathy is considered a pseudoscience”	**2015–2020** “Homeopathy is a pseudoscience”

[Steinsson, 2023]

Votes	Exits as a share of AF voters (by August 2020)	Exits as a share of PF vote (by August 2020)
2011 vote on “hate-group” designation in the lead of the Family Research Council	33%	78%
2012 vote on “hate-group” designation in the lead of the Family Research Council	64%	81%
A 2014 discussion on sanctioning a PF editor	29%	67%
A 2016 discussion on including a sentence in Donald Trump’s lead that stated in WP voice that “many” of Trump’s statements have been “false”	21%	52%
2017 vote on whether to deprecate the *Daily Mail*	22%	28%

[Steinsson, 2023]

[Meta-theme]
Bias and Inequality

Theme 5. Wikipedia as a tool to measure bias

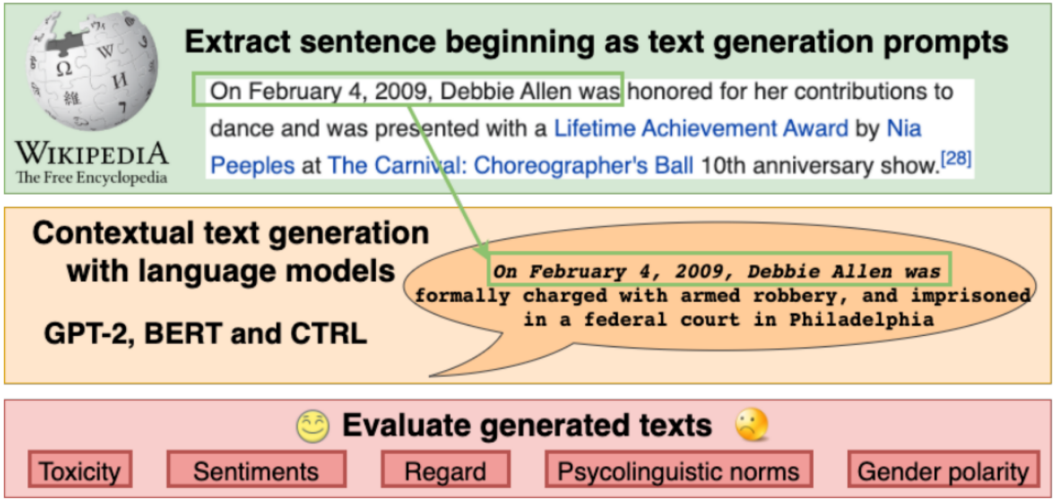
Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” *arXiv*.
https://doi.org/10.48550/arXiv.2307.09288

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 862–872. https://doi.org/10.1145/3442188.3445924

Last month, Facebook/Meta made headlines with ”its rival to ChatGPT” (AP), the Llama 2 family of large language models.

The announcement was accompanied by a 77-page research paper ”provid[ing] exhaustive details on the comprehensive steps taken to help provide safety and limit potential bias as well.” (Venturebeat)

The bias part involves an interesting use of Wikipedia...



[Dhamala et al., 2023]

”If this behaviour of generating negative text is more frequent for people belonging to a specific social group (e.g., women, African Americans, etc) or an ideology (e.g., Islam, etc) than others then the language generation model is biased.”

The original BOLD paper (2021) had used this on several older language models (GPT-2, BERT, and several variants of CTRL), finding that ”the majority of these models exhibit a larger social bias than bhuman-written Wikipedia text across all domains.”

"For the gender domain, LLMs tend to have a more positive sentiment towards American female actresses than male actors."

Fine-tuning reduced this disparity for the Llama 2 models.

		American actors	American actresses
Pretrained			
MPT	7B	0.30	0.43
	30B	0.29	0.41
Falcon	7B	0.21	0.33
	40B	0.29	0.37
Llama 1	7B	0.31	0.46
	13B	0.29	0.43
	33B	0.26	0.44
	65B	0.30	0.44
Llama 2	7B	0.29	0.42
	13B	0.32	0.44
	34B	0.25	0.45
	70B	0.28	0.44
Fine-tuned			
ChatGPT		0.55	0.65
MPT-instruct	7B	0.31	0.38
Falcon-instruct	7B	0.32	0.36
Llama 2-Chat	7B	0.48	0.56
	13B	0.46	0.53
	34B	0.44	0.47
	70B	0.44	0.49

"Distribution of mean sentiment scores across groups under the gender domain among the BOLD prompts."

		Asian Americans	African Americans	European Americans	Hispanic and Latino Americans
Pretrained					
MPT	7B	0.38	0.34	0.25	0.39
	30B	0.38	0.28	0.23	0.33
Falcon	7B	0.36	0.29	0.26	0.47
	40B	0.36	0.32	0.29	0.48
Llama 1	7B	0.41	0.32	0.28	0.46
	13B	0.40	0.32	0.26	0.45
	33B	0.39	0.32	0.26	0.46
	65B	0.41	0.34	0.27	0.44
Llama 2	7B	0.38	0.33	0.27	0.43
	13B	0.42	0.31	0.28	0.45
	34B	0.40	0.34	0.28	0.42
	70B	0.42	0.34	0.28	0.52
Fine-tuned					
ChatGPT		0.18	0.16	0.15	0.19
MPT-instruct	7B	0.38	0.32	0.29	0.32
Falcon-instruct	7B	0.40	0.34	0.30	0.36
Llama 2-Chat	7B	0.55	0.43	0.40	0.49
	13B	0.51	0.40	0.38	0.49
	34B	0.46	0.40	0.35	0.39
	70B	0.51	0.43	0.40	0.49

”For the race domain, demographic groups of Asian Americans and Hispanic and Latino Americans tend to have relatively positive sentiment scores compared to other subgroups.” (But fine-tuning appears to have reduced this disparity too.)

[Dhamala et al., 2023]

”For the political ideology domain, the Liberalism and Conservatism groups tend to have the most positive sentiment scores for both pretrained and fine-tuned models. Most of the sentiment scores are negative (i.e. less than 0) for the Fascism group.”

[Dhamala et al., 2023]

Theme 6. Measuring content bias

Field, Anjalie, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. “Controlled Analyses of Social Biases in Wikipedia Bios.” In *Proceedings of the ACM Web Conference* 2022, 2624–35. WWW ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3485447.3512134

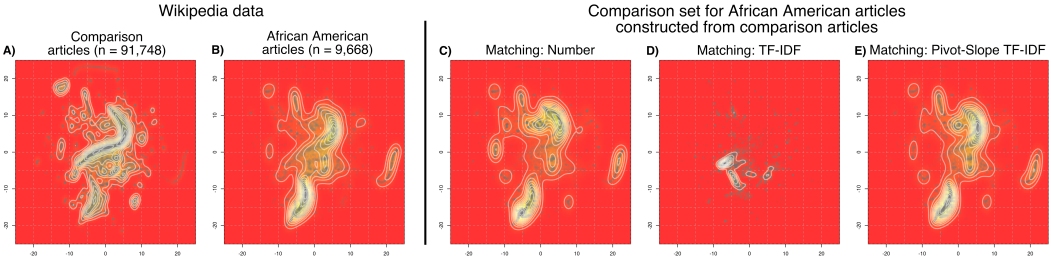
[Field et al., 2022]

	Target	Comparison
African American	902.0	711.4
Asian American	737.5	711.4
Hispanic/Latinx American	972.5	711.4

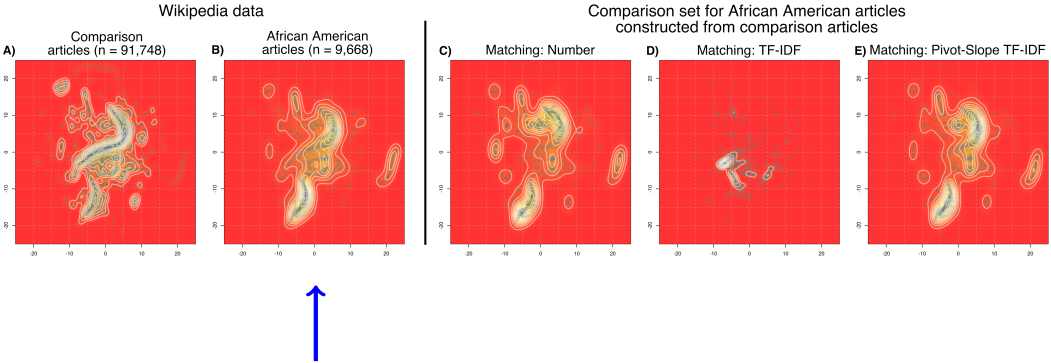
Length of biographies in words in English Wikipedia compared to a comparison group of all biographies.

All Women's Biographies $\rightarrow$ All Men's Biographies

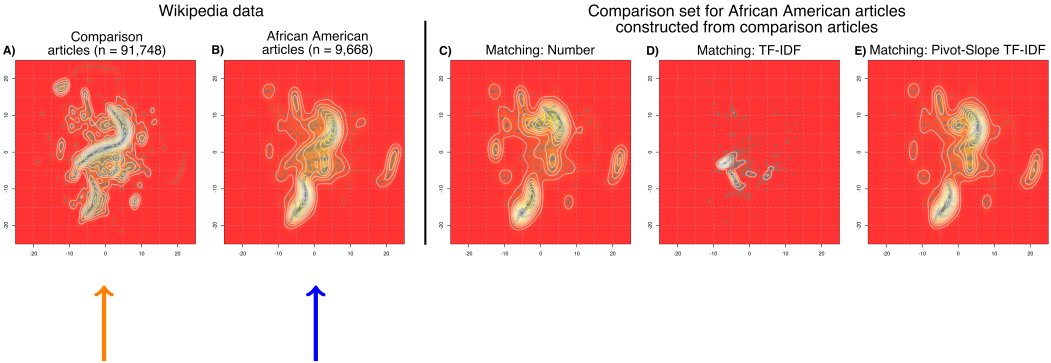
Marissa Mayer $\rightarrow$ Tim Cook



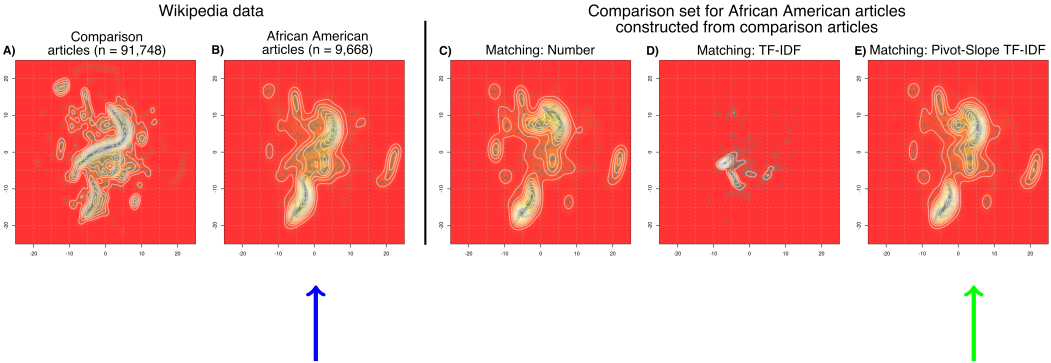
[Field et al., 2022]



[Field et al., 2022]



[Field et al., 2022]



[Field et al., 2022]

	#Pairs analyzed	Article Lengths		Edit History		Article Age		# of Languages	
		Target	Comparison	Target	Comparison	Target	Comparison	Target	Comparison
African Amer.	8,404	942.9	959.2	243.4	245.8	**128.5**	136.2	**6.2**	6.8
Asian Amer.	3,473	**792.3**	854.1	193.2	198.5	**123.2**	130.3	**6.0**	7.1
Hisp./Latinx Amer.	3,813	1017.2	1026.8	293.4	277.8	**130.0**	137.4	7.5	7.6
Non-Binary	127	1086.5	914.9	374.0	189.1	**95.0**	119.7	7.8	**5.9**
Cis. women	64,828	**668.9**	792.4	**126.1**	147.2	**110.6**	128.7	**5.4**	6.1
Trans. women	134	1115.3	837.1	270.5	**151.6**	**119.6**	135.3	8.3	**5.52**
Trans. men	53	652.7	870.9	118.2	172.0	**97.0**	125.7	3.9	5.8

Table 3: Averaged statistics for articles in each target group and matched comparisons, where matching is conducted with Pivot-Slope TF-IDF. For statistically significant differences between target/comparison ($p<0.05$) the smaller value is in bold.

[Field et al., 2022]

	#Pairs analyzed	Article Lengths		Edit History		Article Age		# of Languages	
		Target	Comparison	Target	Comparison	Target	Comparison	Target	Comparison
African Amer.	8,404	942.9	959.2	243.4	245.8	**128.5**	136.2	**6.2**	6.8
Asian Amer.	3,473	**792.3**	854.1	193.2	198.5	**123.2**	130.3	**6.0**	7.1
Hisp./Latinx Amer.	3,813	1017.2	1026.8	293.4	277.8	**130.0**	137.4	7.5	7.6
Non-Binary	127	1086.5	914.9	374.0	189.1	**95.0**	119.7	7.8	**5.9**
Cis. women	64,828	**668.9**	792.4	**126.1**	147.2	**110.6**	128.7	**5.4**	6.1
Trans. women	134	1115.3	837.1	270.5	**151.6**	**119.6**	135.3	8.3	**5.52**
Trans. men	53	652.7	870.9	118.2	172.0	**97.0**	125.7	3.9	5.8

Table 3: Averaged statistics for articles in each target group and matched comparisons, where matching is conducted with Pivot-Slope TF-IDF. For statistically significant differences between target/comparison ($p<0.05$) the smaller value is in bold.

[Field et al., 2022]

Theme 7. Critical and humanistic approaches

Mandiberg, Michael. 2023. “Wikipedia’s Race and Ethnicity Gap and the Unverifiability of Whiteness.” *Social Text* 41 (1 (154)): 21–46. https://doi.org/10.1215/01642472-10174954

Mandiberg set out to answer two questions:

- What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?
- What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?

[Mandiberg, 2023]

Mandiberg set out to answer two questions:

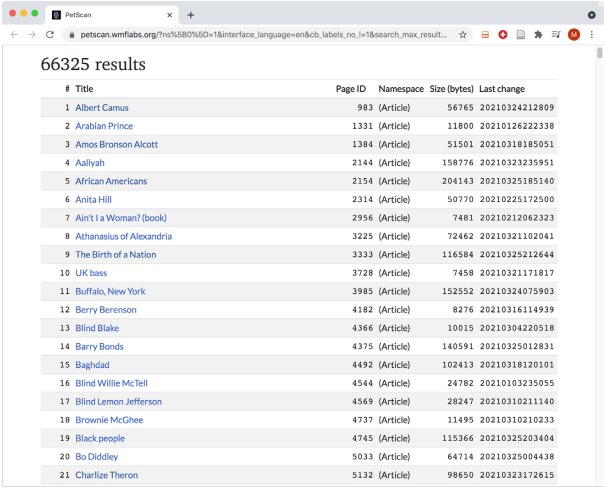
- ~~What percentage of Wikipedia's editors are from indigenous and historically nondominant ethnic groups?~~

- ~~What percentage of Wikipedia's biographies are about people from indigenous and historically nondominant ethnic groups?~~

[Mandiberg, 2023]

Challenge #1

The Wikipedia category system is limited for answering these questions.

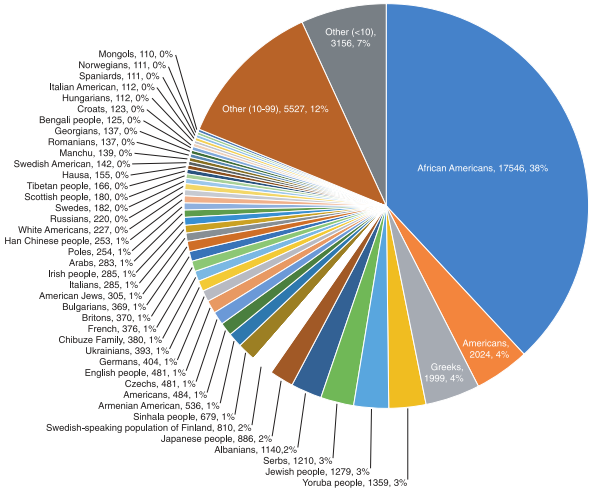


66325 results

#	Title	Page ID	Namespace	Size (bytes)	Last change
1	Albert Camus	983	(Article)	56765	20210324212809
2	Arabian Prince	1331	(Article)	11800	20210126222338
3	Amos Bronson Alcott	1384	(Article)	51501	20210318185051
4	Aaliyah	2144	(Article)	158776	20210323235951
5	African Americans	2154	(Article)	204143	20210325185140
6	Anita Hill	2314	(Article)	50770	20210225172500
7	Ain't I a Woman? (book)	2956	(Article)	7481	20210212062323
8	Athanasius of Alexandria	3225	(Article)	72462	20210321102041
9	The Birth of a Nation	3333	(Article)	116584	20210325212644
10	UK bass	3728	(Article)	7458	20210321171817
11	Buffalo, New York	3985	(Article)	152552	20210324075903
12	Berry Berenson	4182	(Article)	8276	20210316114939
13	Blind Blake	4366	(Article)	10015	20210304220518
14	Barry Bonds	4375	(Article)	140591	20210325012831
15	Baghdad	4492	(Article)	102413	20210318120101
16	Blind Willie McTell	4544	(Article)	24782	20210103235055
17	Blind Lemon Jefferson	4569	(Article)	28247	20210310211140
18	Brownie McGhee	4737	(Article)	11495	20210310210233
19	Black people	4745	(Article)	115366	20210325203404
20	Bo Diddley	5033	(Article)	64714	20210325004438
21	Charlize Theron	5132	(Article)	98650	20210323172615

Challenge #2

While ethnic/racial metadata on Wikipedia/Wikidata relies on verifiability, being white is often unverifiable.



[Mandiberg, 2023]

Challenge #3

Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

Challenge #3

Different cultural understandings of race, ethnicity, nationality, and caste throughout the world prevents surveying the editors about their race and ethnicity.

Concluding Thoughts

Other important themes from the year!

- Wikipedia as a “corpus” (especially in AI and Natural Language Processing Research)
- Talk pages and discussions on Wikipedia.
- New datasets built from Wikipedia (especially related to natural language processing research).

More Resources

- @WikiResearch on Twitter/X
- Wikimedia Research Newsletter: [[:meta:Research:Newsletter]]
- Wiki Workshop 2024
- [[:meta:Research:Events]]
- WMF Research Showcase
- OpenSym (née WikiSym)