WIKIPEDIA AND AI: ETHICAL BOUNDARIES IN LLM TRAINING

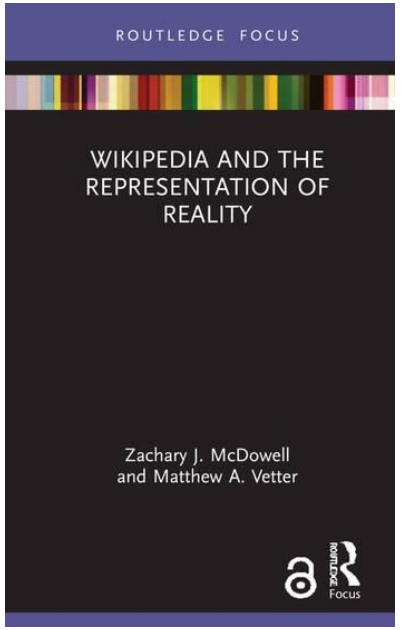
Dr. Matthew Vetter

Deoband Community Wikimedia

DCW Conversation Hour November 2024

ABOUT ME

- Professor of English, Indiana University of Pennsylvania
- Author, *Wikipedia and the Representation of Reality* (Routledge, 2021)
- WMF Regional Grants Committee serving U.S. and Canada since 2021
- Expert in Wikipedia-based education
- Co-chair CCCC Wikipedia Initiative / Wikiproject Writing
- Co-editor, Writing Spaces



RESEARCH PROJECT

1. How are LLMS trained on Wikipedia?
2. How does this training affect the sustainability of Wikipedia?

Research Purpose

The study aims to explore more how Large Language Models (LLMs) are trained on Wikipedia, and how their use in AI-powered chatbots, such as OpenAI's ChatGPT, Microsoft's Copilot, and Google's Gemini affects the sustainability of Wikipedia as a crowed-sourced project, while also raising concerns about information literacy and exploitative digital labor.

On meta: Research:Large Language Models (LLMs) Impact on Wikipedia's Sustainability

Research team

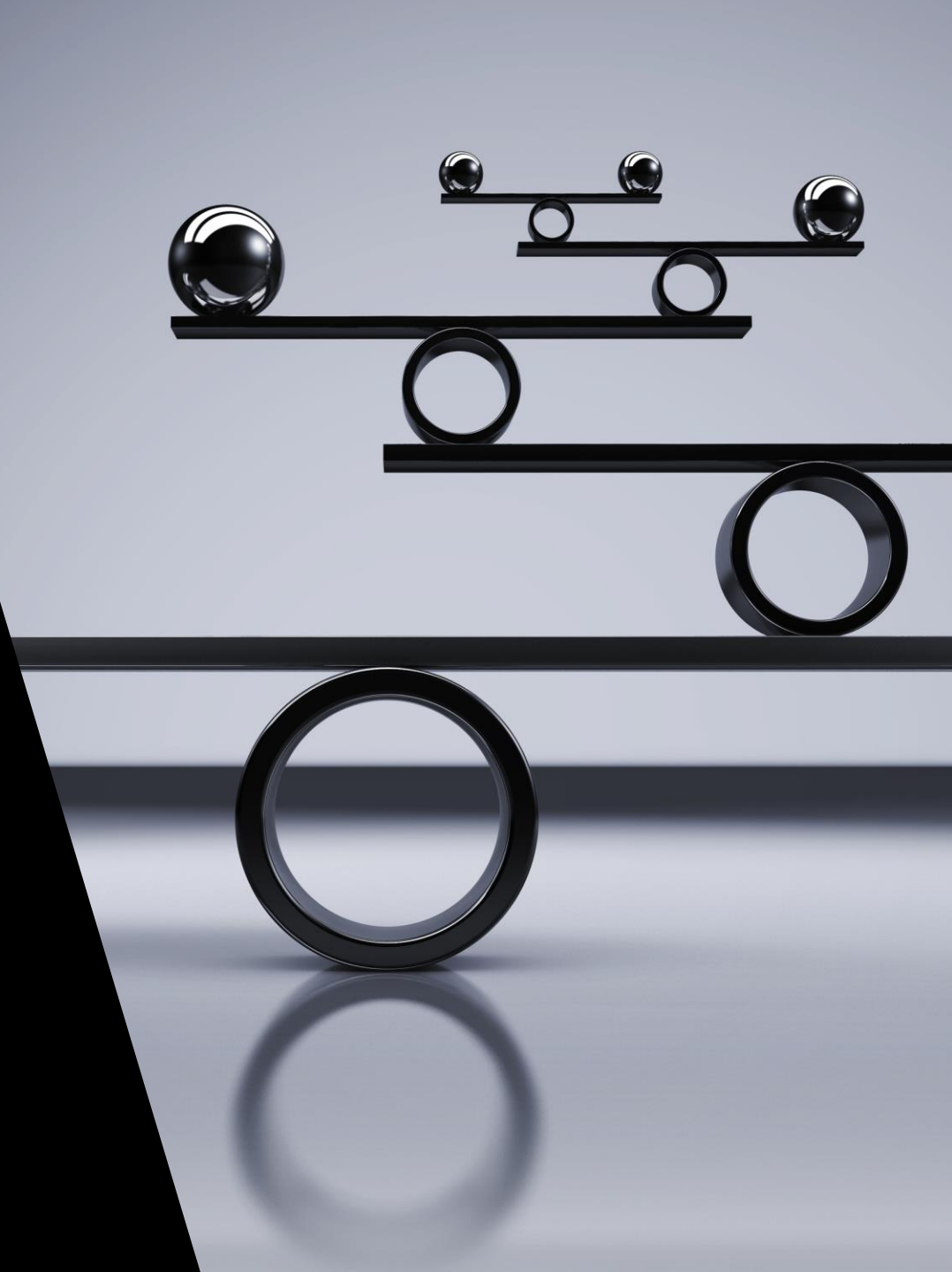
Zach McDowell, PhD – University of Illinois Chicago

Jialei Jiang, PhD – University of Pittsburgh

Article currently under review at AI & Society: "An Endangered Species: How LLMs Threaten Wikipedia's Sustainability"

TODAY'S TALK

- Generative AI & Large Language Models
- Wikipedia as training set
- Research methodology
- Key Findings
- Recommendations
- Questions & Discussion



AI & GEN AI

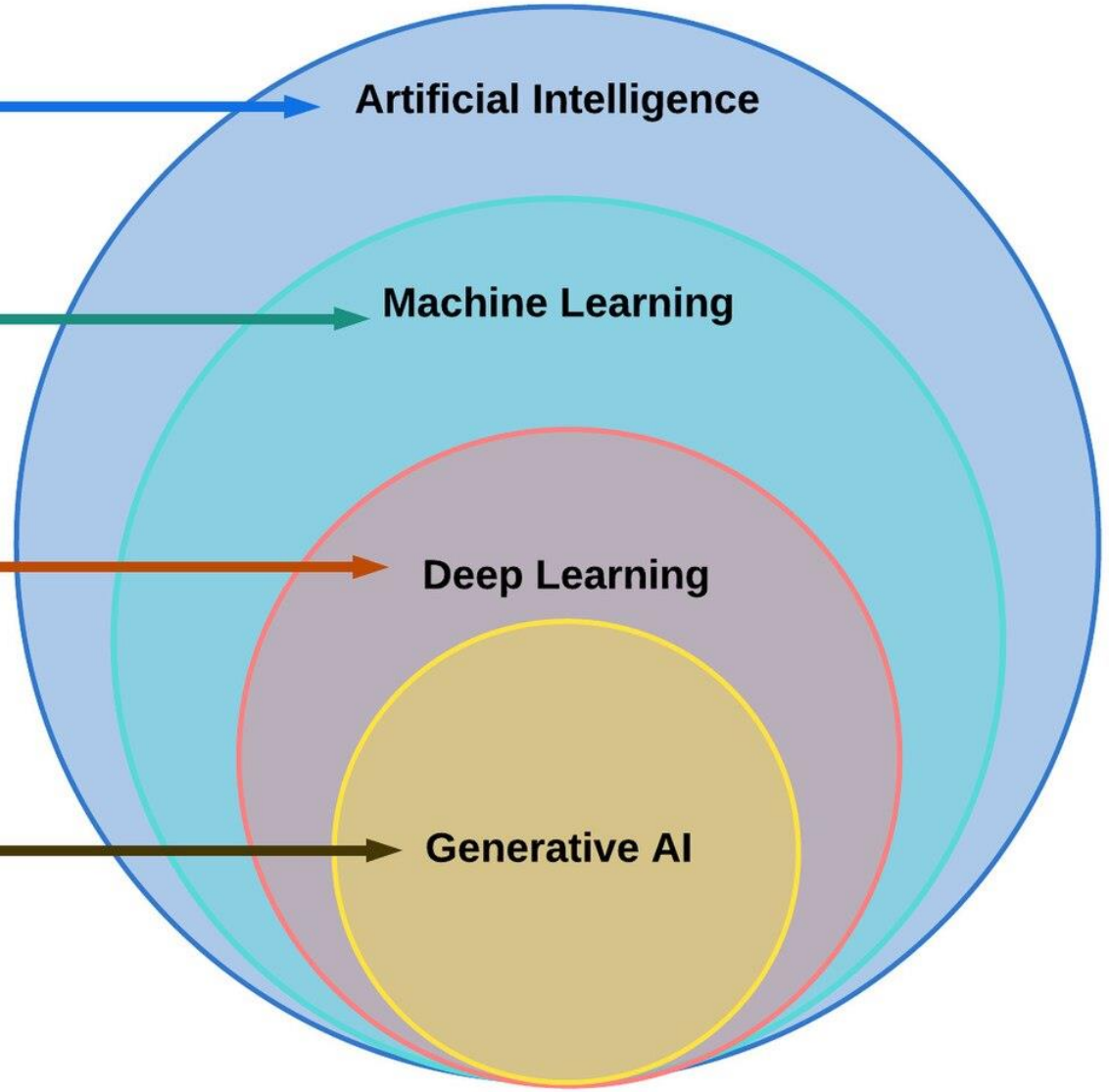
PopovaZhuhadar, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

Artificial Intelligence
AI involves techniques that equip computers to emulate human behavior, enabling them to learn, make decisions, recognize patterns, and solve complex problems in a manner akin to human intelligence.

Machine Learning
ML is a subset of AI, uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt. ML algorithms use supervised or unsupervised learning methods.

Deep Learning
DL is a subset of ML which uses neural networks for in-depth data processing and analytical tasks. DL leverages multiple layers of artificial neural networks to extract high-level features from raw input data, simulating the way human brains perceive and understand the world.

Generative AI
Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

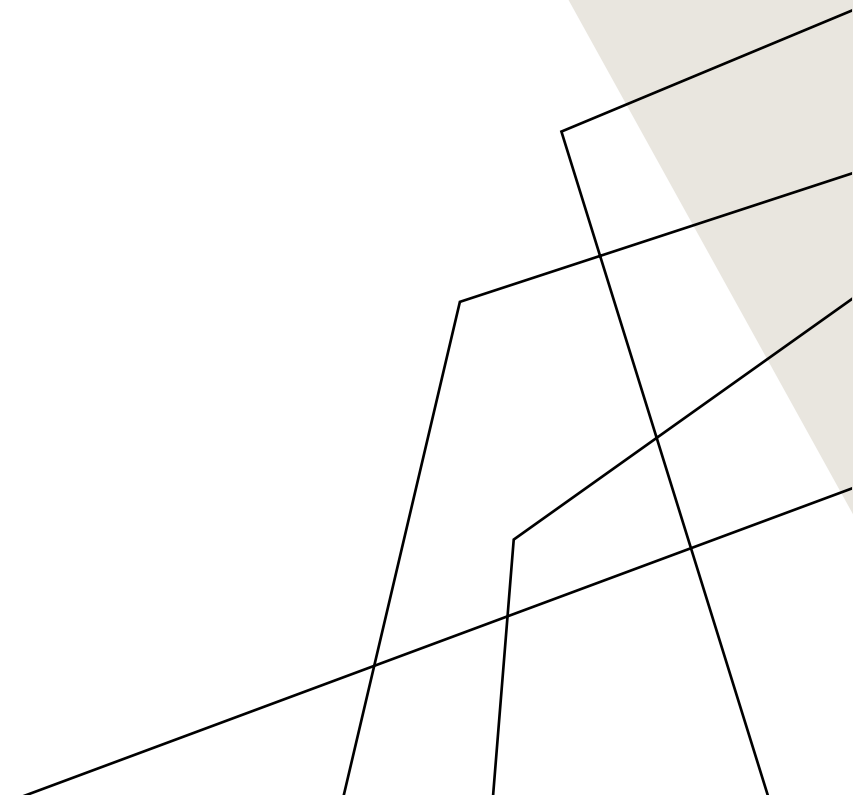


Unraveling AI Complexity - A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI.

(Created by Dr. Lily Popova Zhuhadar, 07, 29, 2023)

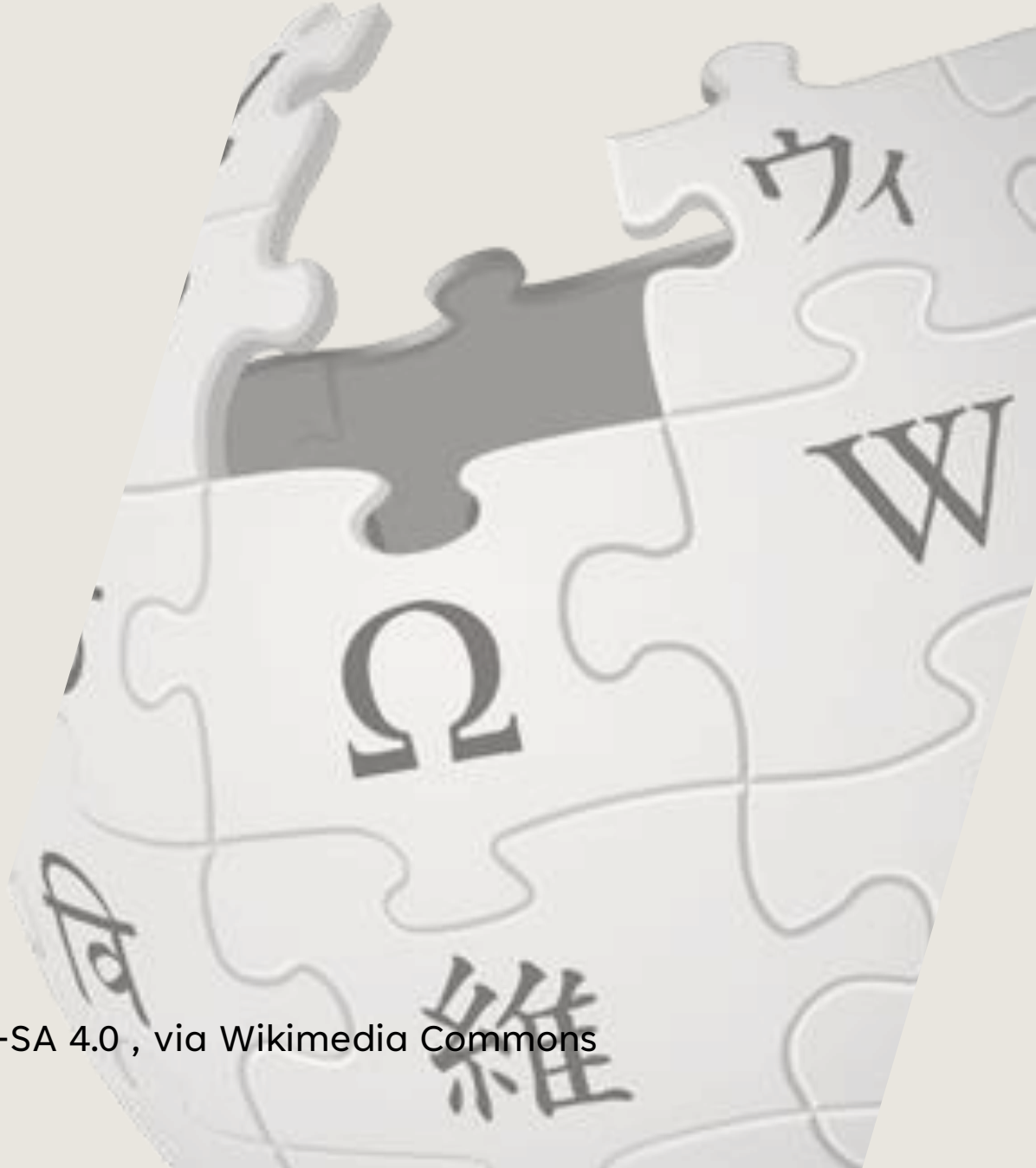
LARGE LANGUAGE MODELS

A large language model (LLM) is a type of computational model designed for natural language processing tasks such as language generation. As language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a self-supervised and semi-supervised training process. (Wikipedia)



WIKIPEDIA AS THE IDEAL TRAINING DATA FOR LARGE LANGUAGE MODELS

- WP:Verifiability
- Multilingual
- Size and scope
- Structured data
- Freely licensed
- Constantly updated



SharryX, CC BY-SA 4.0 , via Wikimedia Commons

RESEARCH METHODOLOGY

Problem-centered expert interview

- Expert = highly specialized knowledge and interest in the intersections between Wikipedia and LLMs

6 expert participants interviewed

- The interview instrument consisted of eight semi-structured questions, allowing for follow-up questions and side discussions

Qualitative thematic analysis

- Transcription
- De-identification
- Familiarization
- Coding and analysis
- Refinement of codes
- Selection of quotes

EXPERT INTERVIEWEES

1. E1 Denny Vrandečić (named) - Head of Special Projects at Wikimedia Foundation

2. E2 Stephen Harrison (named) - Journalist and tech lawyer; author of The Editors

3. E3 Aaron Halfaker (named) - Principal Applied Research Scientist at Microsoft

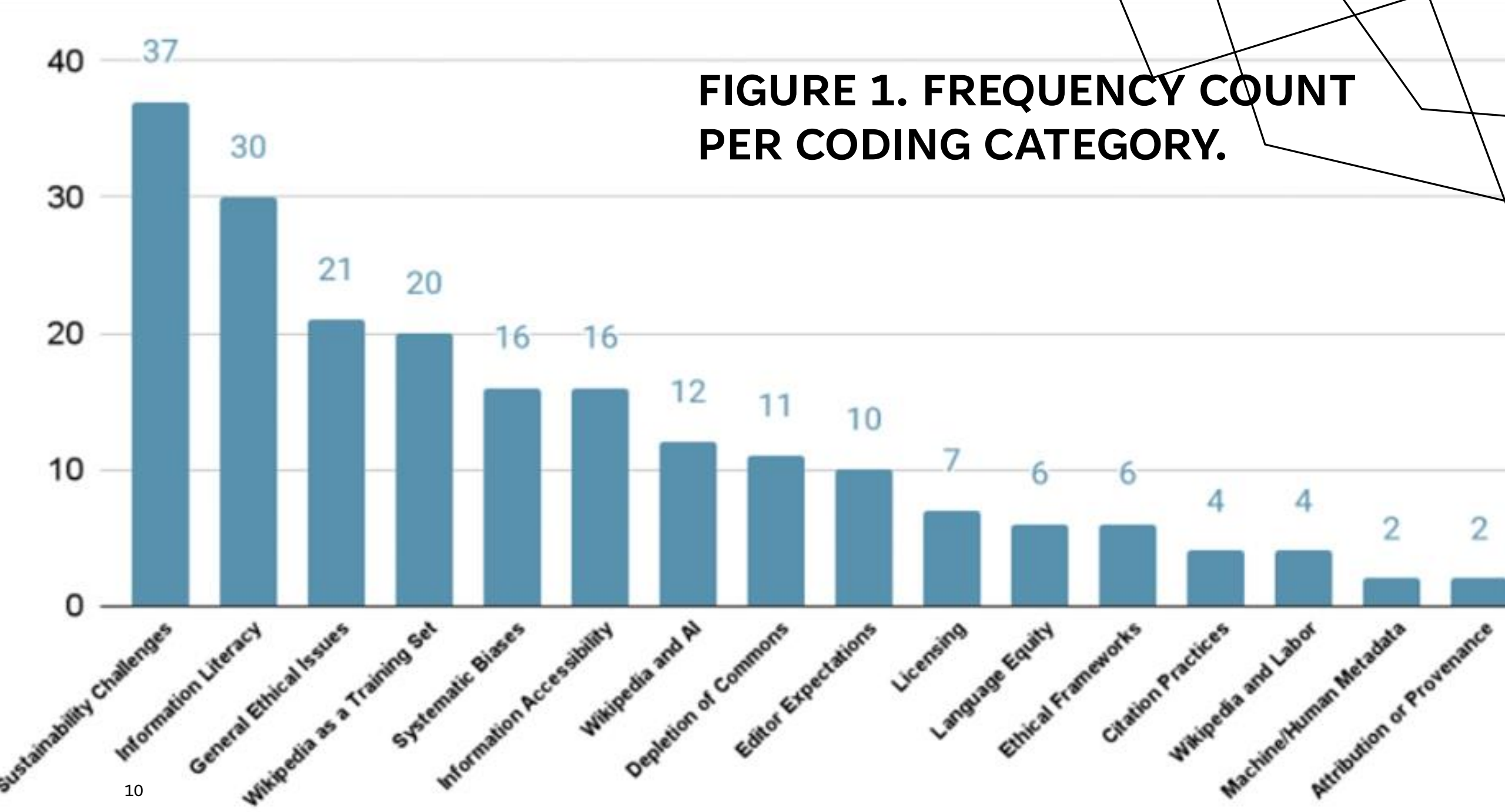
4. E4 Luca de Alfaro (named) - Professor of Computer Science and Engineering, University

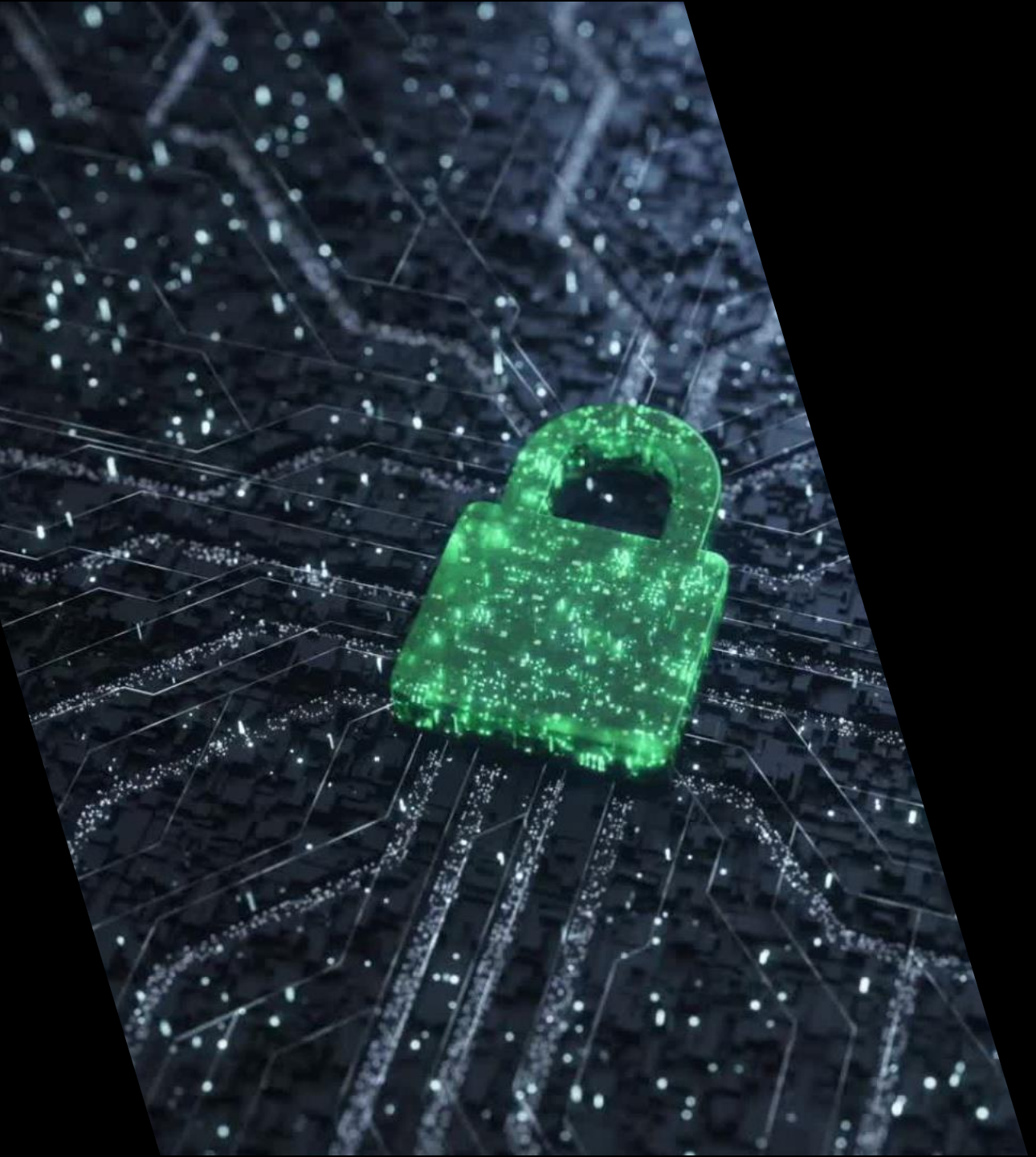
of California Santa Cruz

5. E5 Ava (pseudonym) - Researcher at Wikimedia Foundation

6. E6 Andrew (pseudonym) - Technical staff at Wikimedia Foundation

FIGURE 1. FREQUENCY COUNT PER CODING CATEGORY.





KEY FINDINGS

KEY FINDINGS

1. Wikipedia plays a significant role in the training of LLMs, but the exact process and value it is given is unclear.
2. LLMs act as intermediaries between users and original knowledge sources, often reducing information quality and perpetuating biases, while lacking transparency and proper citation.
3. Wikipedia's sustainability is threatened by LLMs' negative impact on the digital commons, Wikipedia discoverability, community engagement, and disintermediation.
4. The use of Wikipedia as LLM training data involves ethical problems related to contributor expectations, the risk of depleting the commons, and exacerbation of linguistic and cultural inequities.
5. Ethical concerns may be partially addressed via systemic changes to market incentives and license models, financial contributions to Wikipedia from big tech, and technical solutions related to data provenance and attribution.
6. Systemic biases in LLMs, which can be inherited from sources like Wikipedia, are inevitable, but can be mitigated via proactive efforts to diversify communities and content in the digital commons.

DIS/INTERMEDIATION

“LLM applications bring even stronger (dis-)intermediation than the Google Knowledge Panel because they are heavily customized to the question being asked.” (E1).

Wikipedia is bypassed

Information quality itself suffers, whether that be in terms of simplification via a shortened summary or more problematic inaccuracy (misinformation).

Increasing distance between original information and end-user consumption

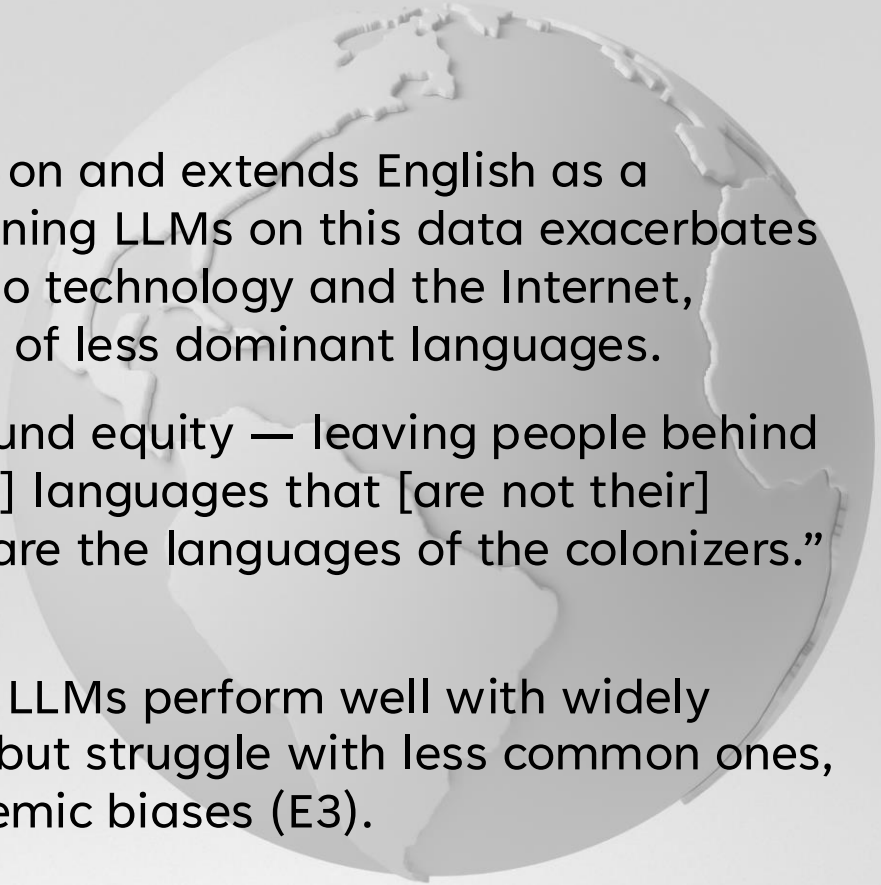
SUSTAINABILITY

- The risk of a shrinking open (commons) environment could isolate Wikipedia and hinder its collaborative nature (E5).
- Users may rely on LLMs for quick consultations, bypassing Wikipedia and reducing opportunities for content improvement and community engagement (E5).
- All interviewees warn that LLMs could diminish the discoverability of Wikipedia, leading to decreased donations and editorial contributions (E1-E6).

CONTRIBUTOR ALIENATION

- Wikipedia contributors never intended for their content to be used by machine learning models (E2, E4).
- “The ethical problem that I hear about most frequently from Wikipedians is that the situation doesn’t seem fundamentally fair. The editors produce this content without compensation, it is openly licensed, and then these big tech companies make so much money from LLMs.” (E2).
-

SYSTEMIC BIAS



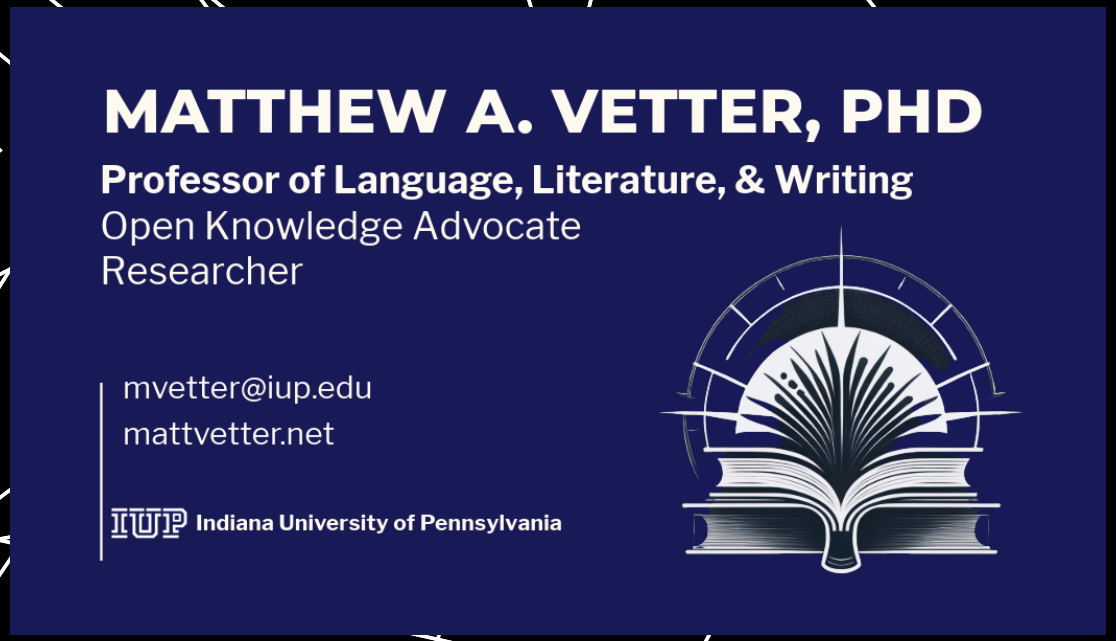
- Wikipedia already relies on and extends English as a dominant language, training LLMs on this data exacerbates existing gaps in access to technology and the Internet, particularly for speakers of less dominant languages.
- "[T]here's a concern around equity — leaving people behind or forcing people to [use] languages that [are not their] native languages. They are the languages of the colonizers." (E5).
- To make matters worse, LLMs perform well with widely documented languages but struggle with less common ones, further entrenching systemic biases (E3).

RECOMMENDATIONS

- Advocate for transparency and attribution in LLM applications.
- Demand reciprocity from big tech in the form of alternate licensing models and philanthropy back to WMF.
- Support initiatives to diversify the Wikipedia editor community, encouraging contributions from underrepresented groups and languages.
- Encourage and continue "Future Audiences" experiments in order to maintain Wikipedia’s feedback loop, where readers become contributors, and caution against tools that replace rather than support Wikipedians.
- Encourage the development of LLMs that prioritize explainability and transparency, providing users with insights into how they retrieve and process information.
- Promote educational initiatives that equip users with the critical thinking skills necessary to evaluate AI-generated content and understand its limitations, as well as to understand Wikipedia's role in the evolving information landscape.
- Promote Wikipedia's unique attributes like its collaborative editing, neutrality, and trusted information to ensure its continued relevance and discoverability in the age of LLMs12.
- Explore AI-assisted tools to enhance human editors' contributions without replacing them3. Focus on tools that support information verification, source identification, and community engagement.

Patryk Kajdrowicz for 24klatkifilmy, commissioned by the Wikimedia Foundation, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

KNOWLEDGE IS HUMAN



THANK YOU

User:Matthewvetter
mvetter@iup.edu

DISCUSSION & QUESTIONS