WIKIMEDIA FOUNDATION

Citation Needed experiment summary

June 2024

Mike Pham, Irene Florez

Work in collaboration with Maryana Pinchuk, Nat Hillard, Chris Albon, Elias Rut, Daniel Erenrich, Michael Raish, and others at WMF and Wikimedia Enterprise

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 is 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
  - Future Audiences is currently working on Citation Needed: ChatGPT API + Cirrus Search API.

High Level Summary

- **Feasibility**: does the technology **make it possible** to verify internet content using Wikipedia?
    - Yes, it works okay even with low investment.
    - Making it work *really* well will mean some substantial investment.
- **Demand**: do people **want** to verify internet content using Wikipedia?
    - Qualitative feedback is all positive
    - Quantitative data is not encouraging – low/no retention
    - Mitigating factors
        - Chrome Extension is not the best vector for this
        - Websites are not the part of the internet where this is needed
        - Extension requires lots of proactivity from the user

Options going forward

Not mutually exclusive

1. Turn it off
2. Improve it (e.g. fact check entire page)
3. Market broadly to get more usage data
4. Show to partners like YouTube and TikTok
5. Extend with “add a fact”

Briefing

- To meet the public's demand for increased accuracy, substantial improvement in model performance is needed and will require dedicated investment in ML engineering (and search API improvements for Retrieval Augmented Generation).

- While sentiment around this work is overwhelmingly positive, there is little organic growth at the moment. Possible factors include: Chrome extensions not being discoverable, AI trust, UX challenges, high communications (publicity) needs.

- With AI products, there is an initial test period for users to gain trust. With Citation Needed we have not yet cleared this hurdle of becoming truly useful in a consistent way.

- The majority of verifications (81%) were not coming from news or social media sites

- Roughly half of verifications could be checked on Wikipedia. Wikipedia does not yet contain all of the possible information that people care about.

Future Audiences: Citation Needed

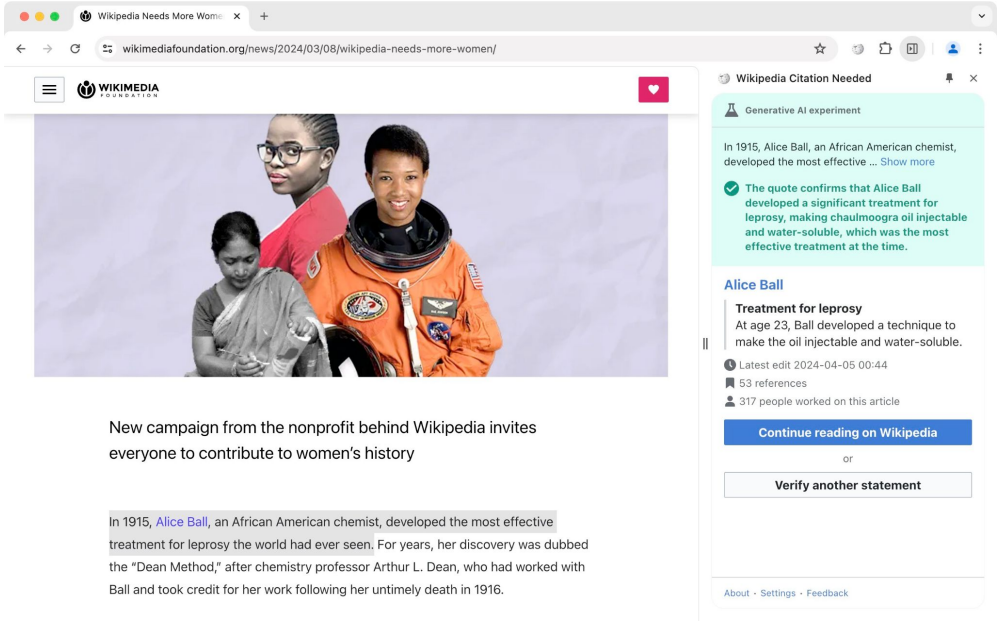
- Users can verify sentence-sized claims using Wikipedia, seeing:
  - Relevant wikipedia article
  - Most relevant quote from article
    - Credibility signals
  - Verification status, and explanation

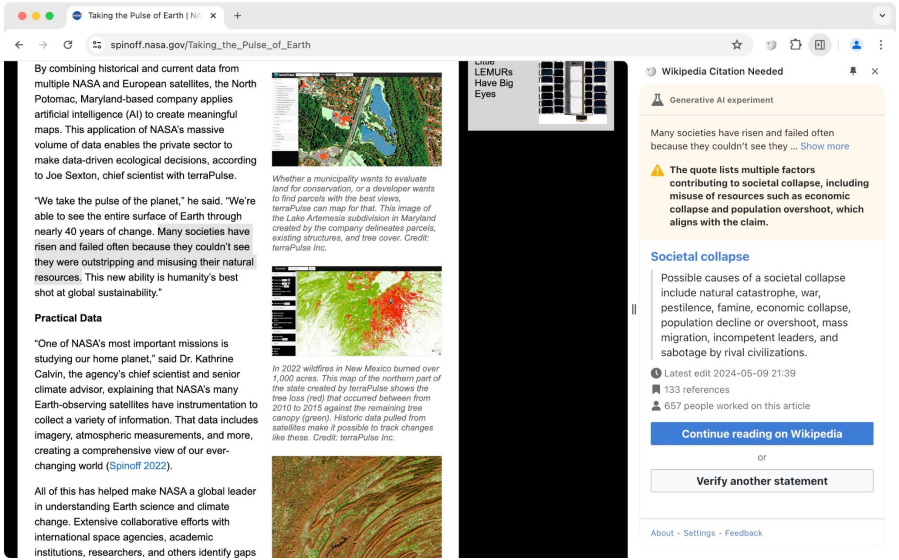
- Chrome browser extension
  - Can reach large audience
  - Don’t need to build from scratch
- ChatGPT API + Cirrus Search API
  - AI-augmented search/information retrieval
  - Retrieval Augmented Generation (RAG)

Scope

- English only
- Text only
- 300 character input claims
- Desktop only
- Chrome browser only

What it looks like





Try it out: chromewebstore:wikipedia-citation-needed

WIKIMEDIA FOUNDATION

Big Questions

What do we need to learn about the **Internet's Conscience** from Citation Needed?

1. **Do people <u>want</u> to verify** website content?
2. **Do people trust Wikipedia content and brand** as a reliable source of information?
3. **Can facilitating off-platform knowledge verification help us reach new audiences or current audiences in new ways?** (i.e., will it create a new vehicle for donations? For knowledge contribution?)
4. **Can we rely on generative AI to assist in knowledge verification?** (i.e., searching for, retrieving, and making inferences about the content on Wikipedia?)

What do we need to learn about the **Internet's Conscience** from Citation Needed?

1. **Do people browsing the web <u>want</u> to verify content** they're seeing on other websites?
    a. People are concerned about misinformation and AI hallucination, and are enthusiastic about Wikipedia's fact checking work. In practice, they are not using our tool that much.

2. **Do people trust Wikipedia content and brand** as a reliable source of information? (i.e., Is it considered credible?)
    a. Mixed bag. Online discourse surrounding Wikipedia's chrome extension 'Citation Needed' is overwhelmingly positive. That said, the most recent preliminary study was inconclusive. More studies are needed to gauge trust levels.

What do we need to learn about the **Internet's Conscience** from Citation Needed?

3. **Can facilitating off-platform knowledge verification help us reach new audiences or current audiences in new ways?** (i.e., will it create a new vehicle for donations? For knowledge contribution?)
    a. Some positive feedback around establishing Wikipedia as a knowledge source, but limited organic growth so far

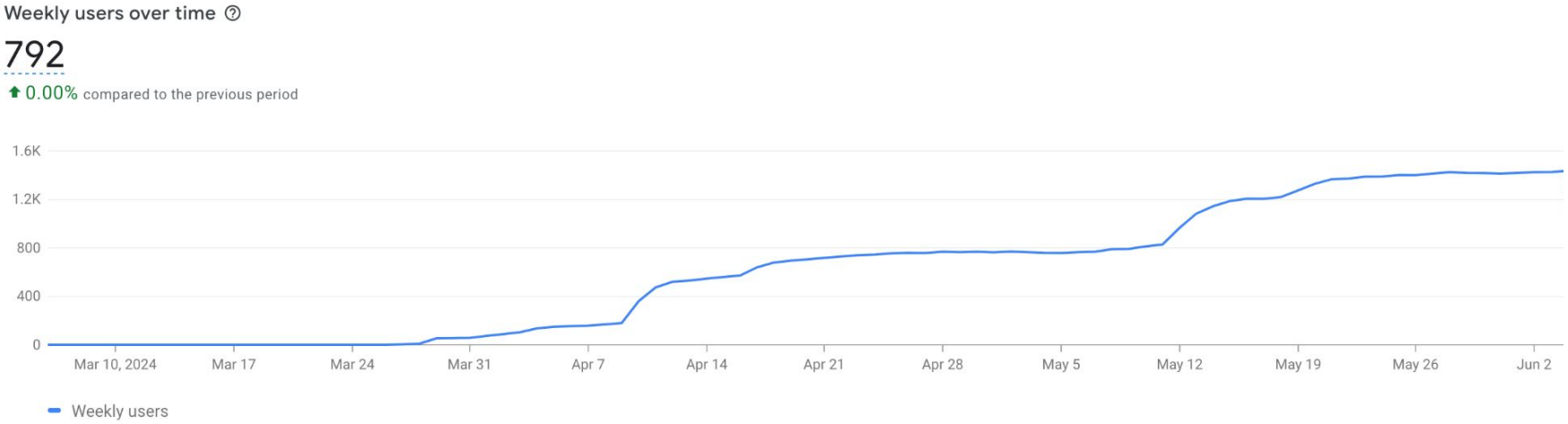
4. **Can we rely on generative AI to assist in knowledge verification?** (i.e., searching for, retrieving, and making inferences about the content on Wikipedia?)
    a. Gen AI is tricky to deal with, but has been overall doing a good job of finding and summarizing relevant information. Some verifications are handled nicely, especially those with direct statements to verify.

WIKIMEDIA FOUNDATION

01 Citation Needed usage and users

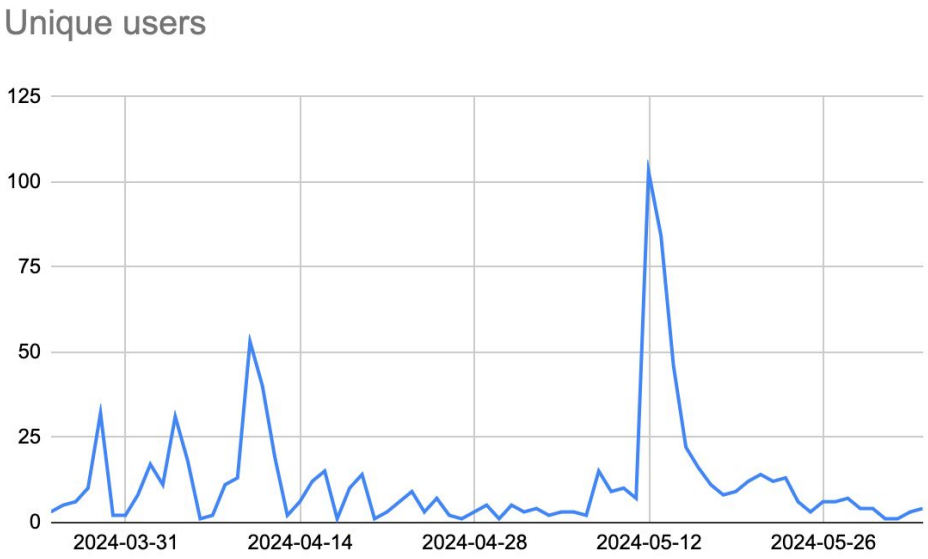
Users

- Spikes in new users following public announcements and posts, but little organic growth
  - Chrome extension webstore is not the best place for organic discovery
- Extension retention: users predominantly install the extension and keep it
  - US users often uninstall extensions that don't meet basic expectations, cause performance issues (browser performance and speed), seem intrusive or have unclear data policies.



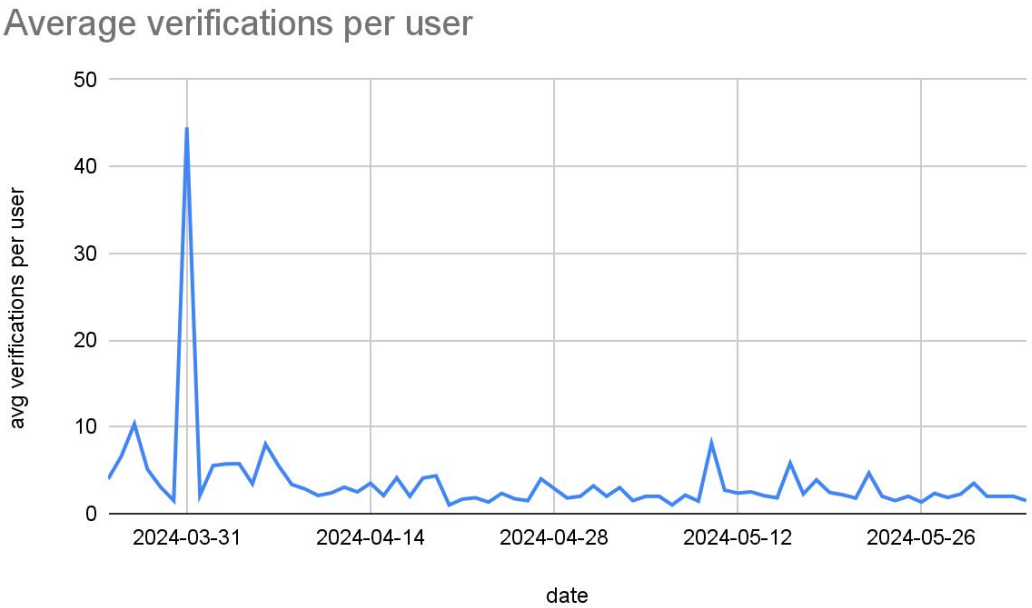
Unique daily users

- Max daily unique users reached 103; far from our goal of 1k+ unique daily users.
- We saw spikes in usage followed by quick decline, implying people did not use it regularly
  - Read about engagement challenges on the slide, “Barriers to use”



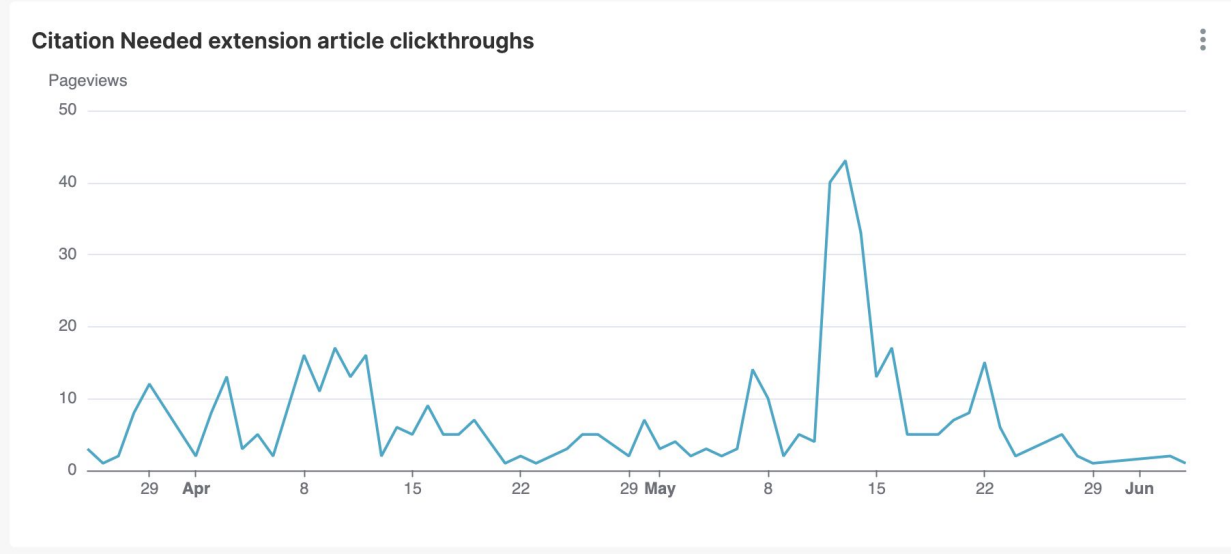
Average verifications/user/day

- Roughly 3-6 verifications per user on average
- Overall users spikes and drops to single digits
- Is this a sign that we're not reaching new audiences? Are power user Wikipedians a minority that use our products, with others showing less interest?



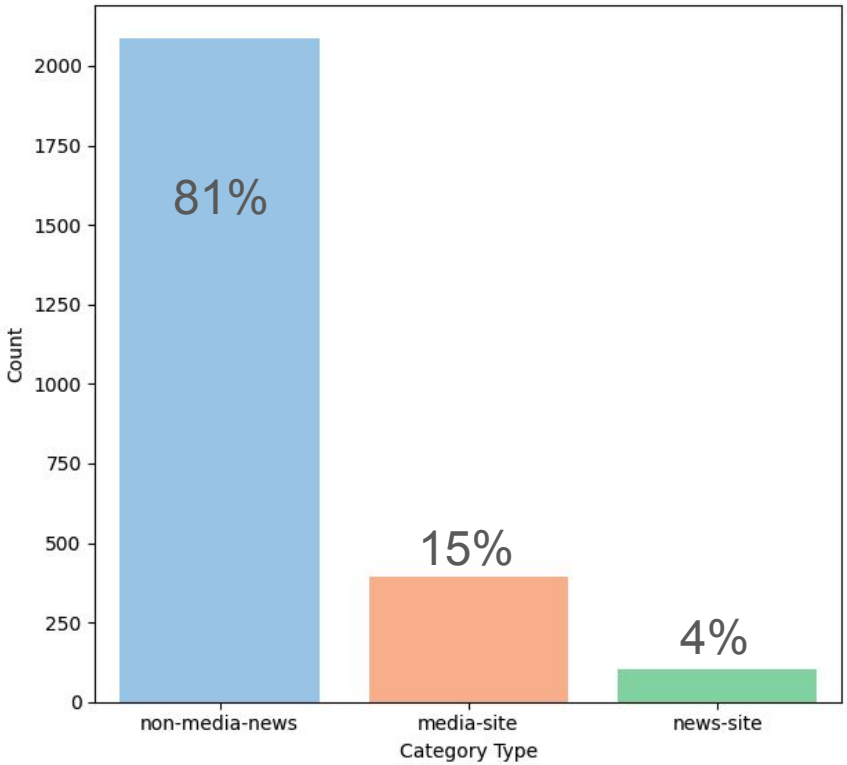
Click Through Rate (CTR)

- CTR proportional to usage, but remain relatively low
- CTR remains steady even at low usage volume, implying that the users who continue to use Citation Needed find links helpful
- Unlikely to encourage drastic increase in Wikipedia visitation



Referral source

- Most usage is happening on sites that are neither news nor social media sites
  - Low social media usage may discourage social media platforms from natively building out their own version
  - In this experiment we did not set out to track specific URLs that were not from media sites (see list in notes)
- Where else are users trying to verify info?
  - One survey taker mentioned using it to check ChatGPT answers



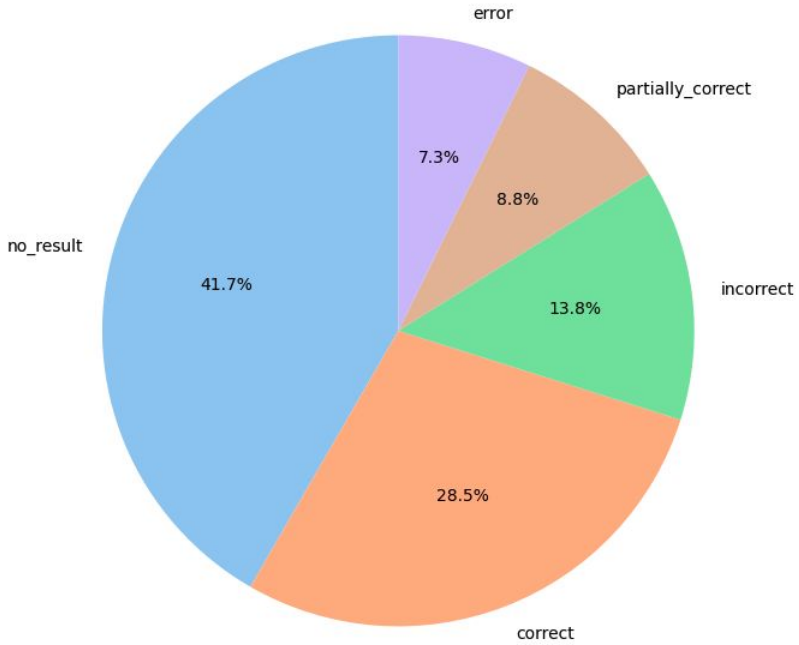
Verification results

- Roughly half of verification requests could be attested (correct, incorrect, partially_correct)

- 29% of verifications were correct
  - Might be a sign that people are checking things they know to be true in order to test the extension

- 42% of verifications have no result
  - These may be topics where Wikipedia does not have enough information to verify these claims: e.g. local news, opinions, recent events, etc.
  - … or where search failed

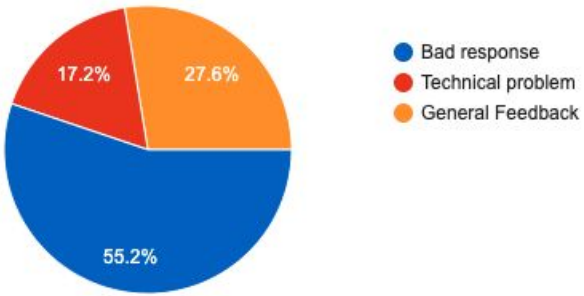
- 7% of verifications resulted in an error



Sentiment

- While use is low, the feedback is encouraging, highlighting the need for this tool
- Opportunity to consider: how could an improved version of this tool integrate into internet usage more seamlessly?
- Citation Needed feedback form responses:

Types of tool feedback provided



Open-text general feedback:

If this goes further, getting it approved so that a "Proceed with caution" message isn't required would probably help.

Just started using this, but overall I think this is a GREAT idea.

I inquired about Pope Francis and received information about Andy Warhol. Tool still needs work.

This is my first use, but I will continue to use the plug-in for more information.

I really like the idea of the plugin and I wanted to provide feedback to help improve it in its infancy :)

It's great! (Love if it worked on firefox, though)

So far i like it :)

Trust in Wikipedia (in progress)

- Additional survey data + interviews are needed
- Sentiment: people are keen on anchoring LLMs to better factuality
- Some users (internal & external) expressed concerns that Wikipedia is not always the best source for factual information (echoed by students in the *ChatGPT in the classroom* study)
    - Showing the quality signals helped instill some trust

WIKIMEDIA FOUNDATION

02 Model Quality

3 optimization parameters

- We focused on improving:
  - **Accuracy** – this used a manually created ~35 item test set that simulated difficult uses of Citation Needed; we also used a 100 item sample of an industry-standard FEVER data set (which strangely included noticeable errors) which was adjacent to our use case
    - After the internal release, we also did a manual annotation task of a 100 query sample of logged queries
  - **Consistency** – we worked to lower the temperature of LLM behavior so that verifications over the same queries would produce as close to the same result as possible
  - **Latency** – we wanted to reduce the lag as much as possible; most of the latency was a direct result of ChatGPT

For our internal release on March 26, 2024:

- Accuracy: 60% (note: ‘partially correct’ was counted as a correct response)
- Consistency: 92%
- Latency (average): 17s

Manual annotation

- When scoring a sample of real queries by hand, 71% were acceptable responses (includes failing gracefully)
- <1% were harmful
- 29% were unreasonable queries: e.g. code snippets, email signatures, other things that were not claims
- 17% were querying information that would not reasonably ever be on Wikipedia
    - This did not include breaking news queries, which would be unlikely to be on Wikipedia at the time of verification, but might eventually be on Wikipedia

See more: Citation Needed test sentences

WIKIMEDIA FOUNDATION

03

Looking forward

Barriers to use

- Hard to use regularly
  - UX
    - Scope components imposed notable requirements, for example: Citation Needed required using Chrome browser extension on desktop in English.
    - Change of habits: Using Citation Needed adds extra steps to online use which can break the flow of reading an article. Also, it requires the user to initiate the process (actively want to verify information). Some users are starting to verify information on third party websites (for example checking SEO keyword validity on SEO websites after receiving an AI generated SEO keyword list). Making the verification process and the habit loop overall more seamless is key to reaching a wider swath of users.
    - Too much work to verify sentence by sentence; creation of inefficiency.
  - Use case. Users want better factuality in the tools they already use (such as ChatGPT), without having to resort to another tool that requires additional effort.
  - Little positive reinforcement. What people want to verify is often not on Wikipedia; only half of claims can be verified (ie: news (breaking, local, specific), opinions, AI hallucinations, etc.).
    - As is, Wikipedia alone is unlikely to provide all relevant information. If we're not yet at a threshold of having enough content, people will default to using tools that are more comprehensive.
    - Verifications may not be on the whole satisfying to users; there may be gap between user expectation and the tools capacity to verify.
  - There may be experiment fatigue if we are reaching the same audience as did the ChatGPT plugin. How long can we keep people's attention with Future Audiences experiments if we release something new every 3 months?
  - Verification success is more likely where clear and succinct sentences are selected. The sentences that most provoke fact checking may be more complex or convoluted.

- Popular information gateways like Google and Siri already search Wikipedia. At present there is no clear reason for users to directly seek a specific Wikipedia-powered information product.

Improving ML models

- Gen AI worked surprisingly well in providing explanations of verifications
  - Turned the AI explanation on by default after noticing it more generally lowered risk of misinterpreting results due to missing context
- We were able to achieve acceptable accuracy based on our internal test sets
  - It was unclear if our bespoke test sets are representative of real world queries
  - Each ML experiment we run will likely require designing and creating a test set based on the task we are doing – this is time consuming
- Real improvements would require dedicated ML engineering efforts
  - We picked off the low hanging fruit improvements, which we exhausted relatively quickly
- For RAG pipelines like ours, CirrusSearch can sometimes be the weakest link
  - We use LLMs to extract search terms, but CirrusSearch can be a little clunky at times
  - It's possible this is one place where vector search could be a benefit: i.e. CirrusSearch API for RAG applications

See more: Citation Needed API experiments spreadsheet

Recommendations

- People want increased accuracy in their current tools. **Work with LLM creators to directly better anchor their models** in credible sources(i.e. can Wikipedia-powered RAG built directly into ChatGPT, Gemini, Claude, etc help instill better trust?), LLM repackagers like iOS, Firefox, etc., or provide an easy to use RAG API that enterprise LLMs can easily utilize
- **Focus on improving search & browsing; and think about Wikipedia-content as *specific* vs general purpose**
  - Encyclopedic content is one of many content types. If we are asking users to use a single-use tool for a multi-use purpose they will experience failure. Notably, general-purpose was the guiding experience for Google search – unify & index all web knowledge. But to aim for coverage, Google does not fact-check.
  - If Wikipedia is a destination that provides better usability/interactivity with our specialized content in a way that Search and AI tools won't replicate (as they are more general purpose), higher quality experience over a more limited search domain may encourage more usage of in-house tools; though these may not reach new audiences.
    - i.e. Netflix search only searches Netflix, which makes it more useful than using Google. Due to Netflix search's desirable domain restriction on search results, most people do not use Google search to find a Netflix movie to watch. Thus, Netflix' in-house search exists side by side with Google search.
- **Improve readers' experience by increasing content types and reducing gaps**
  - Reconsider what kinds of information should be on Wikipedia: is there value in expanding the types of information now on Wikipedia (e.g. include breaking news, etc.)?
  - How do we increase locally relevant content?
- Alternatively, **create a larger alliance with credible news sources** – i.e. news outlets, etc – that would provide a more comprehensive database of up to date credible information (for a RAG API).
- Humans want to give feedback to improve the models. Consider creating an api to make it easy to select and comment on the components involved and **capture actionable feedback systematically**.