Extended: Correcting Social Claims using WikiAgents

Swapneel Mehta
SimPPL and MIT

Dhara Mungra
SimPPL

Raghav Jain
SimPPL and UCSD

Abstract

This proposal seeks to establish Wikipedia as the verifiable knowledge layer for artificial intelligence (AI) agent ecosystems through three innovations: 1) Protocol-agnostic claim verification gateway integrating Wikidata ProVe, 2) Decentralized attestation system for cross-agent audit trails,and 3) Creator Toolkit enabling influencers to validate multimodal content. Through agents tasked with correcting false claims online when requested, Wikimedia will grow public awareness and unlock future audiences for their platform. The 24-month initiative directly advances Wikimedia's multigenerational strategy by creating technical infrastructure for trustworthy AI-human collaboration while expanding youth engagement through verifiable content creation.

Introduction

The introduction of AI agents has reduced the internet to a dataset with companies battling each other over how to generate real user data[1] as a continuous stream of input to keep training larger foundation models without model collapse[2]. With the rise in agentic activity came the introduction of Anthropic AI's Model Context Protocol (MCP), which can colloquially be described as a "set of guiding standards to help agents navigate actions across multiple platforms". Wikipedia is already engaged in operationalizing its knowledge base for startups who are using it as a provenance tracking service, and a number of so-called AI fact-checkers[3] that are also employing it to advance the identification and verification of claims. In this proposal we ask, **can we make Wikipedia the *credibility layer* of the internet?** Emerging AI agent protocols (MCP, A2A, NANDA) lack robust mechanisms to combat hallucination and bias. Agent-generated claims in preliminary audits contained factual errors (Wang, Wang, Iqbal, et al., 2024; Wang, Wang, Manzoor, et al., 2024). Wikipedia's structured knowledge base and provenance tracking systems offer unique solutions but remain underutilized in agentic ecosystems. We explore the following questions through this research:

H1. Can free tools for fact-checking claims in multilingual formats increase public participation in reporting misleading claims?

H2. Could social claims verification elevate the utility (page views, engagement) of the Wikipedia ecosystem?

H3. What incentive structures maximize creator participation in claim verification?

H4. Does decentralized attestation improve trust in multi-agent systems versus centralized alternatives? (could WikiAgents work better than fact-check labels?)

[1] Kylie Robison, Alex Health, Verge, 2025

[2] Model collapse has suffered its recent share of controversy from the claim of synthetic data causing it even in nominal volumes. See: Dohmatob et al., 2025

[3] Facticity, Factiverse, Originality, Sourcebase, others.

Date: Sept 1, 2025 - Sept 1, 2027.

Related work

Wikipedia content is shared across the most unexpected pockets of users across social platforms: for example, we find them on Truth Social, the conservative alternative that President Donald Trump created when he was banned from Twitter/X (Shah et al., 2024). There are pockets of rational discourse engaged in the use of Wikipedia articles even on a platform otherwise painted as ideologically congruent, filled with low-credibility information sharing individuals (Zhang et al., 2025). While there are claims that the advent of AI agents will diminish the value of Wikipedia directly (Wagner & Jiang, 2025), we believe it might also contribute to novel revenue streams given the increased usage of the platform by AI agents in seeking to define and refine research concepts using human-verified data sources–the primary source being Wikipedia.

Methods

The proposed systems advance Wikimedia's strategic pillars by operationalizing three decades of community-driven consensus building. We construct an observational study in which we investigate social media influencers viz. the content they produce online; starting from a list of top creators gathered from SocialBlade, we extract claims and process them as they pertain to different categories. Using publicly available data and social media analytics tools, we compile a list of popular influencers who have large followings and whose content frequently includes statements about brands, products, or widely discussed topics. Once we have this list, we collect a sample of posts from these influencers over a set period. We pay special attention to posts that make factual claims—whether in text, video captions, or images. For example, if an influencer says, “This supplement boosts your immune system,” or shares a meme claiming a certain event happened, we record these statements for further analysis. In particular, results are hashed and stored on-chain, with contextual metadata (timestamp, agent ID, source URLs). The governance of these claims will be important and can draw on crowdsourced models like Reddit’s moderation, Community Notes, or News Detective. We showcase this workflow in the Appendix.

The next step is to verify these claims. We use automated tools that compare the influencer’s statements to information available on Wikipedia and Wikidata. If a claim matches information that is supported by Wikipedia, it is marked as verified. If the claim is not supported or contradicts what is found in Wikipedia, it is flagged for further review. For claims that are disputed or unclear, we also look at how Wikipedia editors and the broader community have discussed these topics on talk pages, which helps us understand if there is consensus or ongoing debate. In addition to automated checking, we monitor how often influencers correct their claims after being presented with verified information. For example, if an influencer is notified that a statement they made is inaccurate, we track whether they update their post or share a correction. We also measure how this process affects engagement with Wikipedia, such as whether there are increases in page views, social media sharing, or more people editing Wikipedia articles on related topics.

Our goal is to demonstrate this use-case for social media influencers as a flagship example of how our agentic protocol can be applied, and allow online AI agents to access an endpoint and use our system for in-conversation claims verification as our study concludes.

What we have already done/have ongoing:
Our team and collaborators have already worked in this area and built the following:

1. Factiverse's claim detection models (89% F1-score in multilingual tests)
2. PublicEditor's browser extension architecture with 320ms median latency
3. Our platform[4], Arbiter's cross-platform data lake containing 250M+ social media interactions.

Our ongoing work and collaborations position us well to demonstrate not only existing impact, but also technical capacity, product vision, and team alignment towards a single goal.

Expected output

We target an audience of AI developers, students, and researchers in the credibility community. For developers, our key deliverable is an SDK (Python, JS) for them to program low-latency (<30 secs) **claims verification agents** as well as **results hashing protocols** when it gets too challenging to maintain locally. For students, we provide a means to fact-check their favorite (top 500 in different categories) influencers and clear their feeds of misleading information up engagement for Wikipedia as a reliable and community-governed source of truth. And for researchers, our combination of social media data, engagement, Wikipedia data, would be an invaluable source of research data across social platforms in tough times for researcher data access from platforms. We aim to publish our work in Management Science, AAAI, ICWSM, or equivalent venues.

Risks

First, the mechanistic issues: token based control of the system including usage-based pricing will be necessary to build an economically sustainable solution and track any abuse of the system by bad actors trying to game it. We will also face the social and reputational challenges. There is also the risk of legal or reputational issues arising from the involvement of influencers. If an influencer shares unverified or controversial claims, it could expose Wikimedia to criticism or regulatory scrutiny. To mitigate this, we will develop clear guidelines for influencer participation, provide training on responsible information sharing, and establish rapid response protocols for addressing disputes or misinformation. This has been done before in terms of the policies that social media platforms define to moderate hateful speech on the platforms[5] and utilised by others to craft civic integrity policies including hate speech.

Community impact plan

We have already established conversations with Wikimedia Brasil and work with WikiCred members in the U.S. who are highly active developers. For Wikimedia developers, we will provide open-source software development kits (SDKs) and detailed documentation, encouraging them to integrate our verification tools into existing bots, gadgets, and workflows. We will host hackathons and developer sprints to support this integration, and offer recognition and support for innovative community-led adaptations. Beyond the Wikimedia movement, we will engage with social media influencers and content creators, providing them with easy-to-use toolkits for claim verification. By partnering with popular creators, especially those with large youth followings, we can promote the adoption of

[4] Formally, SimPPL is the recipient applying for this grant. But we aim to work with the other teams mentioned on the products they have asked us to take over e.g. PublicEditor (no longer actively maintained, but perfectly functional).

[5] Meta's Hateful Conduct Policy, accessed 2025

reliable information practices and drive new audiences to Wikipedia, expecting effects that are similar to prior work in the context of islamophobia (Alrababa'h et al., 2021).

Evaluation

Adoption by both users and influencers will be a topline metric to evaluate our system. Considering our study design, other outcome variables include the system performance metrics, qualitative research studies, and user interviews to gather community feedback directly. Through the engagement and pageview data, our approach allows us to measure not only the accuracy of information shared by influencers but also the effectiveness of Wikipedia as a tool for improving the quality of online discourse.

Throughout the process, we take care to respect privacy and ethical guidelines. We do not collect personal information about influencers or their followers, and we ensure that our analysis is balanced across different viewpoints and communities. By making the verification process transparent and accessible, we hope to encourage more influencers to participate in fact-checking and to help their audiences find trustworthy information online.

Budget

The budget narrative allocates expenses between the enlisted personnel: research staff, postdoctoral (project lead) researcher, project lead, coordinator, and 2 software engineers. Rates are at market price in India, where the team will be based. The software costs are based on bi-annualized actual costs accrued by our current Google Cloud Platform hosted servers and API endpoints including LLM API calls.

References

Alrababa'h, A., Marble, W., Mousa, S., & Siegel, A. A. (2021). Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes. *American Political Science Review, 115*(4), 1111–1128. https://doi.org/10.1017/S0003055421000423

Shah, C., Konka, R., Malpani, G., Mehta, S., & Ng, L. H. X. (2024). *Can Social Media Platforms Transcend Political Labels? An Analysis of Neutral Conservations on Truth Social* (No. arXiv:2406.03354). arXiv. https://doi.org/10.48550/arXiv.2406.03354

Wagner, C., & Jiang, L. (2025). Death by AI: Will large language models diminish Wikipedia? *Journal of the Association for Information Science and Technology, 76*(5), 743–751. https://doi.org/10.1002/asi.24975

Wang, Y., Wang, M., Iqbal, H., Georgiev, G., Geng, J., & Nakov, P. (2024). *OpenFactCheck: Building, Benchmarking Customized Fact-Checking Systems and Evaluating the Factuality of Claims and LLMs* (No. arXiv:2405.05583; Version 2). arXiv. https://doi.org/10.48550/arXiv.2405.05583

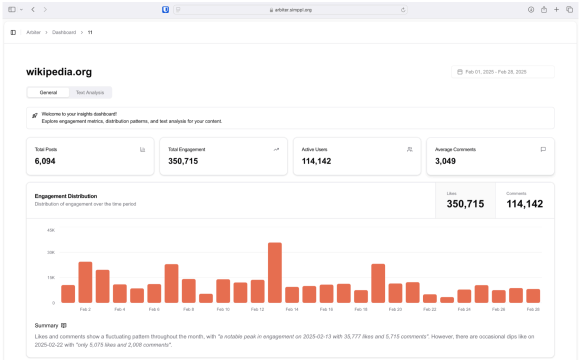
Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G., Das, R. J., & Nakov, P. (2024). *Factuality of Large Language Models: A Survey* (No. arXiv:2402.02420). arXiv. https://doi.org/10.48550/arXiv.2402.02420

Zhang, Y., Lukito ,Josephine, Suk ,Jiyoun, & and McGrady, R. (2025). Trump, Twitter, and Truth Social: How Trump used both mainstream and alt-tech social media to drive news media attention. *Journal of Information Technology & Politics, 22*(2), 229–242. https://doi.org/10.1080/19331681.2024.2328156

Appendix

A. Arbiter

We have designed social media data collection systems to collect hundreds of thousands of posts across Telegram, YouTube, Truth Social, Instagram, and others. Below, we provide screenshots to highlight this system. Click on the header for a demo (needs signup).


arbiter.simppl.org

Arbiter > Dashboard > 11

wikipedia.org

Feb 01, 2025 - Feb 28, 2025

General | Text Analysis

Welcome to your insights dashboard!
Explore engagement metrics, distribution patterns, and text analysis for your content.

Total Posts: 6,094

Total Engagement: 350,715

Active Users: 114,142

Average Comments: 3,049

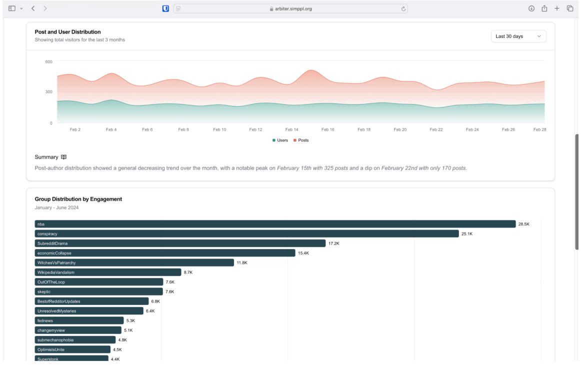
Engagement Distribution
Distribution of engagement over the time period

Likes: 350,715

Comments: 114,142

Summary

Likes and comments show a fluctuating pattern throughout the month, with *"a notable peak in engagement on 2025-02-13 with 35,777 likes and 5,715 comments"*. However, there are occasional dips like on 2025-02-22 with *"only 5,075 likes and 2,008 comments"*.


arbiter.simppl.org

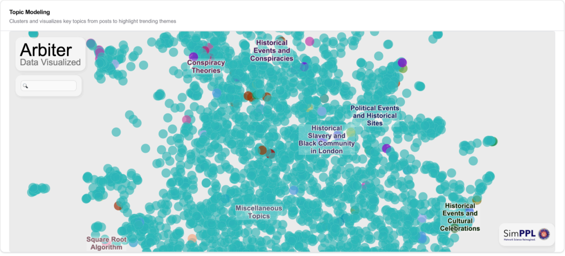
Post and User Distribution
Showing total visitors for the last 3 months

Last 30 days

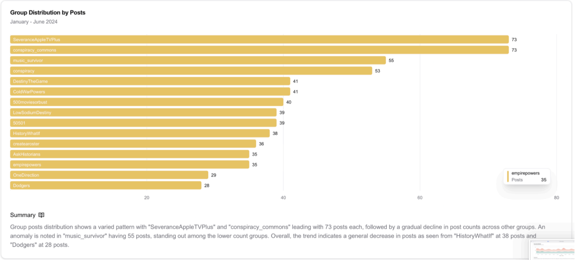
Summary

Post-author distribution showed a general decreasing trend over the month, with a notable peak on *February 15th with 325 posts* and a dip on *February 22nd with only 170 posts.*

Group Distribution by Engagement
January - June 2024


Topic Modeling
Clusters and visualizes key topics from posts to highlight trending themes

Arbiter
Data Visualized
Conspiracy Theories
Historical Events and Conspiracies
Political Events and Historical Sites
Historical Slavery and Black Community in London
Miscellaneous Topics
Historical Events and Cultural Celebrations
Square Root Algorithm
SimPPL

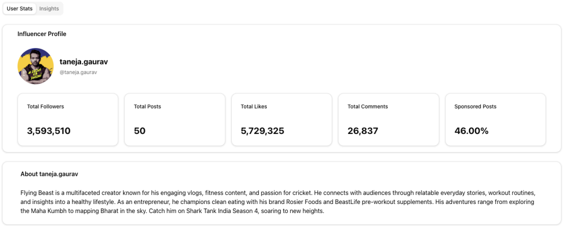

Group Distribution by Posts
January - June 2024

Summary

Group posts distribution shows a varied pattern with "SeveranceAppleTVPlus" and "conspiracy_commons" leading with 73 posts each, followed by a gradual decline in post counts across other groups. An anomaly is noted in "music_survivor" having 55 posts, standing out among the lower count groups. Overall, the trend indicates a general decrease in posts as seen from "HistoryWhatIf" at 38 posts and "Dodgers" at 28 posts.

B. InfluenceCheck

Influencers post multimodal content to a variety of different mainstream and alternative social platforms. Here, we demonstrate that we have already built significantly complex technical architecture for studying Indian influencers and their online claims, such that we will be able to deliver successfully on the proposed work. Click on the header for a demo (no signup needed).


User Stats | Insights

Influencer Profile

taneja.gaurav
@taneja.gaurav

Total Followers: 3,593,510

Total Posts: 50

Total Likes: 5,729,325

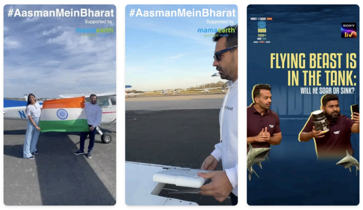
Total Comments: 26,837

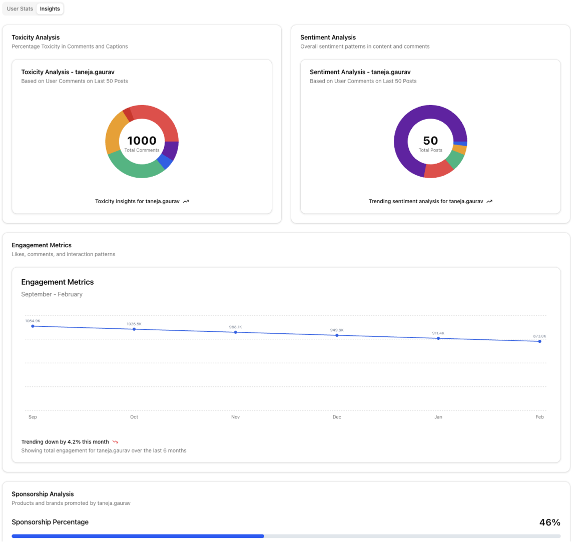
Sponsored Posts: 46.00%

About taneja.gaurav

Flying Beast is a multifaceted creator known for his engaging vlogs, fitness content, and passion for cricket. He connects with audiences through relatable everyday stories, workout routines, and insights into a healthy lifestyle. As an entrepreneur, he champions clean eating with his brand Rosier Foods and BeastLife pre-workout supplements. His adventures range from exploring the Maha Kumbh to mapping Bharat in the sky. Catch him on Shark Tank India Season 4, soaring to new heights.

Media for taneja.gaurav





Toxicity Analysis

Percentage Toxicity in Comments and Captions

Toxicity Analysis - taneja.gaurav

Based on User Comments on Last 50 Posts

Toxicity insights for taneja.gaurav

Sentiment Analysis

Overall sentiment patterns in content and comments

Sentiment Analysis - taneja.gaurav

Based on User Comments on Last 50 Posts

Trending sentiment analysis for taneja.gaurav

Engagement Metrics

Likes, comments, and interaction patterns

Engagement Metrics

September - February

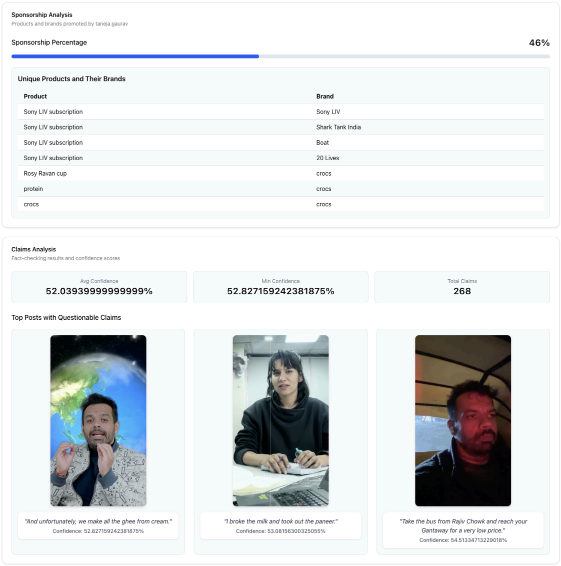
Trending down by 4.2% this month

Showing total engagement for taneja.gaurav over the last 6 months

Sponsorship Analysis

Products and brands promoted by taneja.gaurav

Sponsorship Percentage 46%



Sponsorship Analysis

Products and brands promoted by taneja.gaurav

Sponsorship Percentage 46%

Unique Products and Their Brands

Product	Brand
Sony LIV subscription	Sony LIV
Sony LIV subscription	Shark Tank India
Sony LIV subscription	Boat
Sony LIV subscription	20 Lives
Rosy Ravan cup	crocs
protein	crocs
crocs	crocs

Claims Analysis

Fact-checking results and confidence scores

Avg Confidence 52.039399999999999%

Min Confidence 52.827159242381875%

Total Claims 268

Top Posts with Questionable Claims

"And unfortunately, we make all the ghee from cream."
Confidence: 52.827159242381875%

"I broke the milk and took out the paneer."
Confidence: 53.08156300325055%

"Take the bus from Rajiv Chowk and reach your Gantaway for a very low price."
Confidence: 54.51334713229018%

Figures showcasing how we check for influencers' claims. Work is in small part supported by WikiCred 2021 grant of $10,000.

Folder with full-size photos: https://drive.google.com/drive/folders/1PNvnZfriCl9UOWW40dYP_IOnfoCEwVZn?usp=sharing