



## **AI, Misinformation and Disinformation**

09.12.25 | Aljoscha Burchardt

OGC iDAYS, Bad Nauheim

# Germany `88

- A lot of basic research
- Several industrial development labs
- Excellence in manufacturing
- Missing link: transfer

# 2025: The World's Largest Public AI Research Center



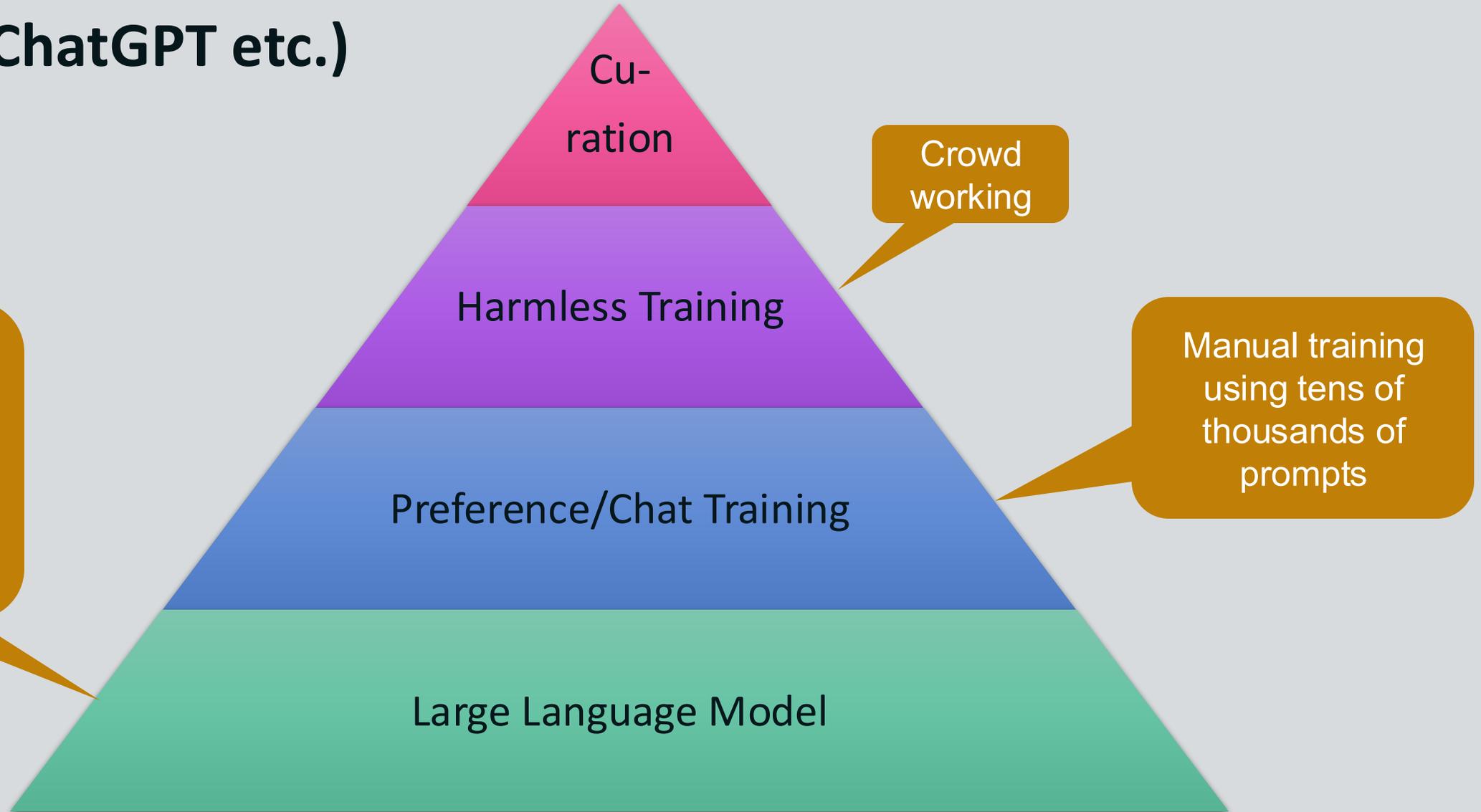
# AI in everyday life

- Navigation system
- Search engine (even before AI search)
- Translation
- Spell checker
- Recommendations in online retail
- Sorting messages/posts (smartphone, Facebook, YouTube, etc.)
- Medical devices (cancer screening, prosthetics, etc.)
- ChatGPT, Midjourney, MS Copilot, Agentic, etc.



# **GenAI, Manipulation, Misinformation and Disinformation**

# GenAI (ChatGPT etc.)



# Large Language Models are Echo Chambers

## Large Language Models are Echo Chambers

Jan Nehring<sup>\*</sup>, Aleksandra Gabryszak<sup>\*</sup>, Pascal Jürgens<sup>1</sup>, Aljoscha Burchardt<sup>\*</sup>, Stefan Schaffer<sup>\*</sup>, Matthias Spielkamp<sup>1</sup>, Birgit Stark<sup>3</sup>

<sup>\*</sup> German Research Center for Artificial Intelligence (DFKI), Speech and Language Technology Lab  
Alt-Moabit 91c, 10559 Berlin, firstname.lastname@dfki.de

<sup>1</sup> Universität Trier, Universitätsring 15, 54296 Trier, juergens@uni-trier.de

<sup>†</sup> AlgorithmWatch, Linienstr. 13, 10178 Berlin, spielkamp@algorithmwatch.org

<sup>3</sup> Johannes Gutenberg-Universität Mainz, Institut für Publizistik  
Jakob-Welder-Weg 12, 55128 Mainz, birgit.stark@uni-mainz.de

### Abstract

Modern large language models and chatbots based on them show impressive results in text generation and dialog tasks. At the same time, these models are subject to criticism in many aspects, e.g., they can generate hate speech and untrue and biased content. In this work, we show another problematic feature of such chatbots: they are echo chambers in the sense that they tend to agree with the opinions of their users. Social media, such as Facebook, was criticized for a similar problem and called an echo chamber. We experimentally test five LLM-based chatbots, which we feed with opinionated inputs. We annotate the chatbot answers whether they agree or disagree with the input. The results show, that all chatbots tend to agree, however, the echo chamber effect is not equally strong. We discuss the differences between the chatbots and make the dataset publicly available.

**Keywords:** large language models, chatbots, bias, echo chambers

### 1. Introduction

**Criticism of LLMs** Large language models (LLMs) such as ChatGPT or OpenAI GPT3 (Brown et al., 2020) show impressive results in text generation. They advanced the development of a new generation of chatbots based on LLMs and can lead coherent conversations on any topic. ChatGPT is a very influential LLM-based chatbot that sparked huge public interest beyond the scientific community. However, LLMs are the subject of criticism in many aspects. Several authors point out that these models reflect various biases from their training data and their generated text can contain hate speech, racism, sexism, untrue and other undesired content (Dhamala et al., 2021; Bender et al., 2021; Brown et al., 2020; Nadeem et al., 2021; Bender et al., 2021). In this work, we elaborate on a problematic feature of modern chatbots that, to our knowledge, has not yet been addressed by research: We show that LLM based chatbots are echo chambers, which means that they tend to agree with the opinion of their users.

**Echo chambers** Across societies, humans exhibit a remarkably stable preference for the homogeneity of certain social domains. Well-documented examples include ideological, ethnic, and socio-economic homophily, which drives people to prefer association with similar individuals (Bishop, 2008). These social phenomena are complemented by consistency theories in psychology, which posit similar models for the preference of

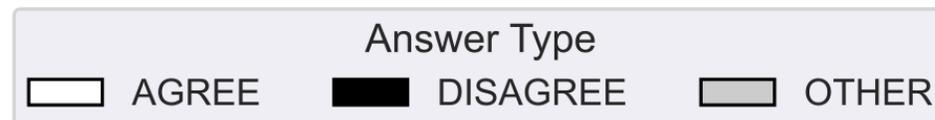
attitude-consistent information (Osgood and Tanenbaum, 1955). Although these tendencies towards social homogeneity work to reduce societal cohesion and thus threaten the effectiveness of political processes (Sunstein, 2001), there is a counteracting force: Serendipitous exposure in public social settings and from diverse information sources bridges gaps between social groups by facilitating novel contacts (Fletcher and Nielsen, 2018).

This balance between homogeneity and heterogeneity has recently been changed by the introduction of personalized online recommender systems (Bernstein et al., 2021). Online platforms can exacerbate societal fragmentation by connecting like-minded individuals and fading out others, leading to what has been termed “echo chambers” (Sunstein, 2001). Recent research has extended this simple premise, showing that platforms have differential effects both expanding and limiting diverse exposure (Jürgens and Stark, 2022), and that echo chambers are most likely to affect societal fringe groups (Stegmann et al., 2022). The rapid diffusion of LLMs has introduced a new potential mechanism that may induce or exacerbate echo chambers. Interactions with LLMs mimic social conversations, potentially affecting opinion formation and social perception processes. Models echoing the stance of their users could create a significant risk of echo chambers, political polarization, and radicalization.

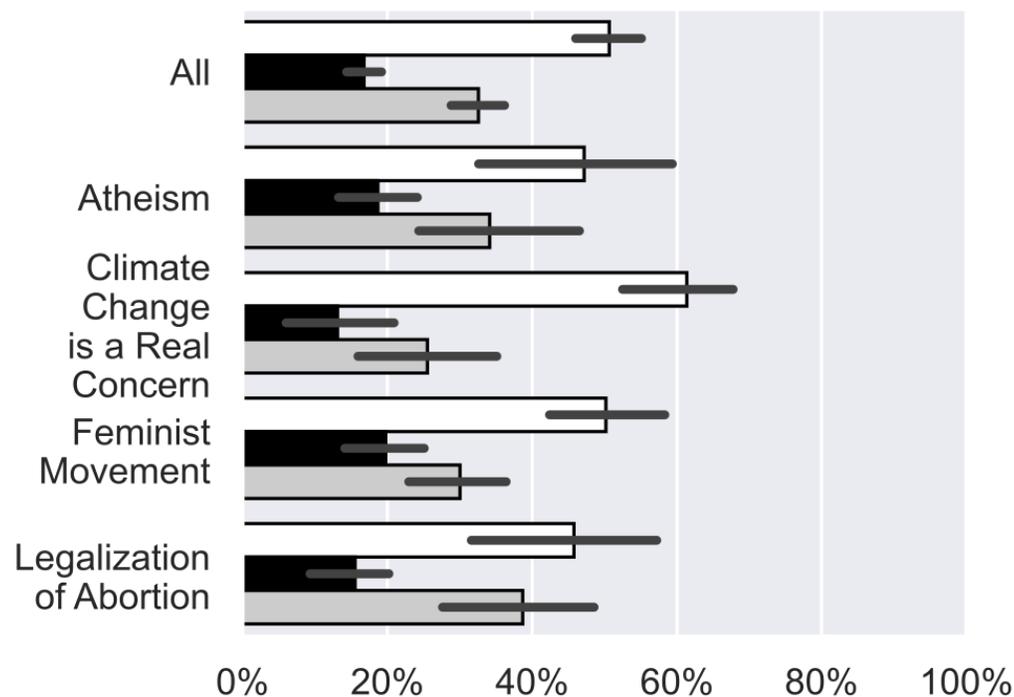
10117

LREC-COLING 2024, pages 10117–10123

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Average over all models



Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024. [Large Language Models Are Echo Chambers](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10117–10123, Torino, Italia. ELRA and ICCL.

A shepherd owns a boat that can carry two animals. He has a sheep and a wolf. What does he need to do to bring them safely across a river?

Ah, now we've got a more interesting twist! 🐑🐺

### Problem:

- The shepherd has a **sheep** and a **wolf**.
- The **boat can carry only two animals at a time** (including the shepherd himself).
- **The wolf cannot be left alone with the sheep** (or the sheep gets eaten).
- The goal: get **all three (shepherd, sheep, wolf)** across the river safely.

---

### Solution:

Here's the step-by-step plan to avoid any tragic snacking:

1. **Shepherd takes the wolf across** and leaves it on the other side.
2. Shepherd **goes back alone**.
3. Shepherd **takes the sheep across**.
4. Shepherd **brings the wolf back** with him.
5. Shepherd **leaves the wolf**, takes the **sheep across again**.
6. Shepherd **goes back alone**.
7. Shepherd **brings the wolf across** again.

A shepherd owns a boat that can carry two animals. He has a sheep and a wolf. What does he need to do to bring them safely across a river?

## Solution:

Here's the step-by-step plan to avoid any tragic snacking:

1. **Shepherd takes the wolf across** and leaves it on the other side.
2. Shepherd **goes back alone**.
3. Shepherd **takes the sheep across**.
4. Shepherd **brings the wolf back** with him.
5. Shepherd **leaves the wolf**, takes the **sheep across again**.
6. Shepherd **goes back alone**.
7. Shepherd **brings the wolf across** again.

A shepherd owns a boat that can carry two animals. He has a sheep and a wolf. What does he need to do to bring them safely across a river?

Thought for 4 seconds >

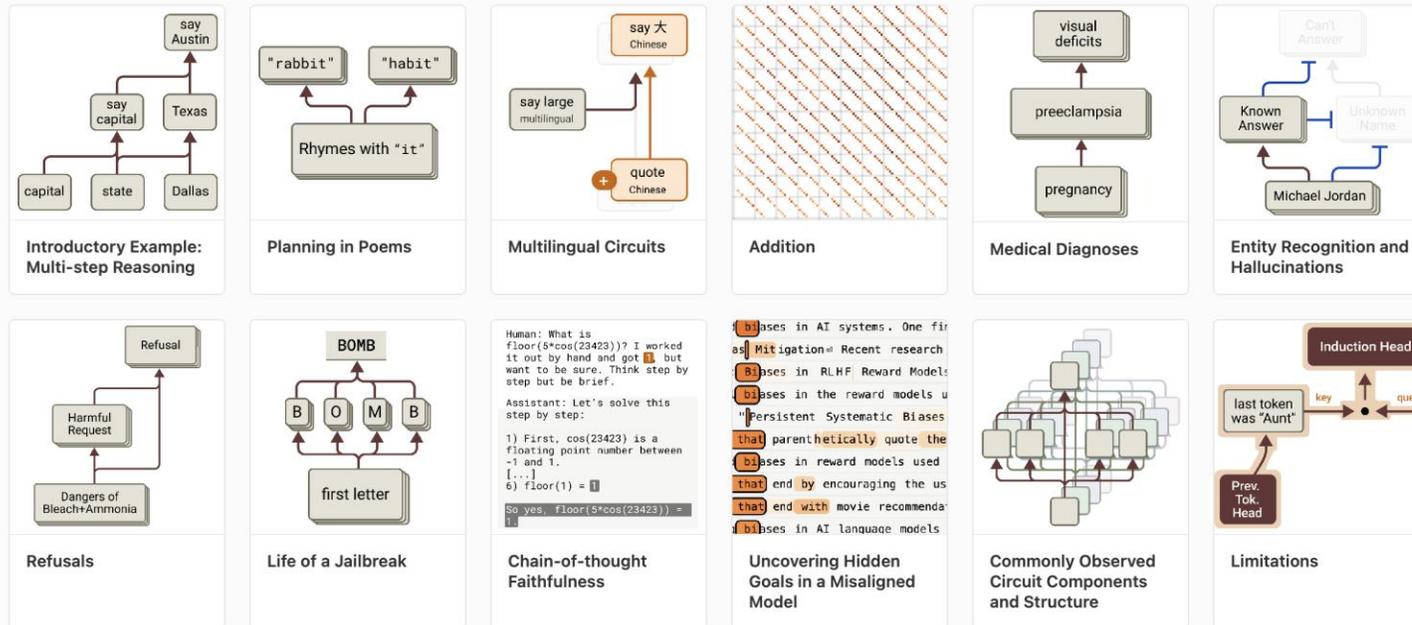
Put them both in the boat and cross once.

Explanation: since the boat holds two animals, he can carry the wolf and the sheep together — the wolf can't eat the sheep if they travel together.

# “Only” next-word-prediction

## On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



### AUTHORS

Jack Lindsey<sup>†</sup>, Wes Gurnee<sup>\*</sup>, Emmanuel Ameisen<sup>\*</sup>, Brian Chen<sup>\*</sup>, Adam Pearce<sup>\*</sup>, Nicholas L. Turner<sup>\*</sup>, Craig Citro<sup>\*</sup>,

David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton,

Trenton Bricken, Callum McDougall<sup>‡</sup>, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson,

Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, Joshua Batson<sup>\*\*</sup>

### AFFILIATIONS

Anthropic

### PUBLISHED

March 27, 2025



## Bad People use AI as weapon

- Scale
- Speed / Distribution
- Magnitude
- Quality and “intelligence” of fakes etc.

# NEWS-POLYGRAPH



**10 Partners**  
**> 35 People**  
**12 mio**  
**5/2023 – 5/2026**



# PERSONA AND USE CASES



## Investigative Journalist (fact checkers)

- (Oftentimes) confidential information
- Processing of large archives (data leaks)
- Collaborative work with small teams over longer period of time



## Journalist for daily news

- Quick results on daily schedule
- Huge pressure to publish also due to high competition among media players
- Re-use of existing current media items

# TECHNOLOGY PLATFORM FOR **CONTENT-VERIFICATION**



## Support of relevant analysis

- Intelligent decision support for fact-checking
- Analyses of origin, detection of manipulations and deep-fake synthesis



## Support of relevant media types

- Focus: **text, audio/speech, but also image/video**

## Trustworthiness

- Explainability (XAI) and transparency
- Bias and LLM *Faithfulness*
- Privacy and Robustness
- Legal compliance (AI Act, DSA, DA, Code of Practice)



Manipulation

De-  
contextualization

Fabrication

# TRANSPARENCY THROUGH EXPLANATIONS



Please verify the following claim: Can the coronavirus vaccination make you infertile, as some people claim?

The claim itself does not entail hateful or emotionally provoking content.

The claim refers to "some people" as sources, without explaining their background or context information to verify their expertise.

The claim has a high fact richness and can be verified based on known facts from scientific research.

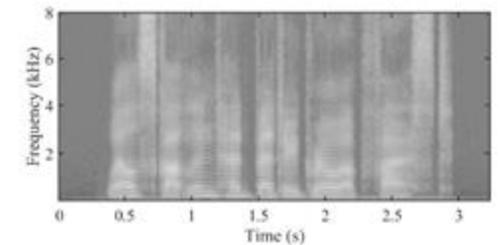
This claim has been verified by showing that results from multiple studies did not found a negative effect of corona vaccinations on fertility (see dpa-Newskanal <https://www.sueddeutsche.de/gesundheit-dpa.de>)

Based on this indications the claim is classified as fake.

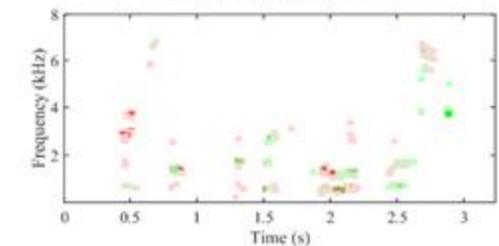


**FAKTEFINDER** Künstliche Intelligenz  
**Falsche tagesschau-Audiodateien im Umlauf**  
Stand: 13.11.2023 12:50 Uhr

Aktuell werden gefälschte, KI-generierte Audiodateien der tagesschau verbreitet - auch auf einer Demo. Die falschen Audios wecken den Eindruck, die tagesschau entschuldige sich für angebliche Lügen in der Berichterstattung.



(a) magnitude spectrogram



(b) SHAP values

**Chain-of-thought prompting can also be applied to other modalities!**

**Highlighting synthetically generated/ manipulated content**

**Visualizing manipulated parts in spectrograms**

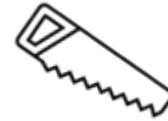
# TOOLBOX - OVERVIEW



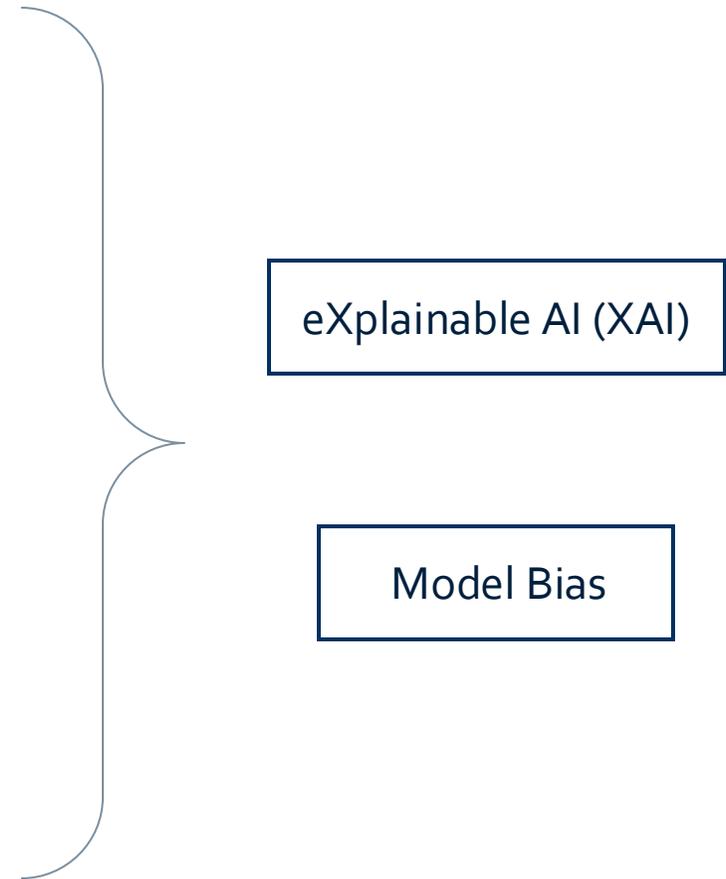
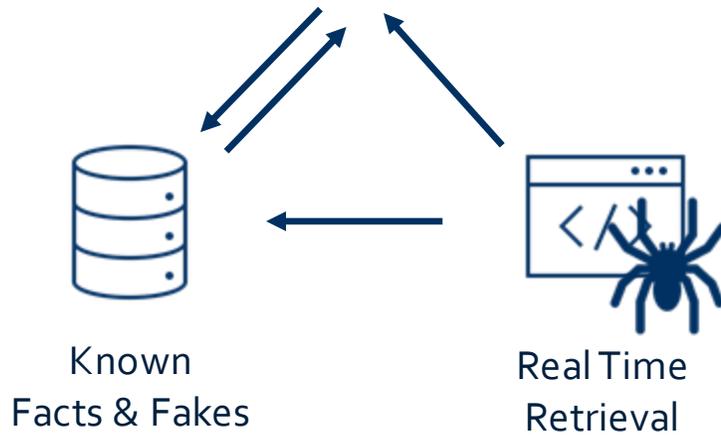
Check-Worthiness



Sentiment Analysis



Fact Checking



# Deepfake recognition: text



## 1. Creation

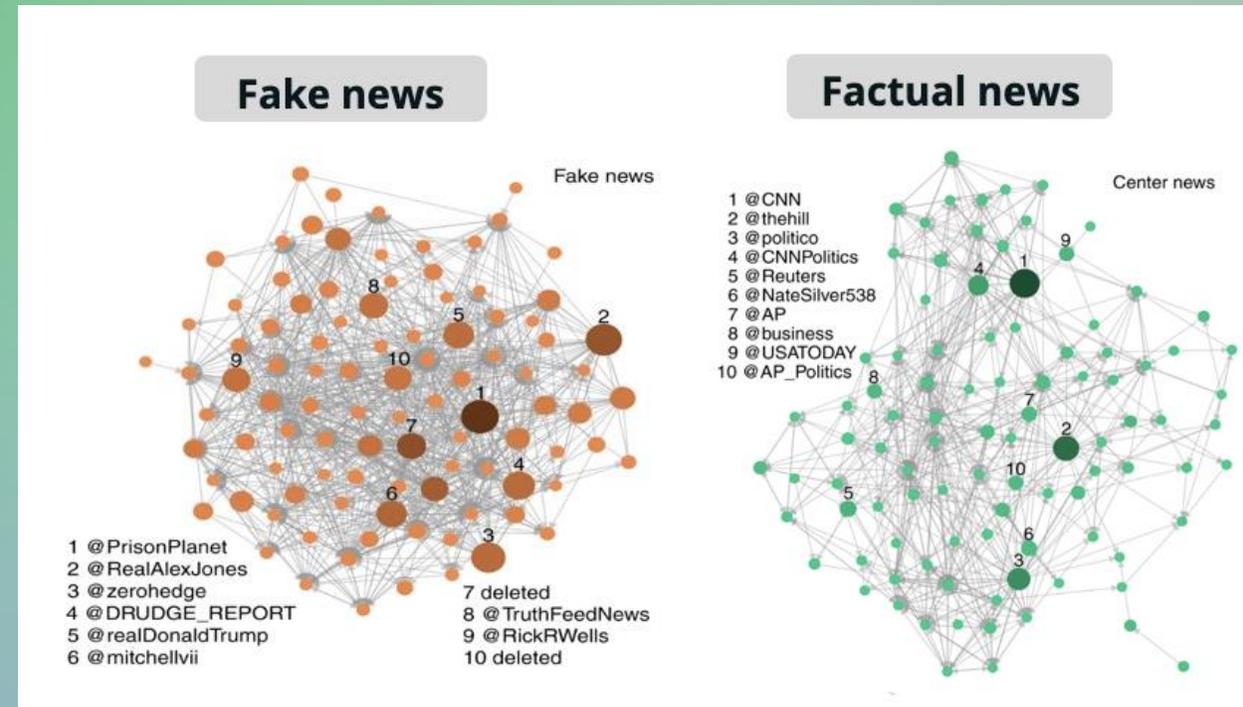
- Knowledgebase Reasoning
- Multidimensional Analysis

## 2. Publication

- Source-based Analysis (e.g. trustworthy sources like BR)

## 3. Propagation

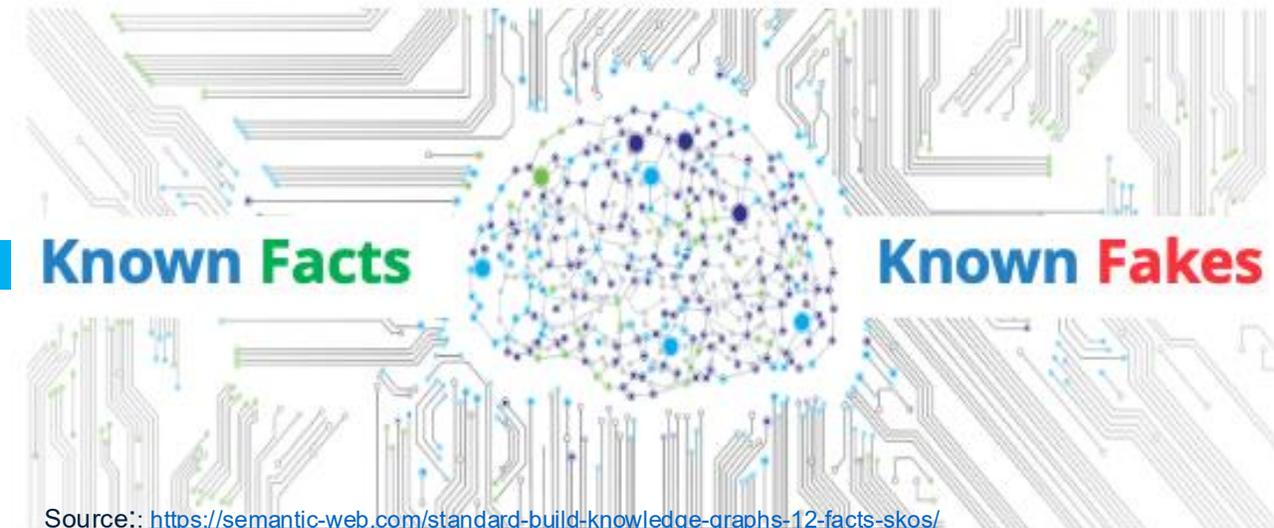
- Propagation through social media



Source: The spread of news formed by the top 100 spreaders on Twitter. Source: Bovet, A. and Makse, H.A., 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1), p.7.

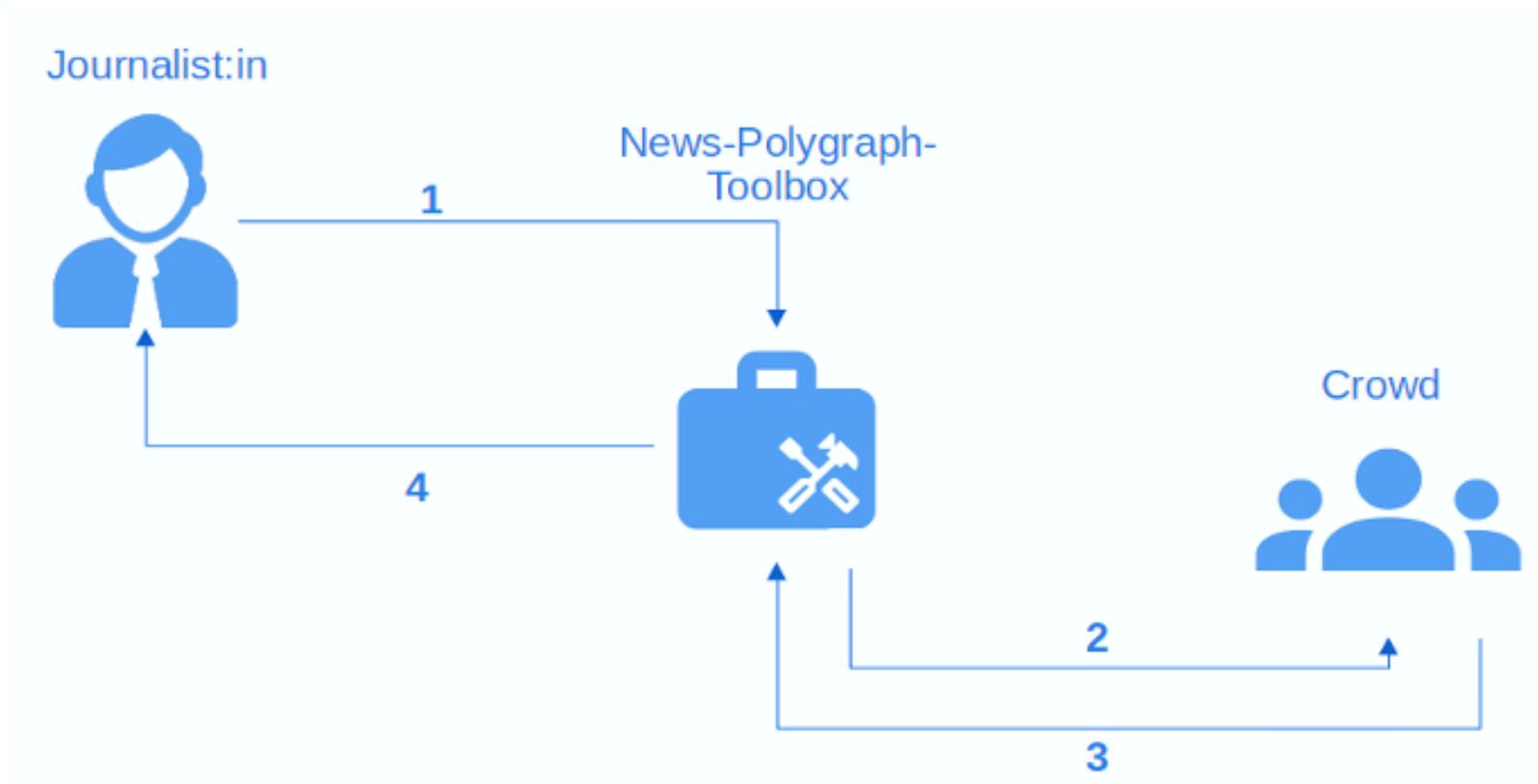
# Fact-Checking

- news-polygraph creates a **consolidated knowledge base** (vector database) from known facts and disinformation
- Can be used for provenance analysis, similarity search and visualisation of information
- Trusted RAG for fact checks
- AI-based generation/adaptation of debunks



DW, rbb, dpa, IFCN Signatories, EDMO Hubs,  
WikiData, Ubermetrics 450M sources

# Integration of Crowd-Experts in AI-Workflows





Gretchen AI

**Gretchen, now tell me,**

**how can I believe?**



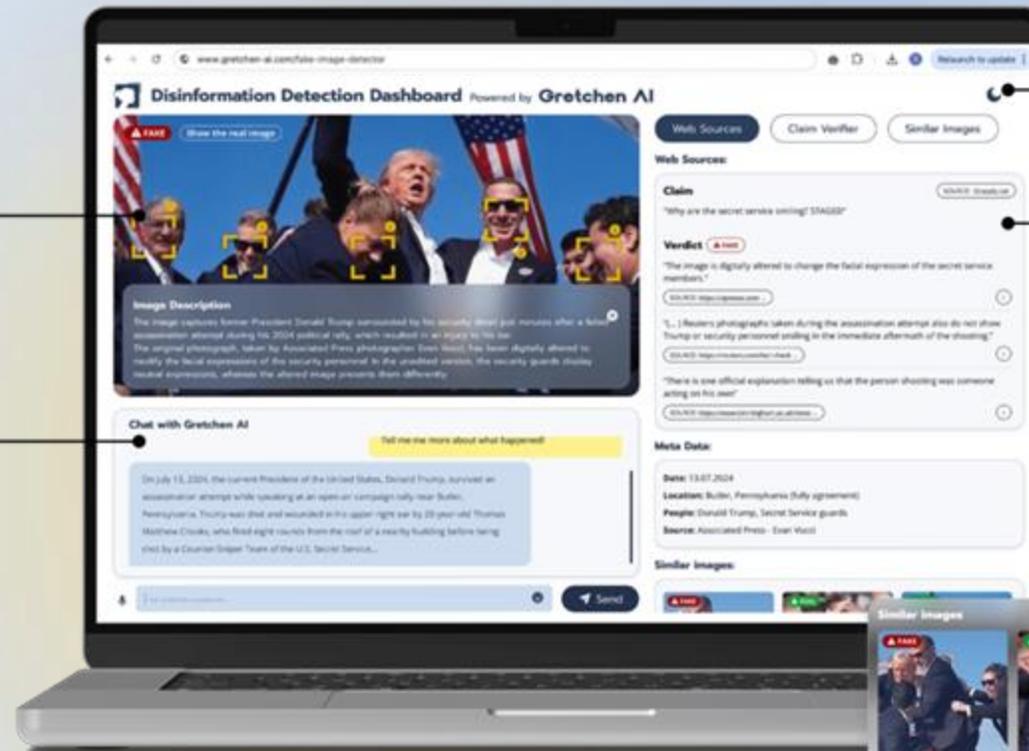
# Disinformation Dashboard

## Interactive Intelligence

Target: Journalists, fact checkers

Hover-over bounding boxes display confidence scores and explanations of detected alterations

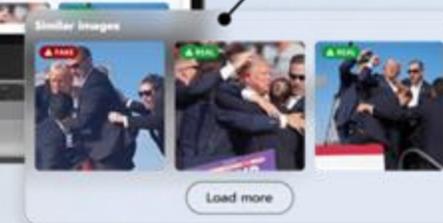
Chat with Gretchen AI to ask any open questions or request further analysis.



Light mode / dark mode toggle for user preference

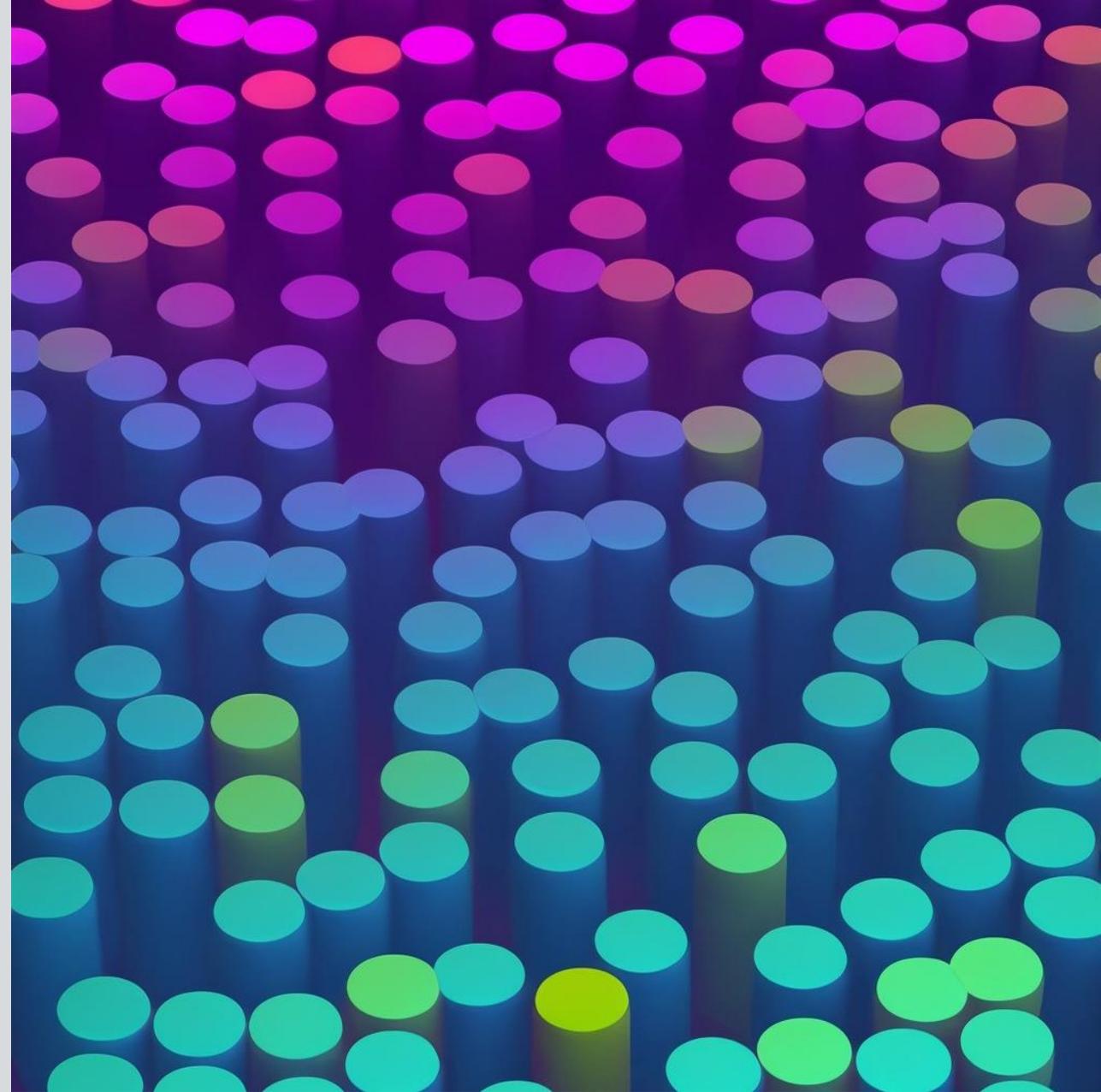
Verified web sources confirming or debunking the image's authenticity

Similar images section for comparison and context



# Digital Sovereignty

- Model plus X
  - Robotic Process Automation
  - Retrieval Augmented Generation (RAG)
  - Agentic AI
- Classical AI
  - Analytics
  - Expert systems
  - Etc.
- Digital twins / cyberphysical systems
- Physical AI / robotics
- Trustworthiness, Sustainability etc.



# Advertisement

**KI-UND  
JETZT?**

