WIKIMEDIA
FOUNDATION

Wikipedia ChatGPT plugin experiment summary

Mike Pham, Irene Florez

Feb 2024

Future Audiences Overview

- **Future Audiences**: quick experiments to learn about strategies the Wikimedia Foundation can pursue to continue to attract and retain knowledge seekers and sharers as technology and user behavior online changes.
    - Not trying to build full products!

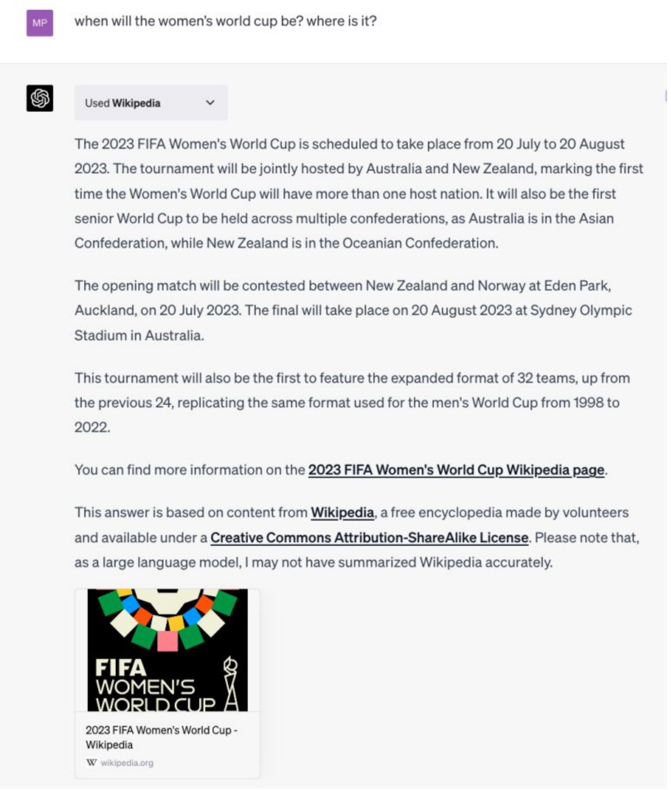
- Large Language Models (LLMs) and OpenAI's ChatGPT have changed how people can query advanced Machine Learning (ML) models to get text/code/images back

- Creating and training in-house LLMs is expensive and requires a lot of expertise. Experimentation on a third-party chatbot was an opportunity for us to learn quickly/cheaply.
    - Future Audiences partnered with OpenAI to test hypotheses around the future of chat assistants for information seeking, using their already built technology:
        - July 2023: launched Wikipedia plugin for ChatGPT 4.0[1]

See footnotes

What the Wikipedia plugin did:

- ChatGPT called the Wikipedia plugin based on user input and plugin description[2]
- Plugin uses the existing search API to find relevant Wikipedia articles
- Returns top 12 passages of top 4 relevant articles
- Summarize response using only this info
- Include boilerplate disclaimer (includes licensing)
- Include links to cited articles



when will the women's world cup be? where is it?

The 2023 FIFA Women's World Cup is scheduled to take place from 20 July to 20 August 2023. The tournament will be jointly hosted by Australia and New Zealand, marking the first time the Women's World Cup will have more than one host nation. It will also be the first senior World Cup to be held across multiple confederations, as Australia is in the Asian Confederation, while New Zealand is in the Oceanian Confederation.

The opening match will be contested between New Zealand and Norway at Eden Park, Auckland, on 20 July 2023. The final will take place on 20 August 2023 at Sydney Olympic Stadium in Australia.

This tournament will also be the first to feature the expanded format of 32 teams, up from the previous 24, replicating the same format used for the men's World Cup from 1998 to 2022.

You can find more information on the 2023 FIFA Women's World Cup Wikipedia page.

This answer is based on content from Wikipedia, a free encyclopedia made by volunteers and available under a Creative Commons Attribution-ShareAlike License. Please note that, as a large language model, I may not have summarized Wikipedia accurately.

2023 FIFA Women's World Cup - Wikipedia
W wikipedia.org

See footnotes

Executive Summary

- **Are LLMs and/or chat assistants a new paradigm for information seekers?**
  - No/not yet. We do not see evidence at this time that people are using this in a way that is drastically different from search.
- **Are chat assistants displacing traffic to Wikipedia?**
  - No
- **Are LLMs chat assistants reaching new audiences?**
  - Generally, the plugin reached the same audience demographics, many of whom reported using the website more. While we see the potential to reach additional audiences in areas where we want to create/build audiences, this potential is mitigated by OpenAI: paywall, internal strategy, public opinion, legal rulings, etc..
- **How accurate are current LLMs in retrieving/summarizing answers to informational queries?**
  - Pretty accurate and relevant, with variation for different languages
- **What are the perceptions of credibility?**
  - People generally trust information more when they know it's sourced from Wikipedia

WIKIMEDIA
FOUNDATION

01 ChatGPT plugin usage and users

Data & Sources: Logging + Survey

Trends and product features move quickly in this space; insights shared are snapshots.

Logging data[3] (in yellow boxes)

- Collected through plugin instrumentation

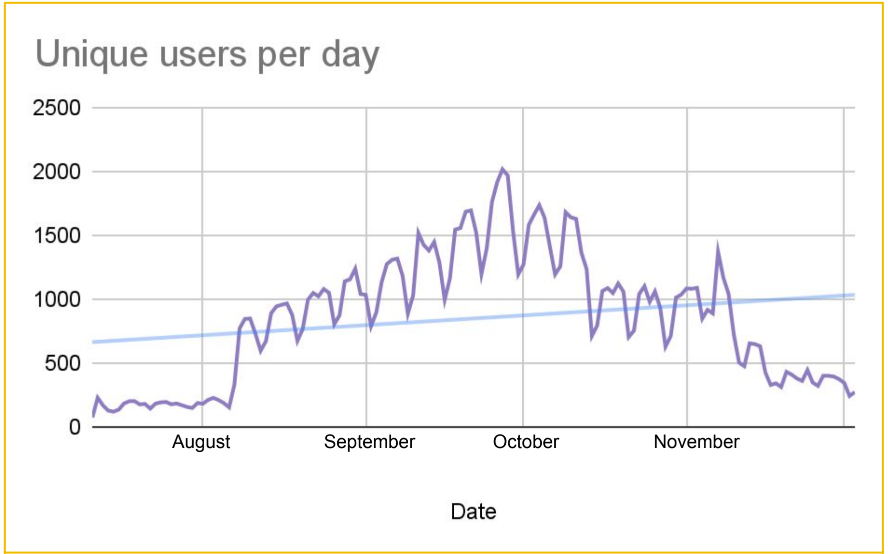
Survey data (in green boxes)

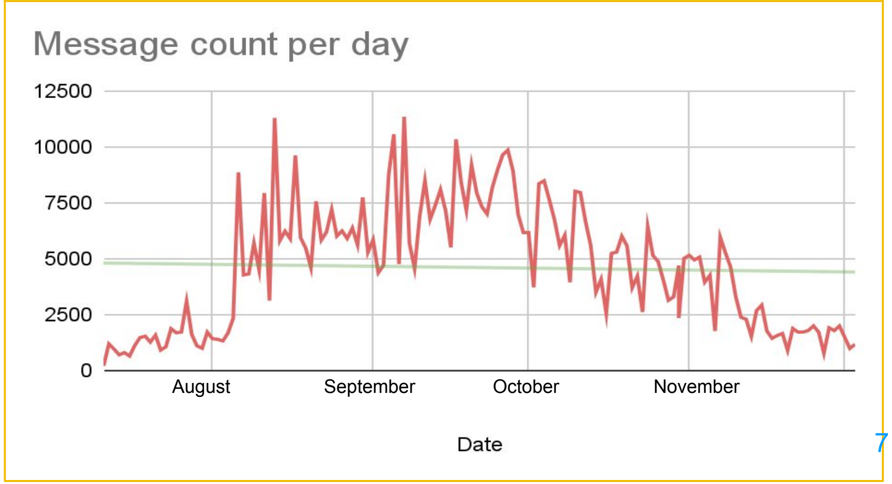
- Survey ran for ~2 weeks in September 2023
- 71 participants
- English language
- Plugin users were shown a link in the ChatGPT response, prompting them to take the survey
- Survey hosted on Google Forms

See footnotes

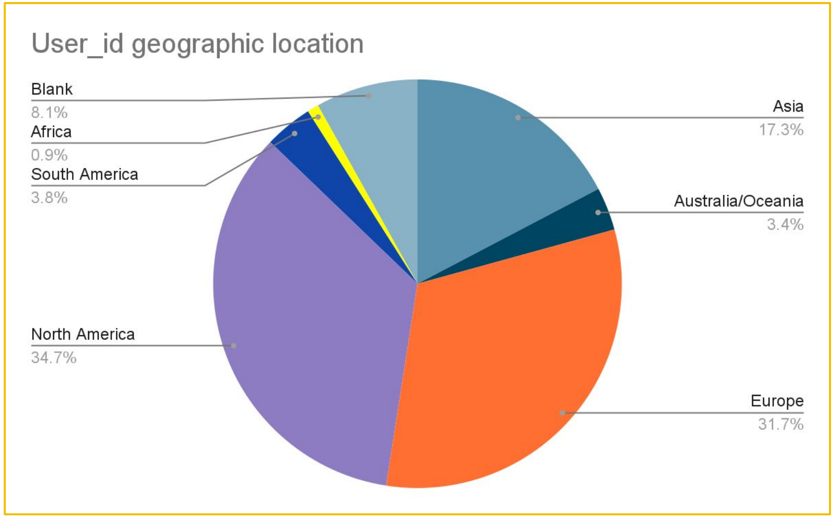
Users & Messages

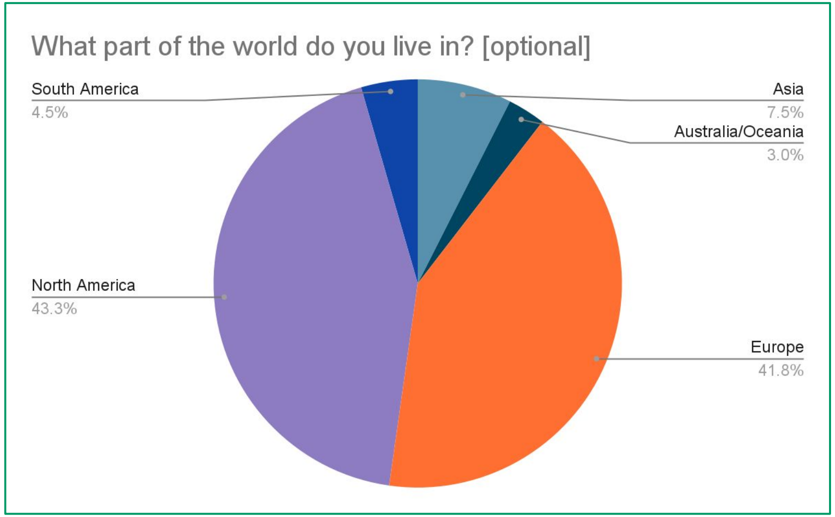
- Initial usage spike in early August 2023 resulting from media publicity
- Slow decline starting mid-September 2023
- Rapid decline to minimal usage in mid November 2023 when OpenAI deprecated plugins to move to GPTs


Unique users per day


Message count per day

2/3 of users are located in Europe & North America

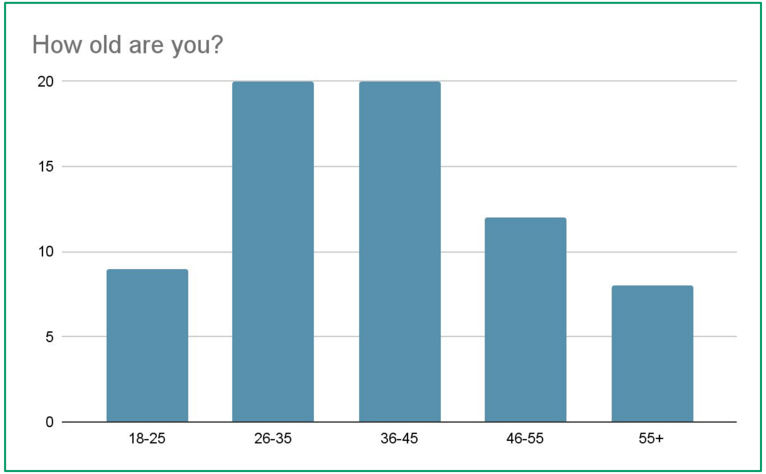


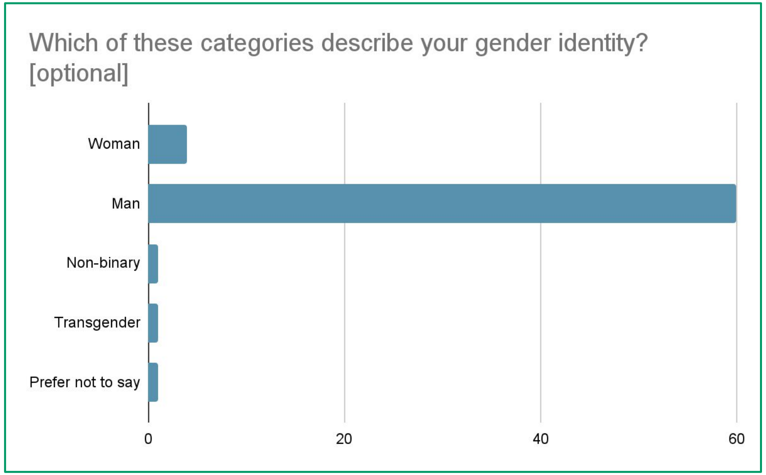


Top languages by number of messages: (relative ranking changes week to week)
English (en), Chinese (zh), German (de), French (fr), Japanese (ja)

Gender & Age

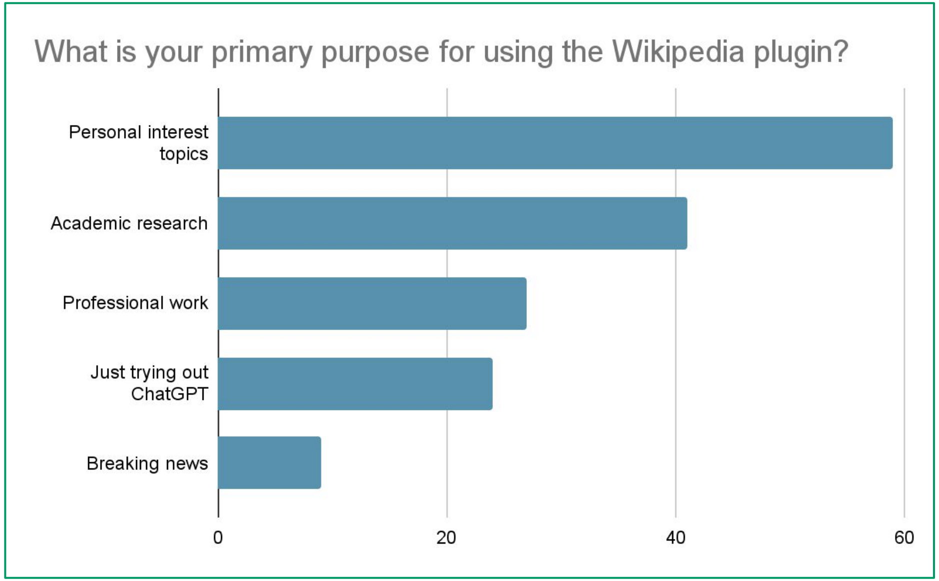
26-45 year old men comprise the majority of survey respondents





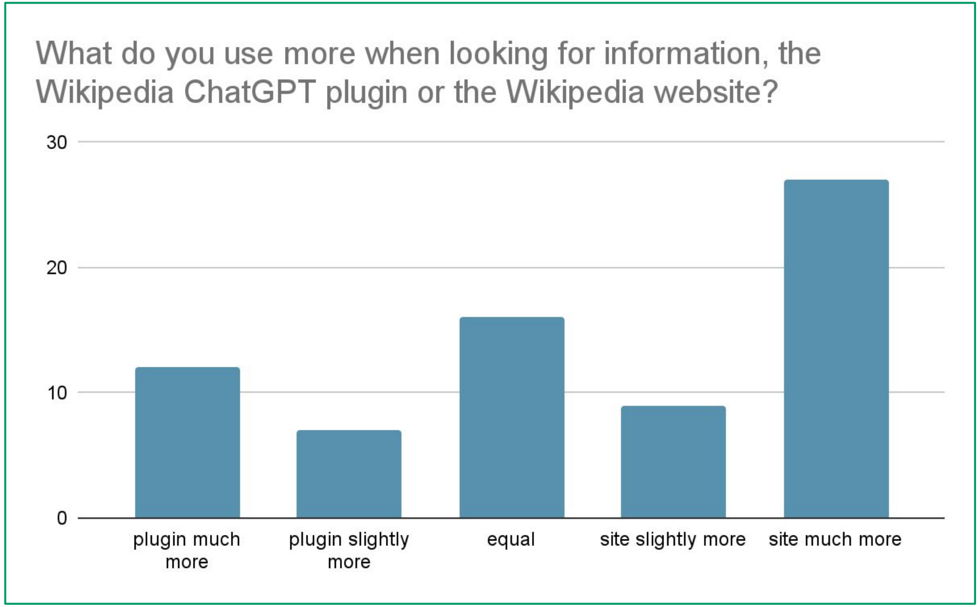
Use cases differ

People reported using the plugin for a variety of reasons



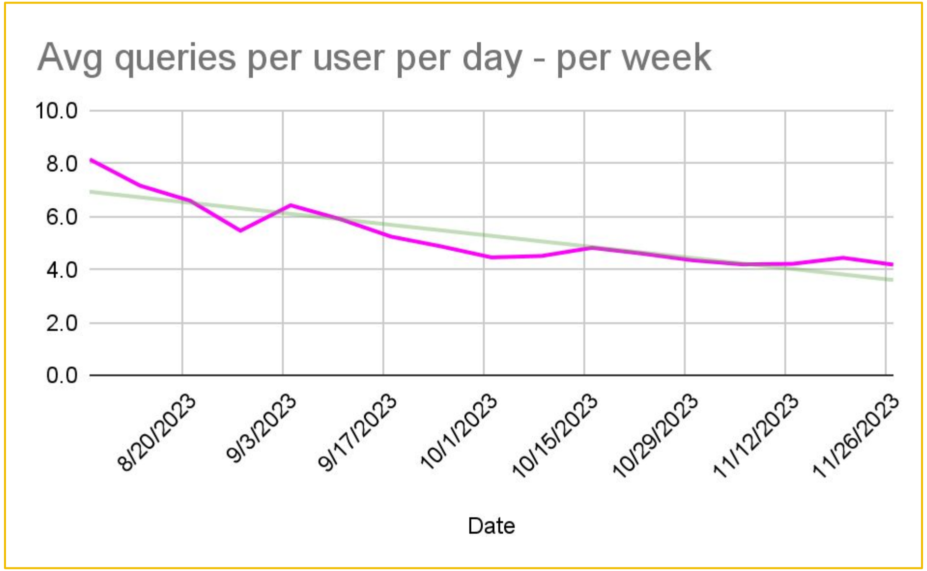
Generally, users didn't shift to the plugin to access information

Most users report using the Wikipedia website equally or more than the Wikipedia ChatGPT plugin



Plugin Click-Through Rate (CTR) comparable to Google search

- **Hypothesis:** lower CTR and higher queries/user may be indicative of conversational information-seeking where users ask more questions and make more comments instead of clicking through to read long form articles.
    - CTR from the plugin to English Wikipedia is roughly similar to CTR to our projects from Google[4].
    - Avg daily queries/user fell to ~4
- Takeaway: Users may be using the plugin like search rather than conversationally



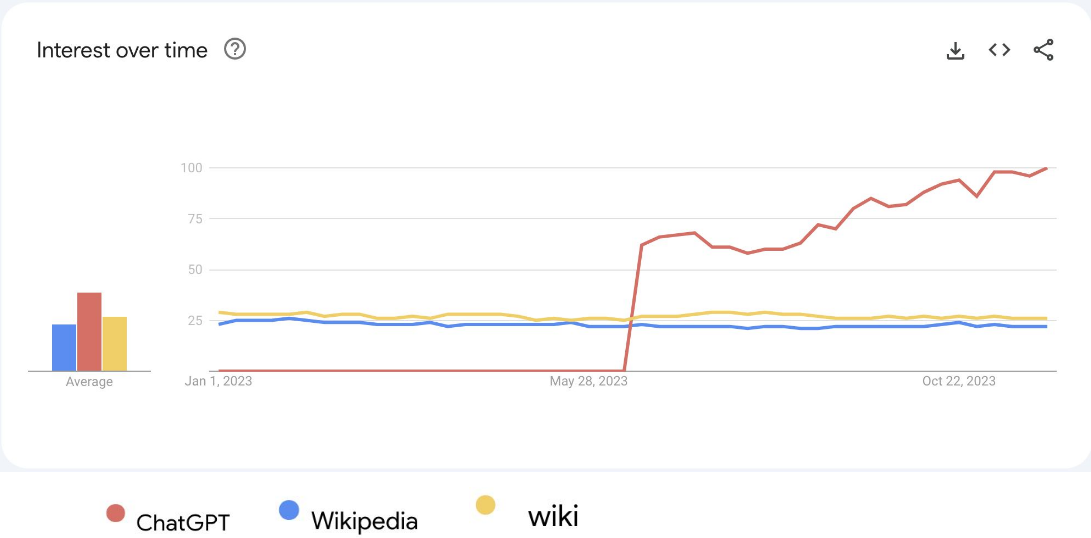
See footnotes

ChatGPT’s rise did not impact searches for Wikipedia

Increasing interest in ChatGPT did not affect interest in Wikipedia per Google Trends

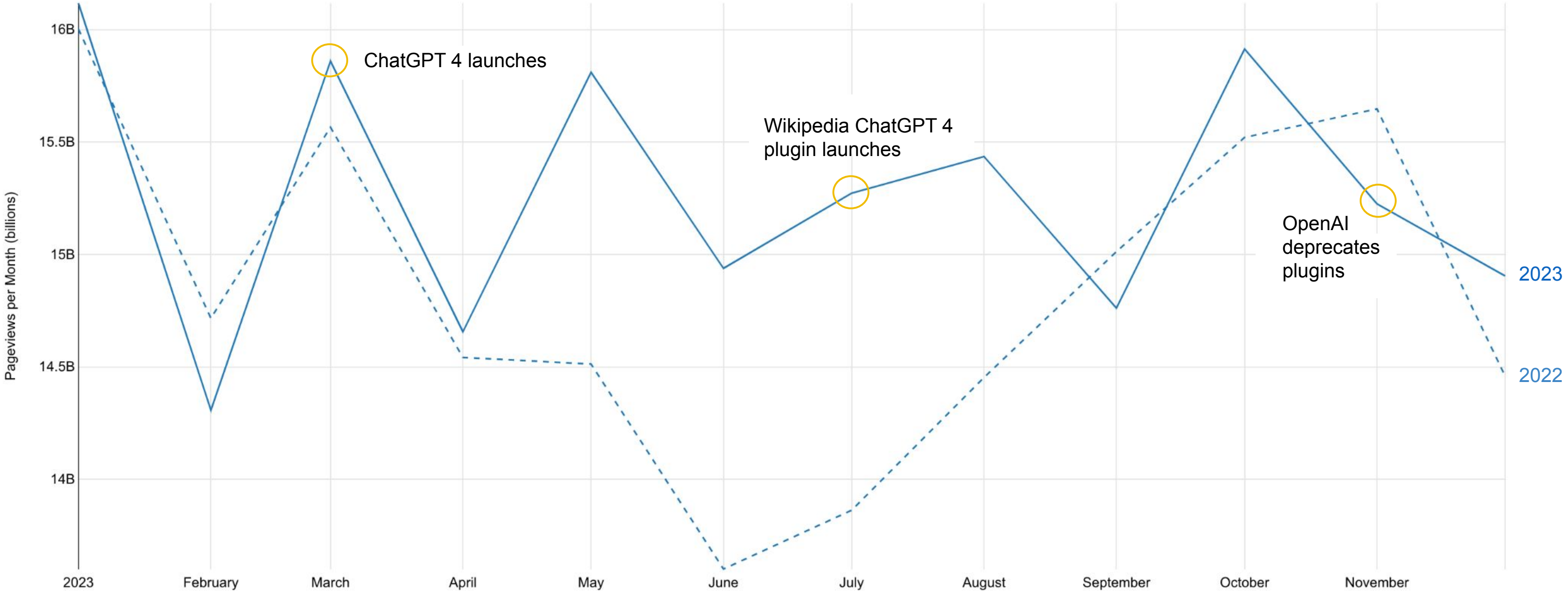
- No notable change in how often ‘wikipedia’ and ‘wiki’ show up in search terms[5]

Google Trends: searches over time



See footnotes

No major drop in readers during the meteoric rise in ChatGPT use[6]



See footnotes

Main questions

- **Are LLMs and/or chat assistants a new paradigm for information seekers?**
  - No/not yet. We do not see evidence at this time that people are using this in a way that is drastically different from search.
- **Are chat assistants displacing traffic to Wikipedia?**
  - No
- **Are LLMs chat assistants reaching new audiences?**
  - Generally, the plugin reached the same audience demographics, many of whom reported using the website more. While we see the potential to reach additional audiences in areas where we want to create/build audiences, this potential is mitigated by OpenAI: paywall, internal strategy, public opinion, legal rulings, etc..
- How accurate are the current generation of models in retrieving/summarizing answers to informational queries?
  - What are the perceptions of credibility?

02 Model Quality

Model quality - data sources

Internal annotation data (pink boxes)

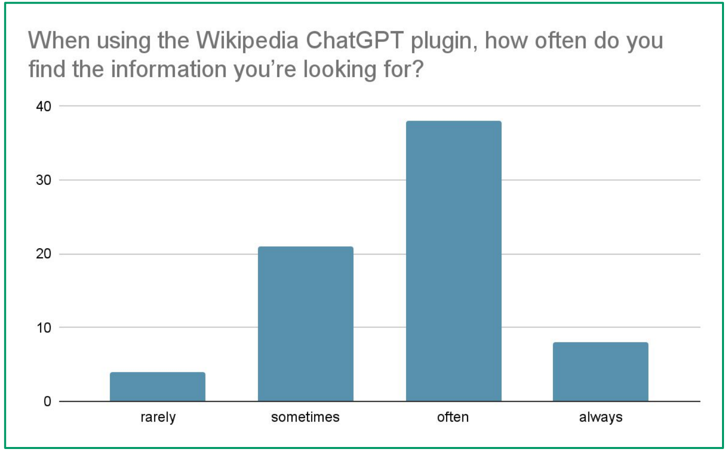
- Used a random sample of logged queries
- Focused on half dozen major languages (most represented, and we had speakers for)
- Did not review exact ChatGPT responses, but re-generated responses
- **Note:** Take these numbers with a grain of salt!

Survey data (in green boxes)

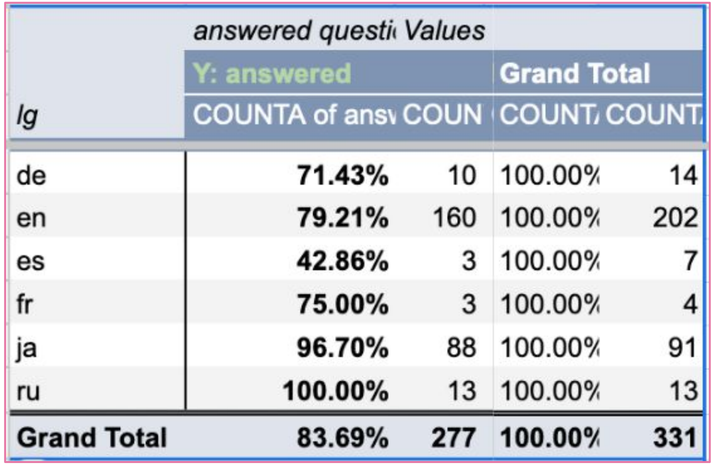
- Survey ran for ~2 weeks in September 2023
- 71 participants
- English language
- Plugin users were shown a link in the ChatGPT response, prompting them to take the survey
- Survey hosted on Google Forms

Relevancy

- **Relevancy:** Is the response to the query on topic (regardless of whether it is accurate or not)?
- Users often found information they were looking for
- ChatGPT answers were often relevant to the query
- This may reflect search relevancy more than LLM capabilities



When using the Wikipedia ChatGPT plugin, how often do you find the information you're looking for?

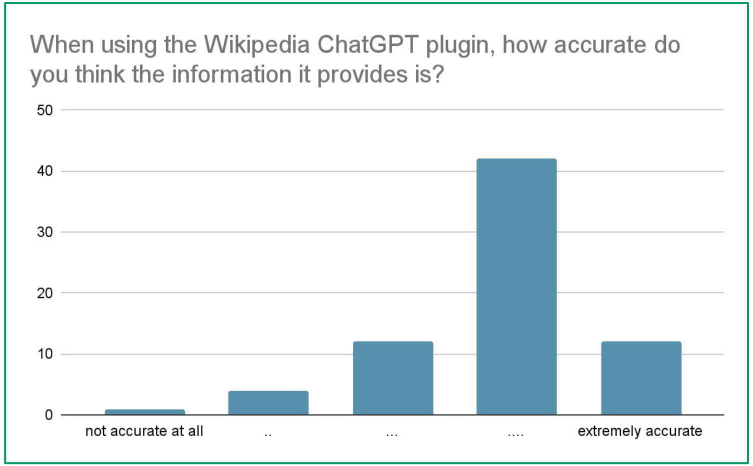


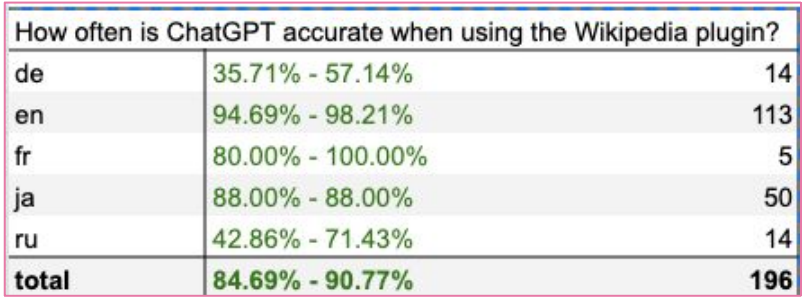
	answered questi Values			
	Y: answered		Grand Total	
lg	COUNTA of ansı	COUN	COUNT/	COUNT
de	**71.43%**	10	100.00%	14
en	**79.21%**	160	100.00%	202
es	**42.86%**	3	100.00%	7
fr	**75.00%**	3	100.00%	4
ja	**96.70%**	88	100.00%	91
ru	**100.00%**	13	100.00%	13
Grand Total	**83.69%**	**277**	**100.00%**	**331**

Accuracy

- **Accuracy**: Is the information in the plugin’s response factually true, according to the Wikipedia sources?
- Users perceive the plugin as being accurate
- Corroborates with internal evaluation that answer factuality is fairly accurate[7]
  - Low factuality in German and Russian may also be from individual annotators and low number of samples

Overall, ChatGPT does a good job staying factual when specifically prompted to restrict answers to Wikipedia content


When using the Wikipedia ChatGPT plugin, how accurate do you think the information it provides is?



How often is ChatGPT accurate when using the Wikipedia plugin?		
de	35.71% - 57.14%	14
en	94.69% - 98.21%	113
fr	80.00% - 100.00%	5
ja	88.00% - 88.00%	50
ru	42.86% - 71.43%	14
total	**84.69% - 90.77%**	**196**

See footnotes

Main questions

- Are LLMs and/or chat assistants a new dominant paradigm for information seekers?
- Are LLMs chat assistants reaching new audiences?
- **How accurate are current LLMs in retrieving/summarizing answers to informational queries?**
  - Pretty accurate and relevant, with variation for different languages
  - **What are the perceptions of credibility?**
    - People generally trust information more when they know it’s sourced from Wikipedia

03

Looking forward

OpenAI

- OpenAI has gotten rid of plugins to move towards a market of GPTs (personalized chat assistants)
    - Move towards more specific task-oriented AI
    - Low code/no code programming
    - Like plugins, this marketplace will be paywalled behind ChatGPT's subscription fees
- ChatGPT is not the only LLM in a highly competitive industry

While we should continue our partnership with them to further explore LLM/AI technology without needing to built it in house from scratch, as well as keeping an eye on how the conversational chat market emerges, **we should be cautious about being locked in with them too closely**

- Being dependent on a single partner, such as Google Search, can make us overly vulnerable

WMF Future Audiences: ‘Citation Needed’ extension experiment

- Experimental browser extension that reduces user effort to verify information on the internet using Wikipedia as a credible source
  - LLM behind the scenes, used for specific tasks (not a chat interface)
- Use AI to bring Wikipedia content to people not on the Wikipedia website (e.g. social media platforms)
- Position Wikipedia as a voice in the “Internet’s Conscience”: a beacon of credible information, and a model for creating it, in a new age of (AI-generated) misinformation

WIKIMEDIA
FOUNDATION

04 Conclusions

Conclusions - LLM technology

- LLMs, including ChatGPT, are pretty good at summarizing retrieved information
  - But people may be using them more like search currently
- LLM technology will likely be folded into more familiar internet tools and pathways
  - Traditional search will inevitably incorporate this technology
  - Retrieval Augmented Generation (RAG) – i.e. pairing search/information retrieval with generative AI, like the plugin did – is becoming an industry norm for keeping AI-generated content more on the rails

  while enabling other new ones:
  - Low/no code programming
  - Future Audiences: Wikipedia 'Citation Needed' extension

Conclusions - Market & Strategy

- People trust information more when it's transparently coming from Wikipedia
  - OpenAI partnership shows LLM brands can be willing to work directly with us, and there is a place for Wikipedia in the development of AI
  - As Retrieval Augmented Generation becomes more widespread, there is a growing WMEnterprise opportunity to provide a high quality pipeline to WMF project content for customers wanting to ground their generative AI in reliable information
- As more LLMs become available, developing an in-house WMF LLM will need to be carefully weighed for costs and benefits, and justified with specific use cases
- Wikipedia may reach some new audiences with conversational AI, but being too dependent on OpenAI mitigates this potential
- Hallucination and copyright issues with LLMs will continue to be ongoing legal and societal issues, and WMF has an opportunity to position itself as the Internet's Conscience

WIKIMEDIA FOUNDATION

Thanks!

Work in collaboration with Maryana Pinchuk, Nat Hillard, Chris Albon, Isaac Johnson, Michael Raish, and others at WMF and Wikimedia Enterprise

WIKIMEDIA FOUNDATION

mpham@wikimedia.org

Footnotes

1. Plugins were only available to ChatGPT Plus users on desktop.
2. ChatGPT uses its own internal AI reasoning to decide when to use installed plugins, such as the Wikipedia one. To do this, it uses both the user inputted text, and the text description that we provide to explain what the plugin does. When a user inputs a query into ChatGPT, the AI will decide which installed plugin to use, if any.
3. Logging Data related link: https://gitlab.wikimedia.org/repos/machine-learning/chatgpt-plugin/-/tree/dev
4. See more on web search engines.
5. Source: Google Trends

   As interest in ChatGPT increased, as measured by search trends, there was no notable change in how often 'wikipedia' and 'wiki' continued to show up in search terms, further pointing towards no noticeable shift away from current trends for getting to Wikipedia content -- even if new ones may be emerging
6. *Content interactions and pageviews are proxies for Wikimedia consumers. They reflect a combination of how many people come to our projects as well as how engaged they are. They do not capture consumption of Wikimedia content outside of our websites and apps.*

   In September 2023, ChatGPT had around 180.5 million users. The OpenAI website generated 1.6 billion visits in December 2023.

   https://www.forbes.com/sites/jodiecook/2023/12/06/6-giveaway-signs-of-chatgpt-generated-content/

   https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/
7. Some very active Wikipedians also tried out the plugin and gave us qualitative feedback that they found the results to be irrelevant/inaccurate. This may indicate a there are potential differences in expectations of quality between individuals within our community and an average knowledge seeker.