Critical AI Literacy in Practice

Lessons from Current DH Projects

Moritz Mähr

moritz.maehr@gmail.com

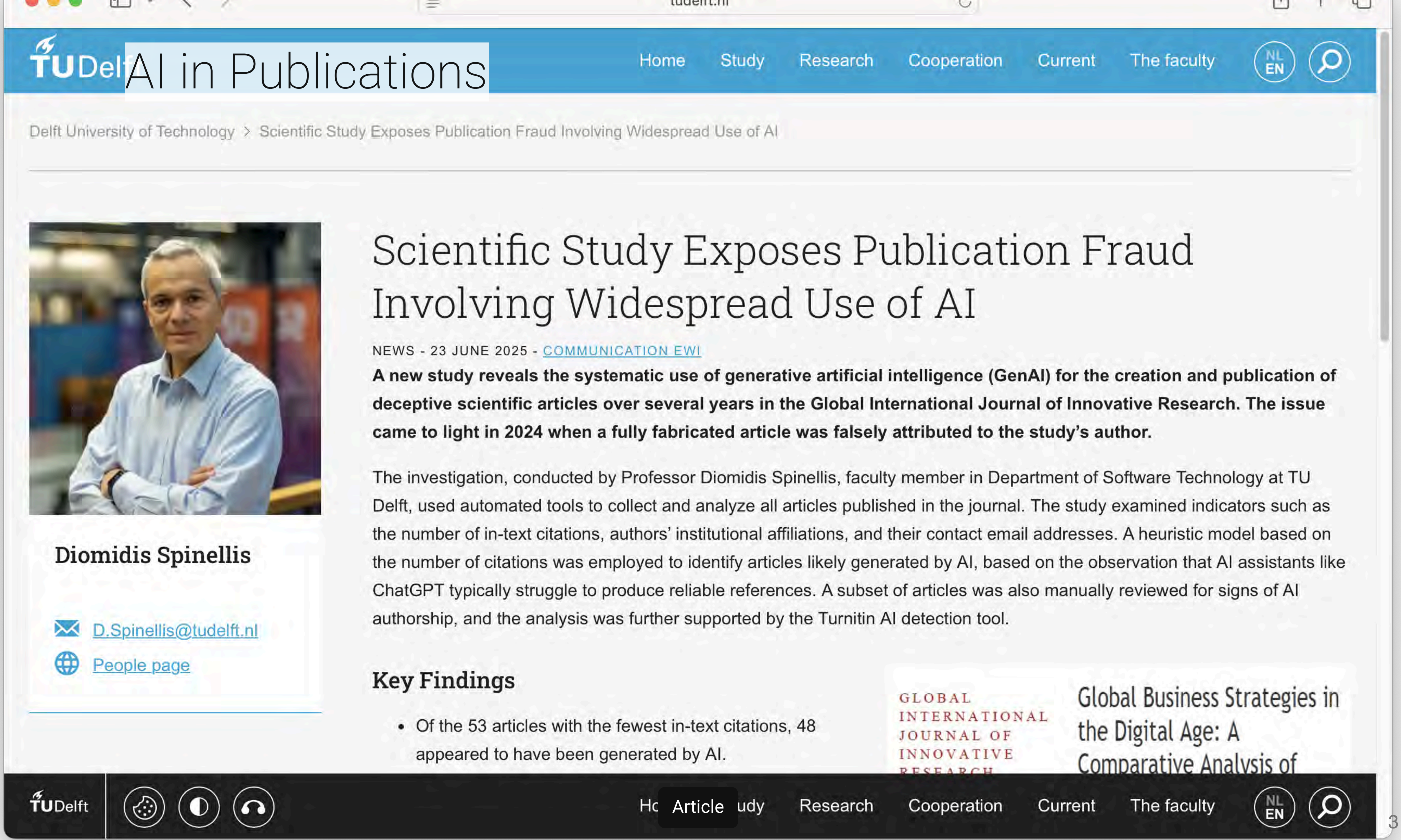
University of Basel

University of Bern

September 9, 2025

AI is Everywhere, also in Science

AI in Publications



Delft University of Technology > Scientific Study Exposes Publication Fraud Involving Widespread Use of AI

Scientific Study Exposes Publication Fraud Involving Widespread Use of AI

NEWS - 23 JUNE 2025 - COMMUNICATION EWI

A new study reveals the systematic use of generative artificial intelligence (GenAI) for the creation and publication of deceptive scientific articles over several years in the Global International Journal of Innovative Research. The issue came to light in 2024 when a fully fabricated article was falsely attributed to the study's author.

The investigation, conducted by Professor Diomidis Spinellis, faculty member in Department of Software Technology at TU Delft, used automated tools to collect and analyze all articles published in the journal. The study examined indicators such as the number of in-text citations, authors' institutional affiliations, and their contact email addresses. A heuristic model based on the number of citations was employed to identify articles likely generated by AI, based on the observation that AI assistants like ChatGPT typically struggle to produce reliable references. A subset of articles was also manually reviewed for signs of AI authorship, and the analysis was further supported by the Turnitin AI detection tool.

Diomidis Spinellis

✉ D.Spinellis@tudelft.nl

People page

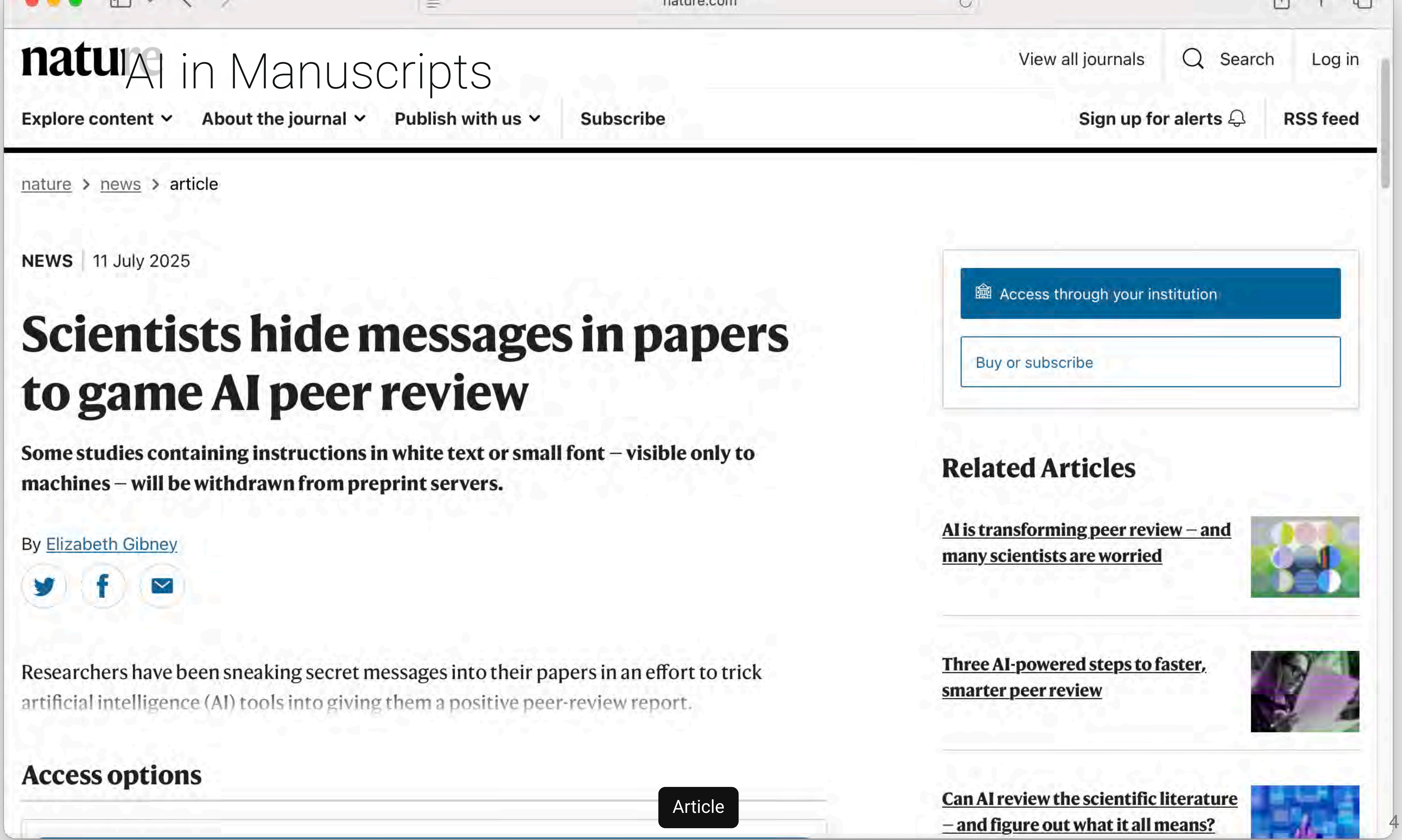
Key Findings

- Of the 53 articles with the fewest in-text citations, 48 appeared to have been generated by AI.

GLOBAL INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

Global Business Strategies in the Digital Age: A Comparative Analysis of

AI in Manuscripts



NEWS | 11 July 2025

Scientists hide messages in papers to game AI peer review

Some studies containing instructions in white text or small font – visible only to machines – will be withdrawn from preprint servers.

By Elizabeth Gibney

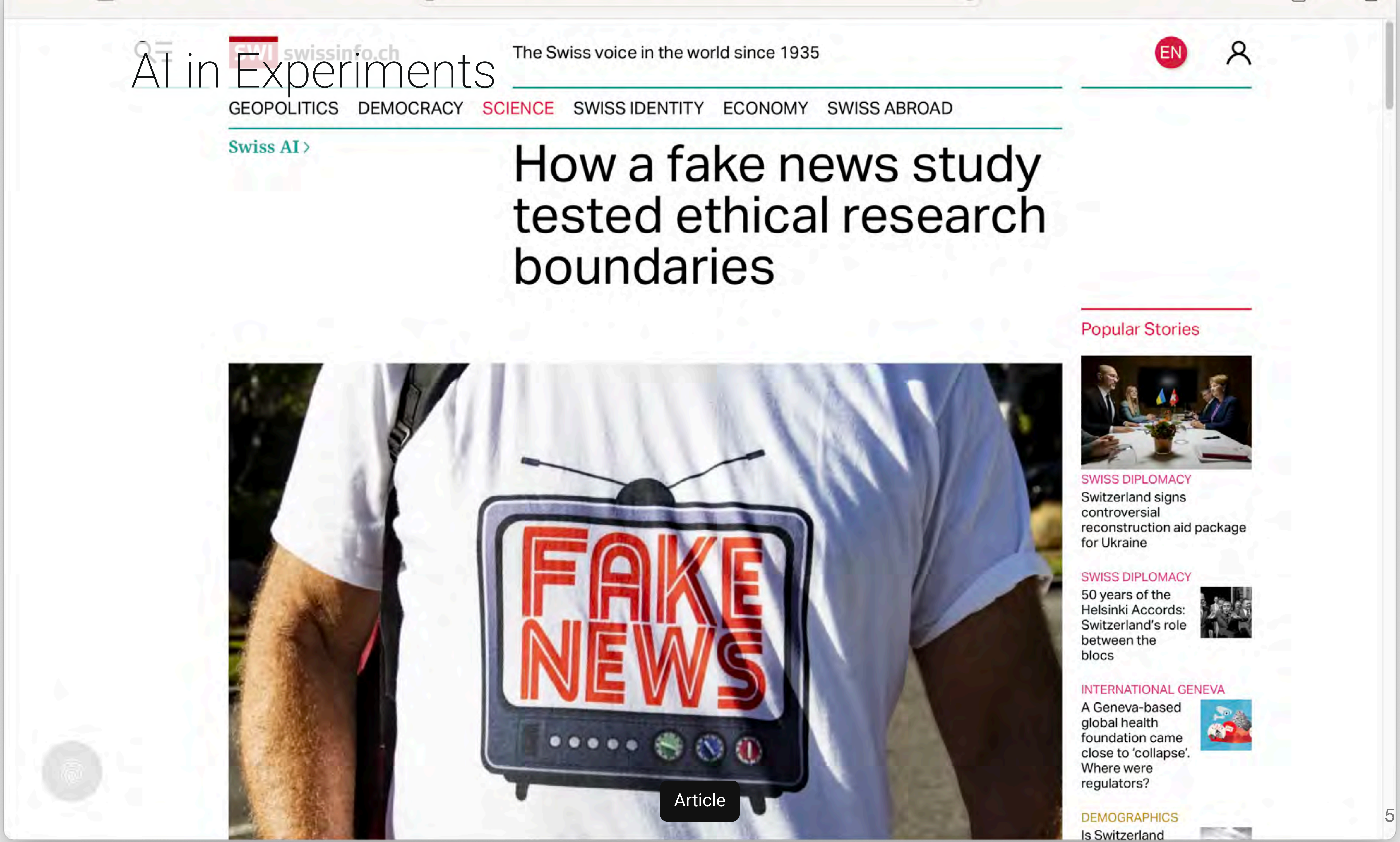
Researchers have been sneaking secret messages into their papers in an effort to trick artificial intelligence (AI) tools into giving them a positive peer-review report.

Access options

Related Articles

- AI is transforming peer review – and many scientists are worried
- Three AI-powered steps to faster, smarter peer review
- Can AI review the scientific literature – and figure out what it all means?

AI in Experiments



SWI swissinfo.ch

The Swiss voice in the world since 1935

EN

GEOPOLITICS DEMOCRACY SCIENCE SWISS IDENTITY ECONOMY SWISS ABROAD

Swiss AI ›

How a fake news study tested ethical research boundaries

Popular Stories

SWISS DIPLOMACY
Switzerland signs controversial reconstruction aid package for Ukraine

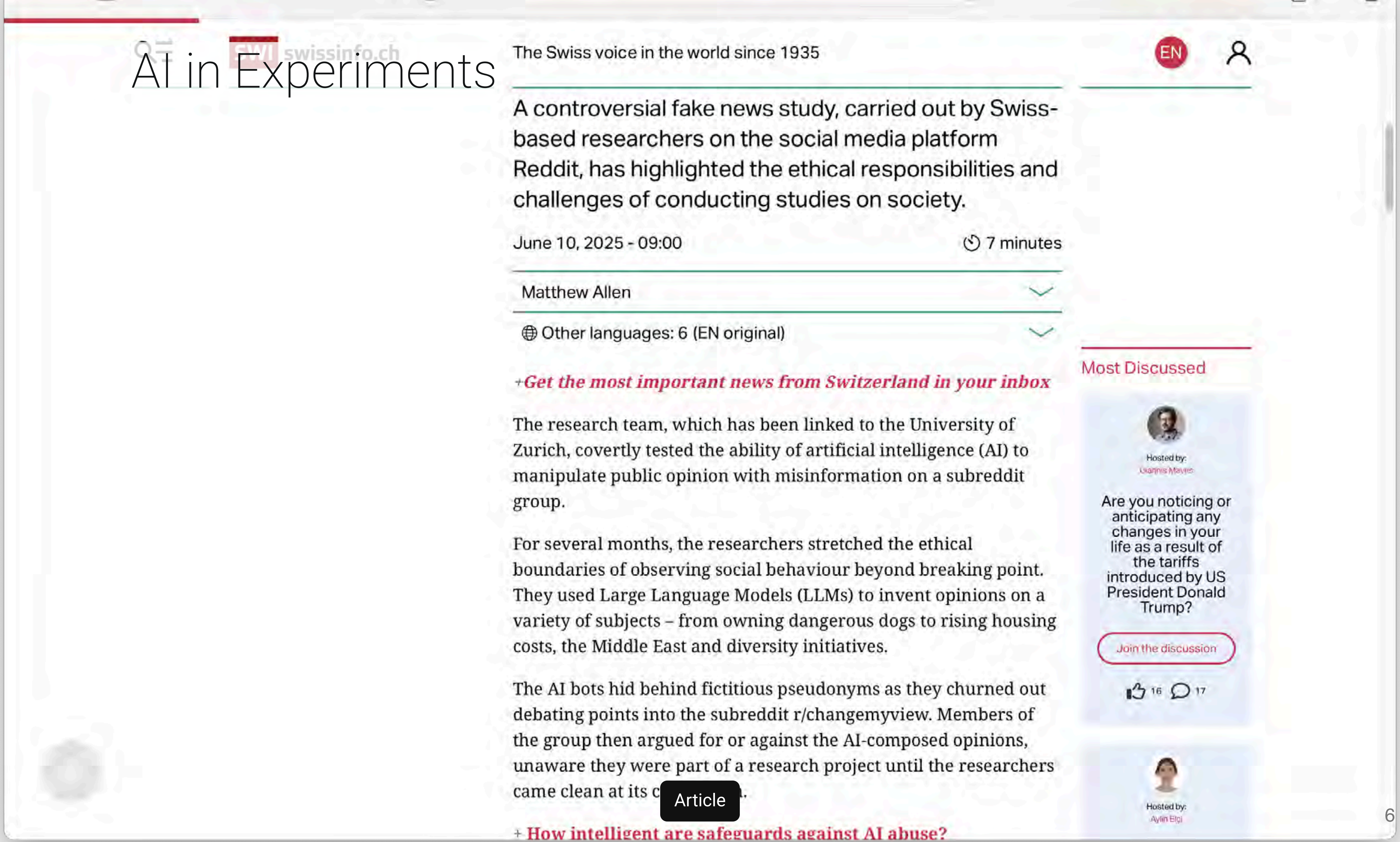
SWISS DIPLOMACY
50 years of the Helsinki Accords: Switzerland's role between the blocs

INTERNATIONAL GENEVA
A Geneva-based global health foundation came close to 'collapse'. Where were regulators?

DEMOGRAPHICS
Is Switzerland

Article

AI in Experiments



The Swiss voice in the world since 1935

A controversial fake news study, carried out by Swiss-based researchers on the social media platform Reddit, has highlighted the ethical responsibilities and challenges of conducting studies on society.

June 10, 2025 - 09:00

7 minutes

Matthew Allen

Other languages: 6 (EN original)

+*Get the most important news from Switzerland in your inbox*

The research team, which has been linked to the University of Zurich, covertly tested the ability of artificial intelligence (AI) to manipulate public opinion with misinformation on a subreddit group.

For several months, the researchers stretched the ethical boundaries of observing social behaviour beyond breaking point. They used Large Language Models (LLMs) to invent opinions on a variety of subjects – from owning dangerous dogs to rising housing costs, the Middle East and diversity initiatives.

The AI bots hid behind fictitious pseudonyms as they churned out debating points into the subreddit r/changemyview. Members of the group then argued for or against the AI-composed opinions, unaware they were part of a research project until the researchers came clean at its c[illegible].

+ How intelligent are safeguards against AI abuse?

Most Discussed

Hosted by:
Giannis Mavris

Are you noticing or anticipating any changes in your life as a result of the tariffs introduced by US President Donald Trump?

Hosted by:
Aylin Elçi

What can we do about it?

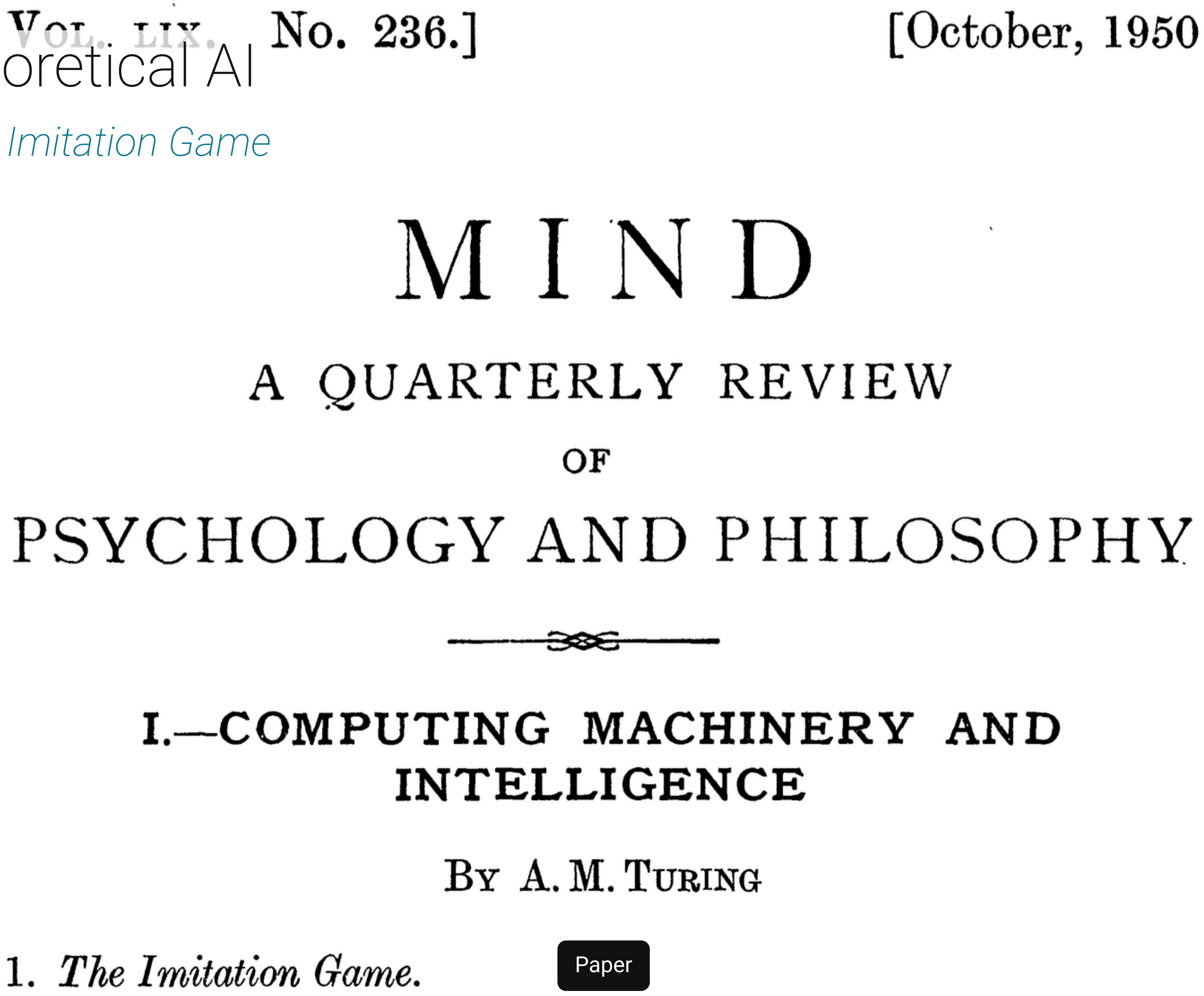
We can

1. Understand the technology and its history

A short history of AI

Theoretical AI

1950: Imitation Game



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

Paper

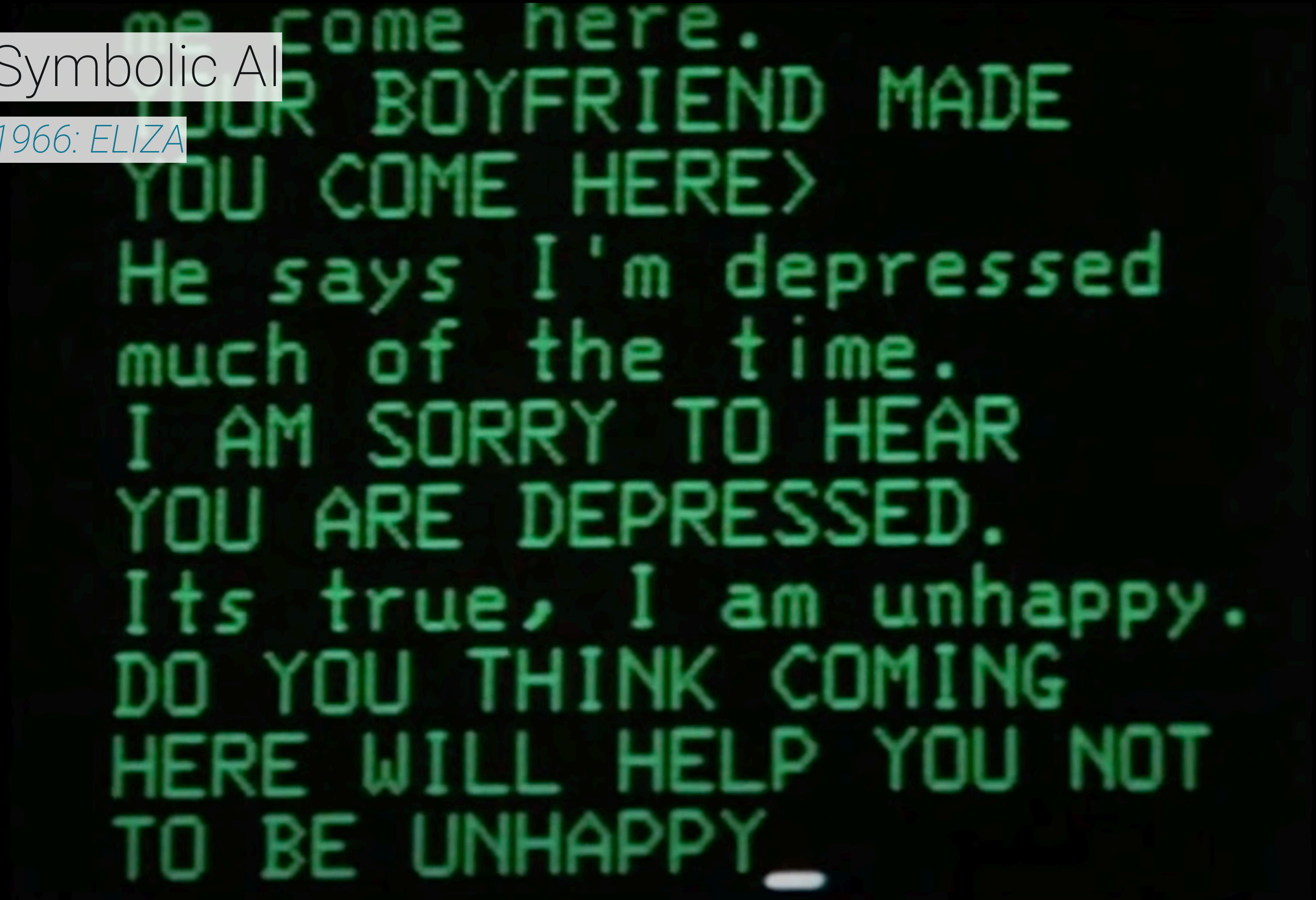
Theoretical AI

1950: Imitation Game

> The new form of the problem can be described in terms of a game which we call the ‘imitation game’. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two.
> (...)
> We now ask the question, ’What will happen when a machine takes the part of A in this game?

Symbolic AI

1966: ELIZA

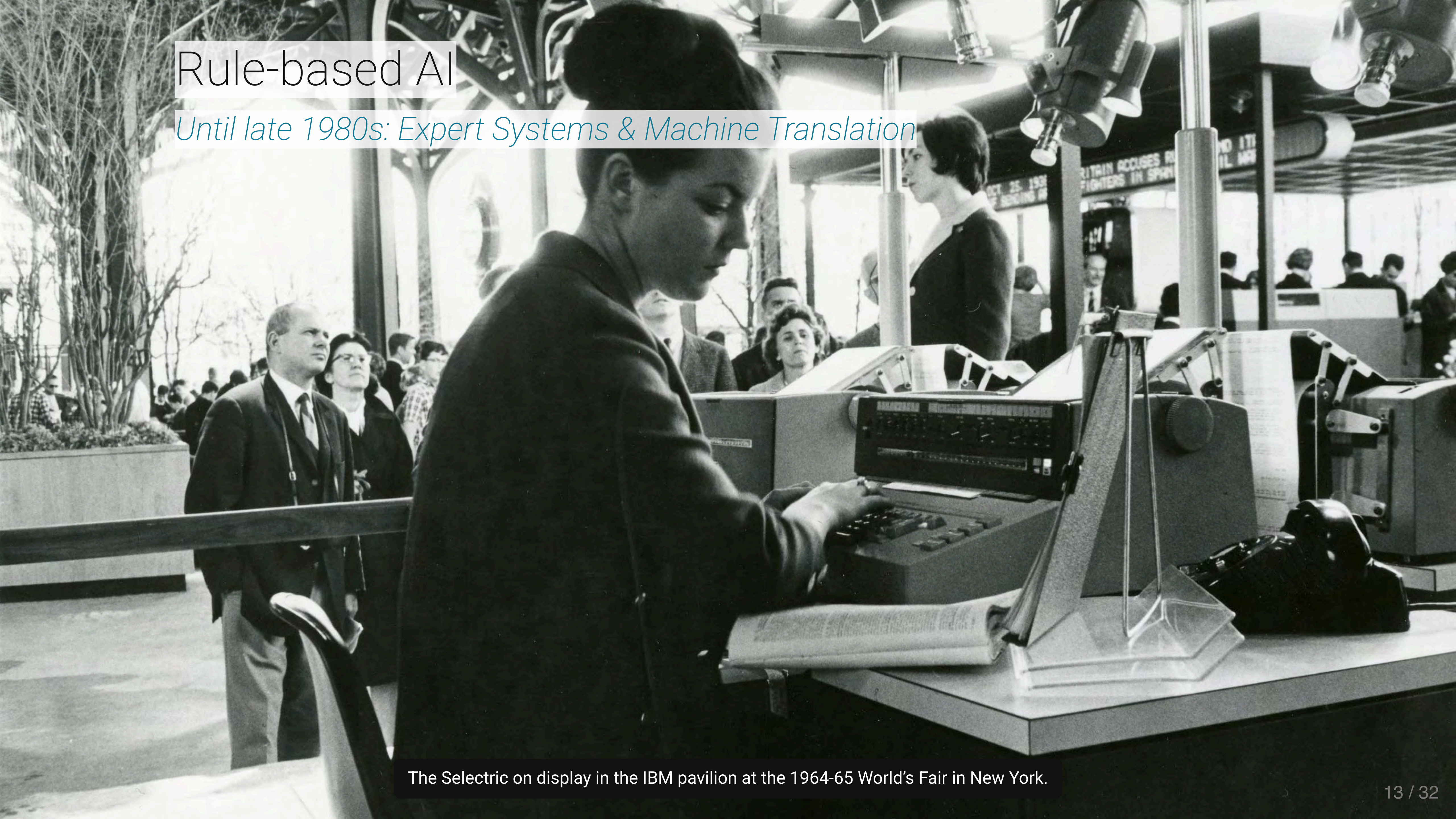


```
me come here.
YOUR BOYFRIEND MADE
YOU COME HERE>
He says I'm depressed
much of the time.
I AM SORRY TO HEAR
YOU ARE DEPRESSED.
Its true, I am unhappy.
DO YOU THINK COMING
HERE WILL HELP YOU NOT
TO BE UNHAPPY
```

BBC footage

Rule-based AI

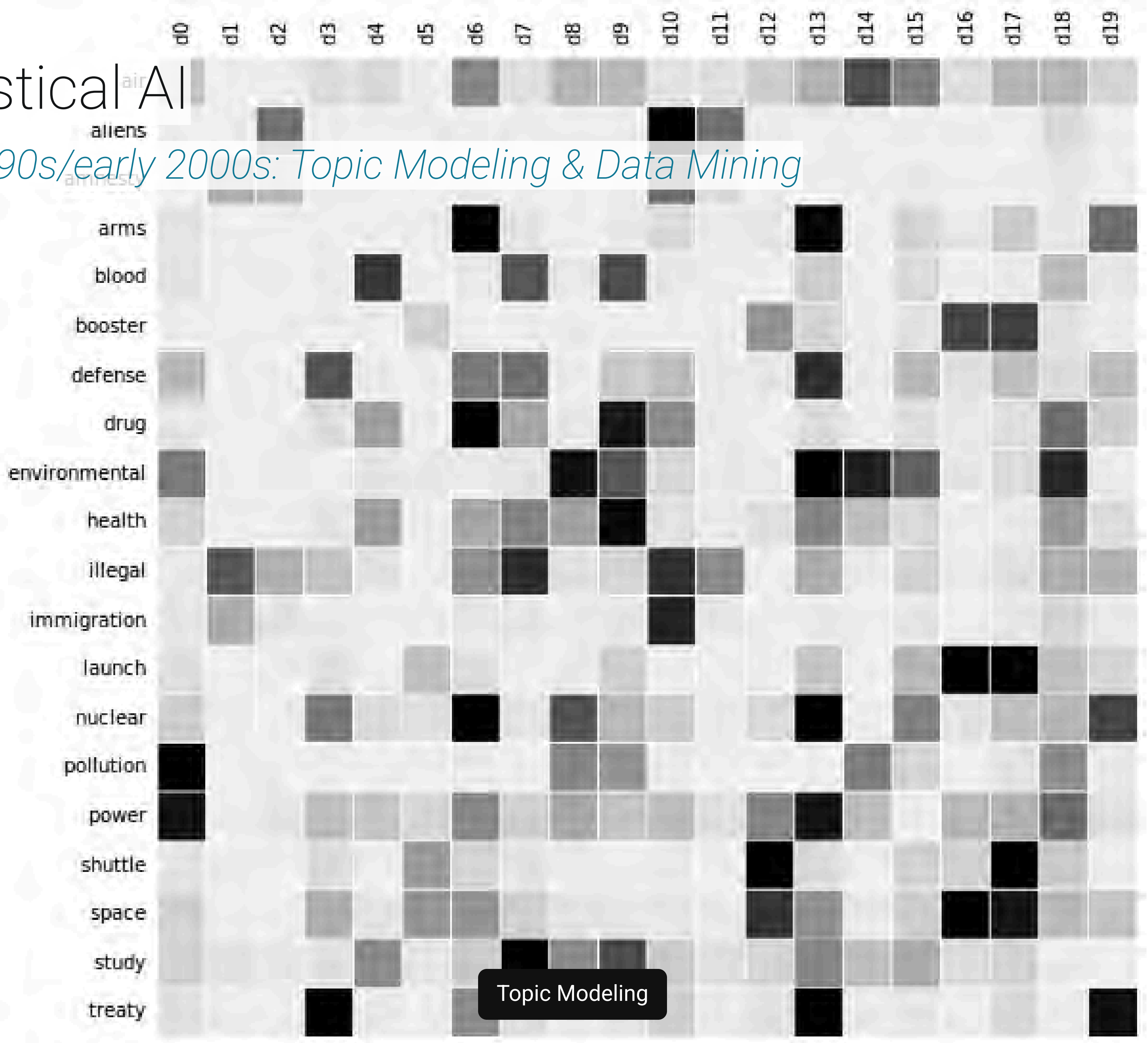
Until late 1980s: Expert Systems & Machine Translation



The Selectric on display in the IBM pavilion at the 1964-65 World’s Fair in New York.

Statistical AI

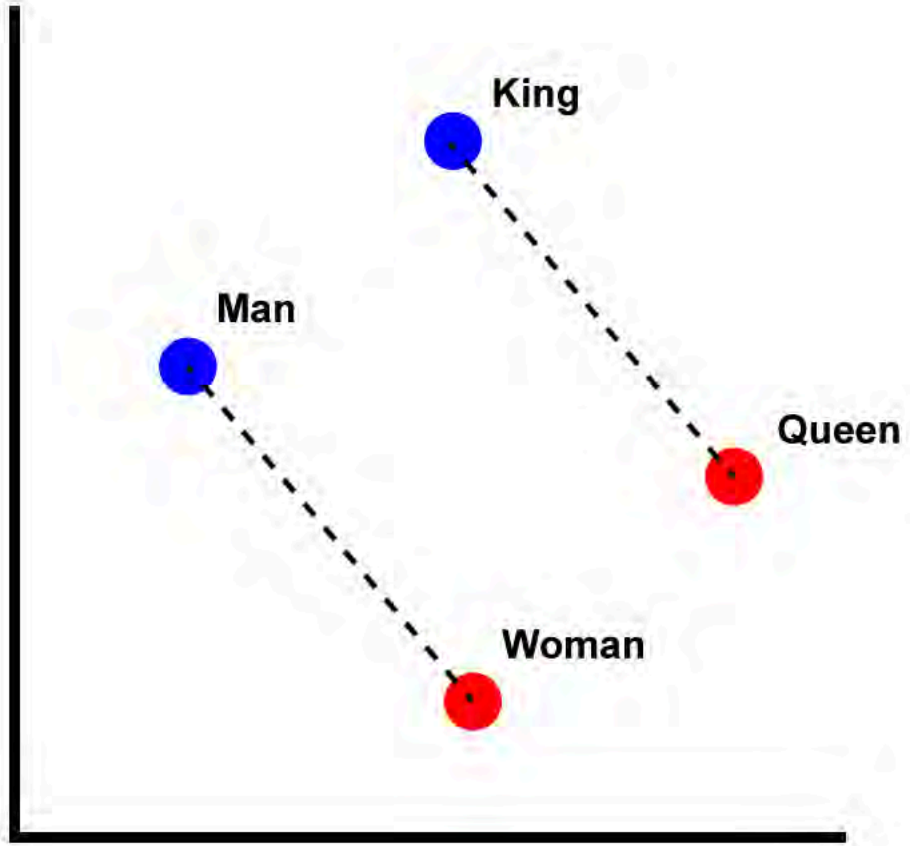
Late 1990s/early 2000s: Topic Modeling & Data Mining



Topic Modeling

Neural AI

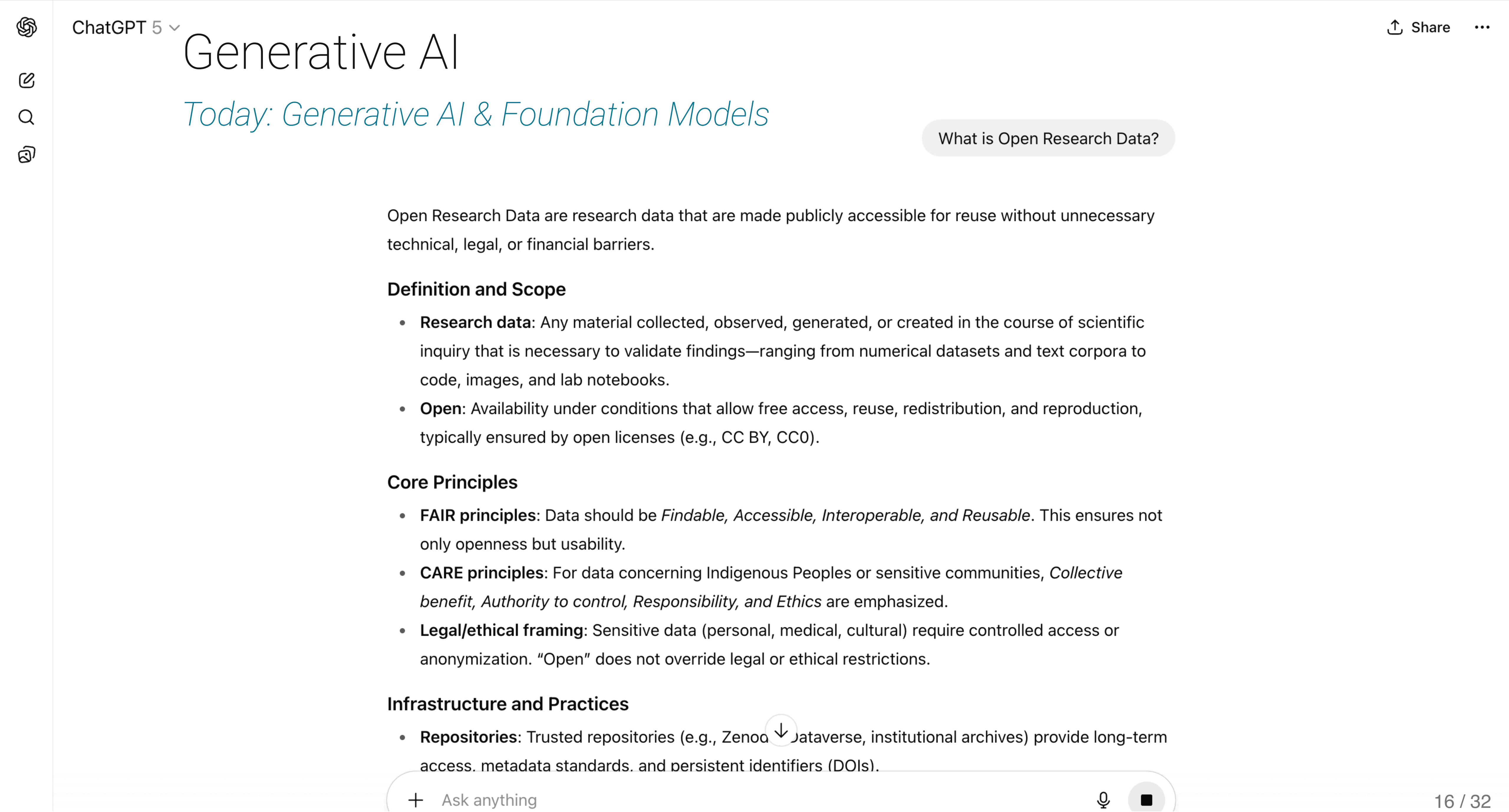
2010s: Deep Learning & Large Language Models



Word2Vec

Generative AI

Today: Generative AI & Foundation Models



What is Open Research Data?

Open Research Data are research data that are made publicly accessible for reuse without unnecessary technical, legal, or financial barriers.

Definition and Scope

- **Research data**: Any material collected, observed, generated, or created in the course of scientific inquiry that is necessary to validate findings—ranging from numerical datasets and text corpora to code, images, and lab notebooks.
- **Open**: Availability under conditions that allow free access, reuse, redistribution, and reproduction, typically ensured by open licenses (e.g., CC BY, CC0).

Core Principles

- **FAIR principles**: Data should be *Findable, Accessible, Interoperable, and Reusable*. This ensures not only openness but usability.
- **CARE principles**: For data concerning Indigenous Peoples or sensitive communities, *Collective benefit, Authority to control, Responsibility, and Ethics* are emphasized.
- **Legal/ethical framing**: Sensitive data (personal, medical, cultural) require controlled access or anonymization. "Open" does not override legal or ethical restrictions.

Infrastructure and Practices

- **Repositories**: Trusted repositories (e.g., Zenodo, Dataverse, institutional archives) provide long-term access, metadata standards, and persistent identifiers (DOIs).

Known problems of Generative AI

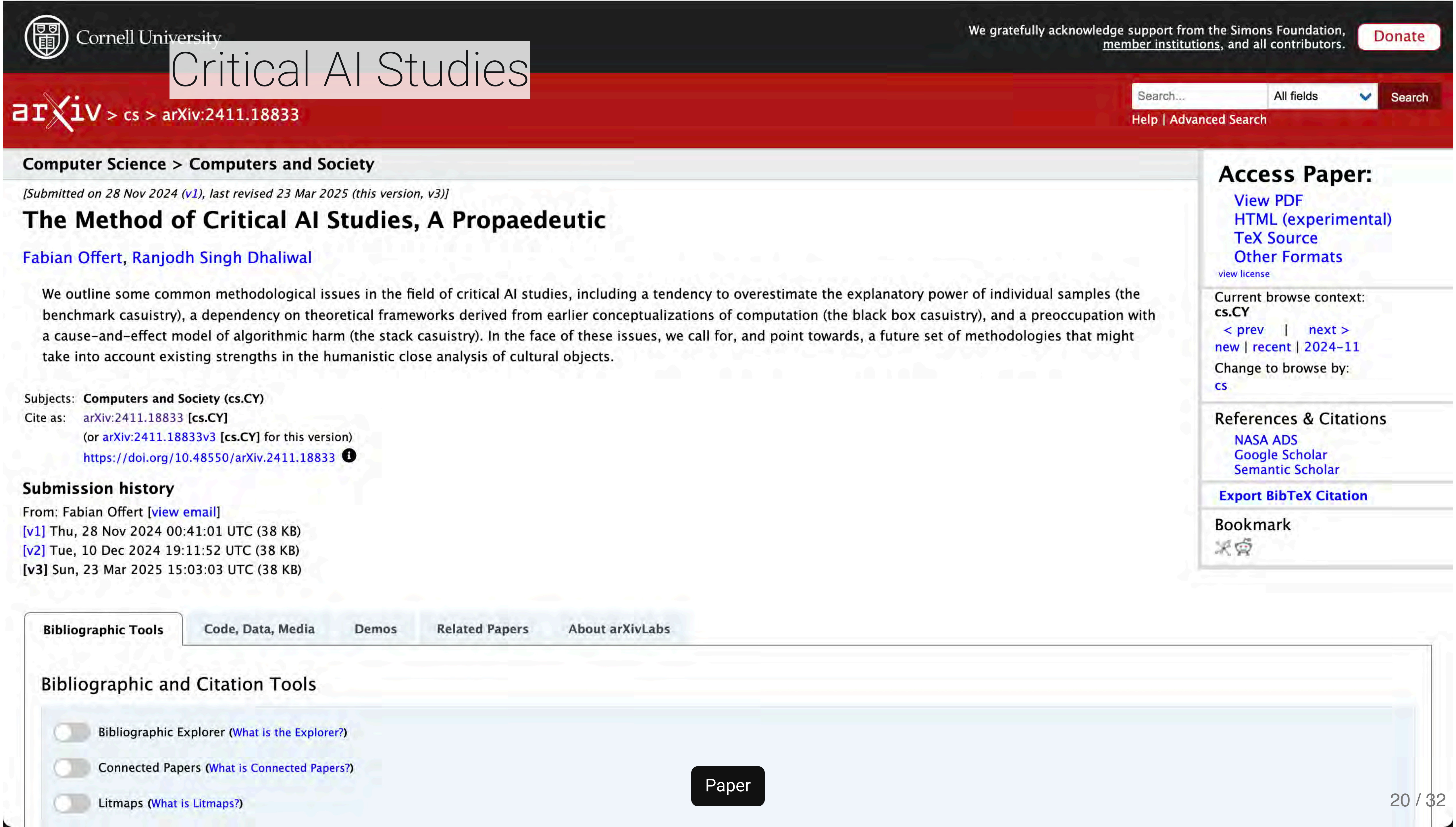
- **Bias** in training data
- Lack of **explainability**
- Lack of **transparency**
- Lack of **accountability**
- Lack of **reproducibility**
- **Environmental impact**
- **Ethical** issues
- **Legal** issues
- **Social** issues
- **Epistemological** issues
- ...

What can we do about it?

We can

1. Understand the technology and its history
2. Understand the limitations and problems of AI

Critical AI Studies



Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors.

Donate

arXiv > cs > arXiv:2411.18833

Search... All fields Search

Help | Advanced Search

Computer Science > Computers and Society

[Submitted on 28 Nov 2024 (v1), last revised 23 Mar 2025 (this version, v3)]

The Method of Critical AI Studies, A Propaedeutic

Fabian Offert, Ranjodh Singh Dhaliwal

We outline some common methodological issues in the field of critical AI studies, including a tendency to overestimate the explanatory power of individual samples (the benchmark casuistry), a dependency on theoretical frameworks derived from earlier conceptualizations of computation (the black box casuistry), and a preoccupation with a cause-and-effect model of algorithmic harm (the stack casuistry). In the face of these issues, we call for, and point towards, a future set of methodologies that might take into account existing strengths in the humanistic close analysis of cultural objects.

Subjects:	**Computers and Society (cs.CY)**
Cite as:	arXiv:2411.18833 **[cs.CY]**
	(or arXiv:2411.18833v3 **[cs.CY]** for this version)
	https://doi.org/10.48550/arXiv.2411.18833 ℹ

Submission history

From: Fabian Offert [view email]

[v1] Thu, 28 Nov 2024 00:41:01 UTC (38 KB)

[v2] Tue, 10 Dec 2024 19:11:52 UTC (38 KB)

[v3] Sun, 23 Mar 2025 15:03:03 UTC (38 KB)

Access Paper:

- View PDF
- HTML (experimental)
- TeX Source
- Other Formats

view license

Current browse context:
cs.CY

< prev | next >

new | recent | 2024-11

Change to browse by:

cs

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark

Bibliographic Tools | Code, Data, Media | Demos | Related Papers | About arXivLabs

Bibliographic and Citation Tools

○ Bibliographic Explorer (What is the Explorer?)

○ Connected Papers (What is Connected Papers?)

○ Litmaps (What is Litmaps?)

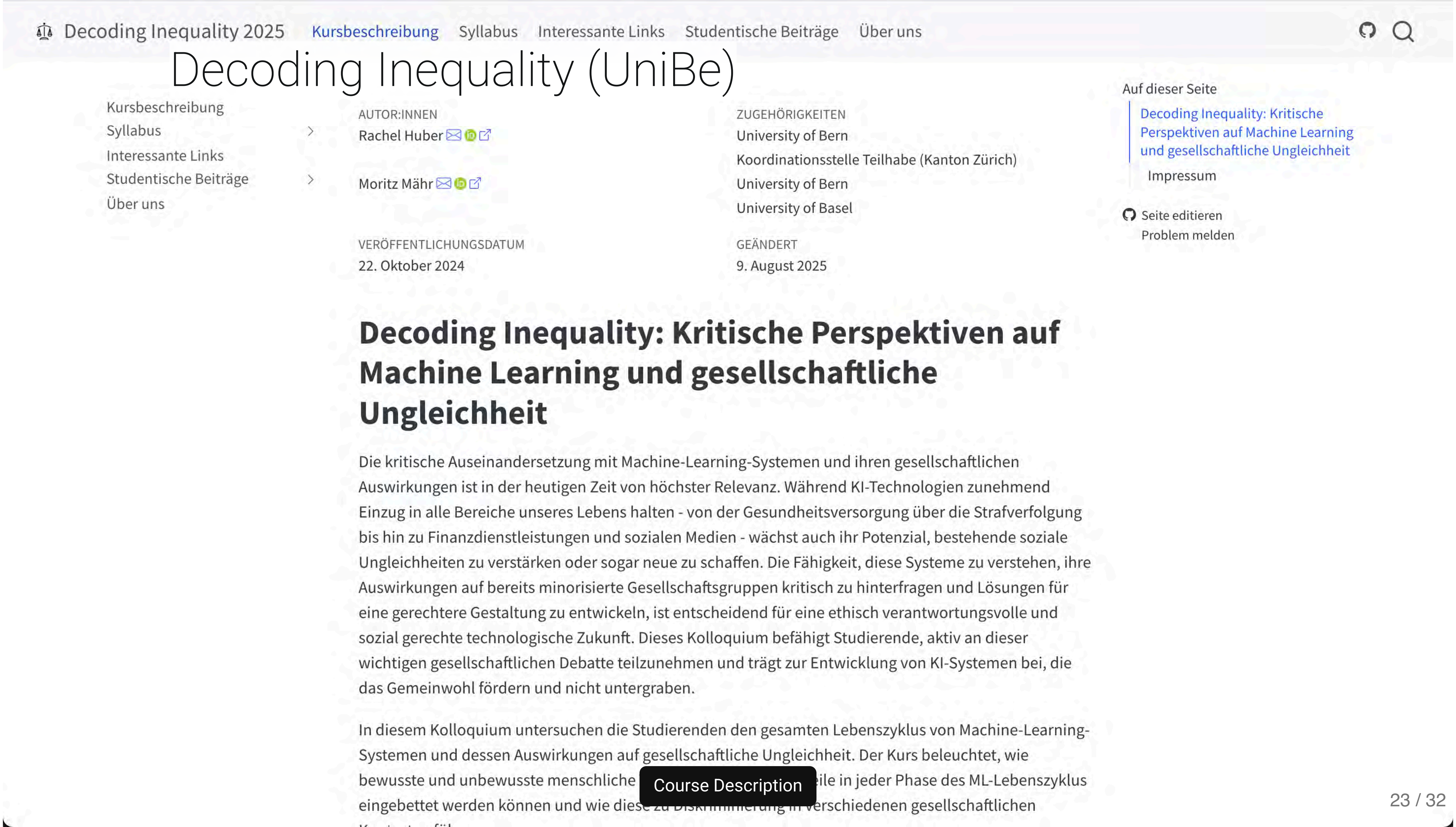
Paper

Teaching AI Literacy

Critical AI Literacy

- **Technical literacy**: Understanding how AI systems work, their capabilities and limitations
- **Epistemological awareness**: Questioning what counts as knowledge and how AI shapes it
- **Ethical evaluation**: Considering consent, privacy, transparency, and accountability
- **Social impact assessment**: Examining power structures, equity, and broader implications
- **Practical application**: Developing workflows that maintain scholarly rigor
- **Continuous learning**: Staying informed as technology evolves rapidly

Decoding Inequality (UniBe)



AUTOR:INNEN	ZUGEHÖRIGKEITEN
Rachel Huber	University of Bern Koordinationsstelle Teilhabe (Kanton Zürich)
Moritz Mähr	University of Bern University of Basel

VERÖFFENTLICHUNGSDATUM	GEÄNDERT
22. Oktober 2024	9. August 2025

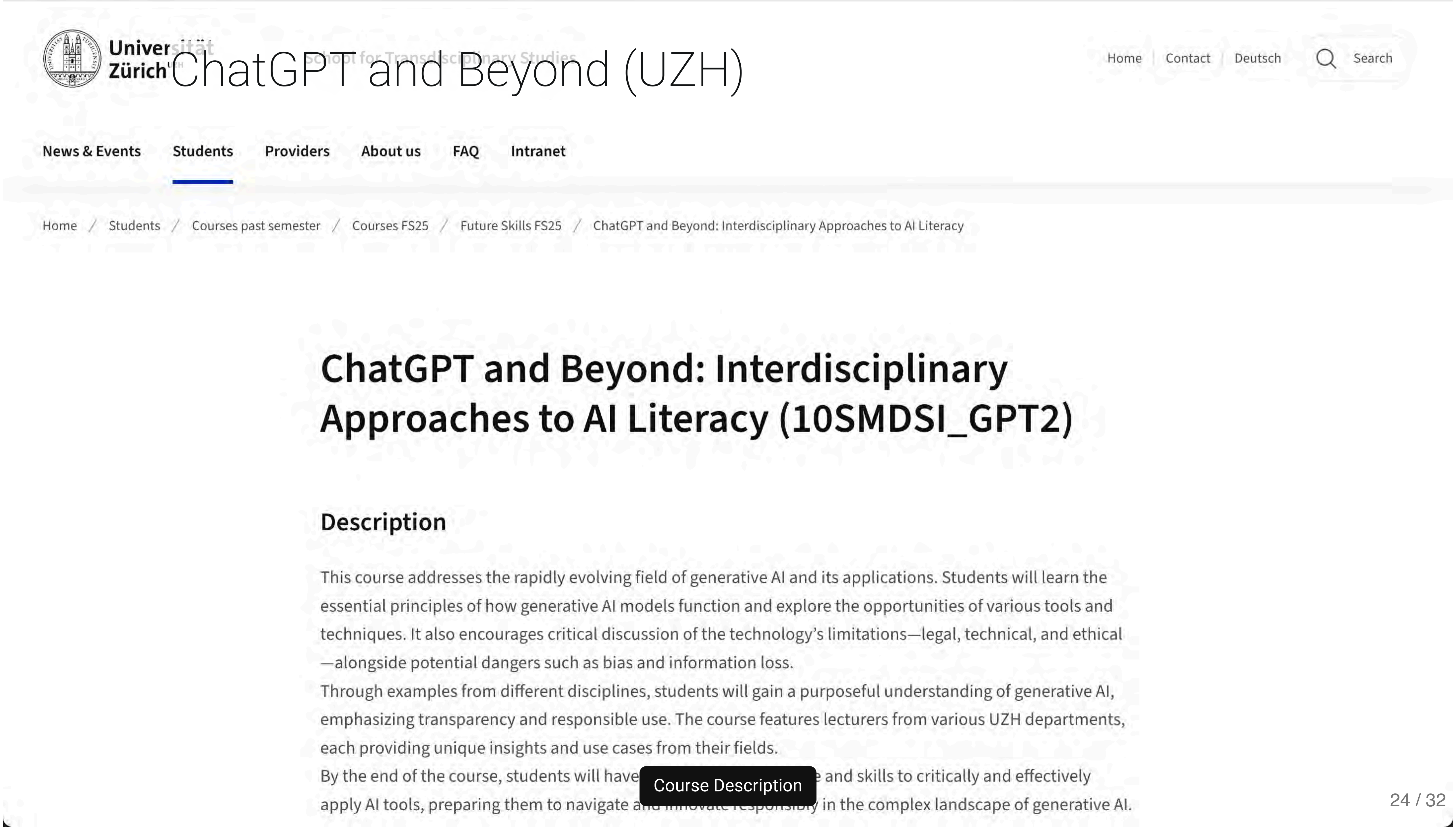
Decoding Inequality: Kritische Perspektiven auf Machine Learning und gesellschaftliche Ungleichheit

Die kritische Auseinandersetzung mit Machine-Learning-Systemen und ihren gesellschaftlichen Auswirkungen ist in der heutigen Zeit von höchster Relevanz. Während KI-Technologien zunehmend Einzug in alle Bereiche unseres Lebens halten - von der Gesundheitsversorgung über die Strafverfolgung bis hin zu Finanzdienstleistungen und sozialen Medien - wächst auch ihr Potenzial, bestehende soziale Ungleichheiten zu verstärken oder sogar neue zu schaffen. Die Fähigkeit, diese Systeme zu verstehen, ihre Auswirkungen auf bereits minorisierte Gesellschaftsgruppen kritisch zu hinterfragen und Lösungen für eine gerechtere Gestaltung zu entwickeln, ist entscheidend für eine ethisch verantwortungsvolle und sozial gerechte technologische Zukunft. Dieses Kolloquium befähigt Studierende, aktiv an dieser wichtigen gesellschaftlichen Debatte teilzunehmen und trägt zur Entwicklung von KI-Systemen bei, die das Gemeinwohl fördern und nicht untergraben.

In diesem Kolloquium untersuchen die Studierenden den gesamten Lebenszyklus von Machine-Learning-Systemen und dessen Auswirkungen auf gesellschaftliche Ungleichheit. Der Kurs beleuchtet, wie bewusste und unbewusste menschliche ile in jeder Phase des ML-Lebenszyklus eingebettet werden können und wie diese zu Diskriminierung in verschiedenen gesellschaftlichen

Course Description

ChatGPT and Beyond (UZH)



Universität Zürich UZH
School for Transdisciplinary Studies
Home | Contact | Deutsch
Search
News & Events
Students
Providers
About us
FAQ
Intranet
Home / Students / Courses past semester / Courses FS25 / Future Skills FS25 / ChatGPT and Beyond: Interdisciplinary Approaches to AI Literacy

ChatGPT and Beyond: Interdisciplinary Approaches to AI Literacy (10SMDSI_GPT2)

Description

This course addresses the rapidly evolving field of generative AI and its applications. Students will learn the essential principles of how generative AI models function and explore the opportunities of various tools and techniques. It also encourages critical discussion of the technology's limitations—legal, technical, and ethical—alongside potential dangers such as bias and information loss.

Through examples from different disciplines, students will gain a purposeful understanding of generative AI, emphasizing transparency and responsible use. The course features lecturers from various UZH departments, each providing unique insights and use cases from their fields.

By the end of the course, students will have e and skills to critically and effectively apply AI tools, preparing them to navigate and innovate responsibly in the complex landscape of generative AI.

Course Description

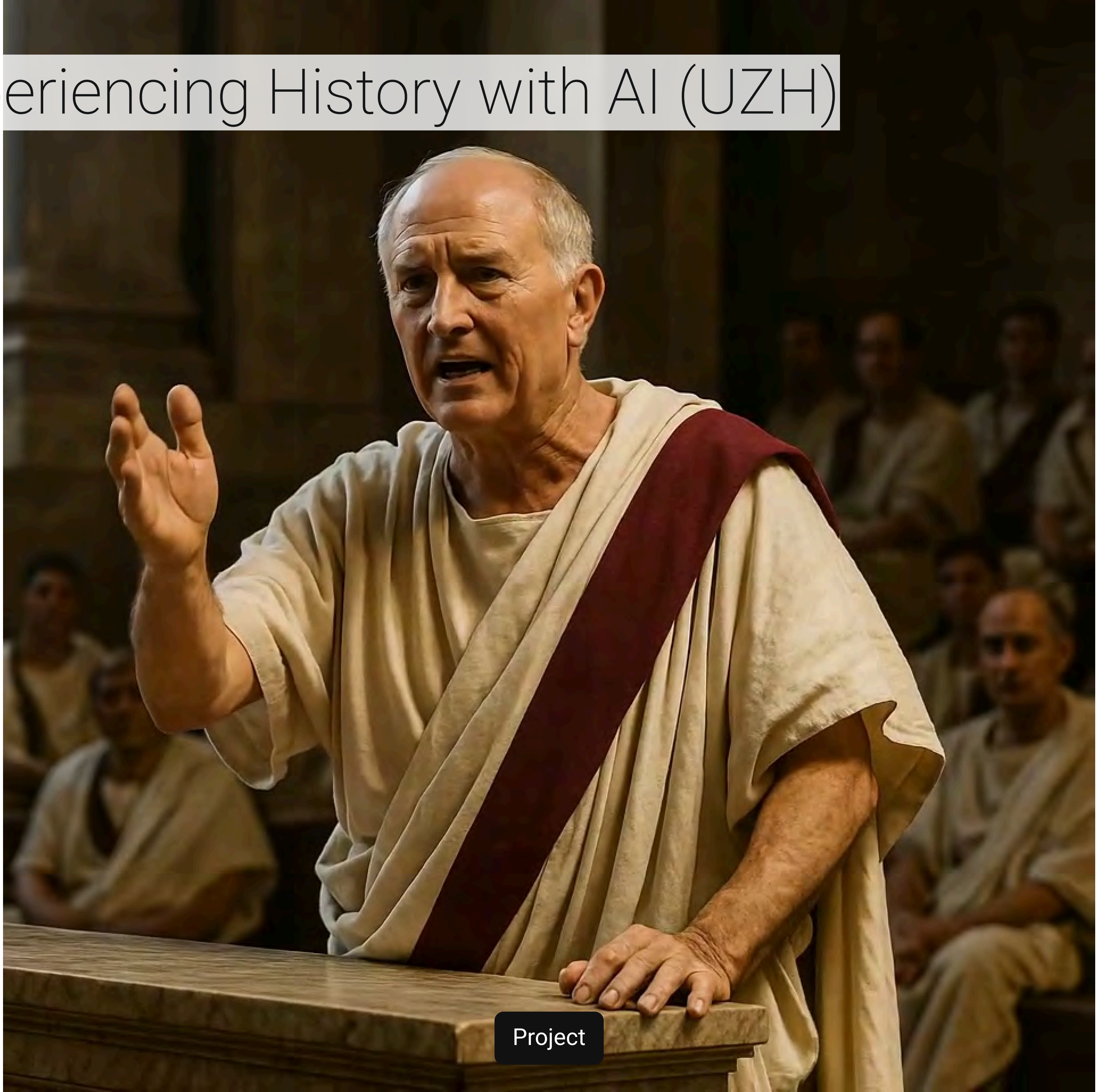
What can we do about it?

We can

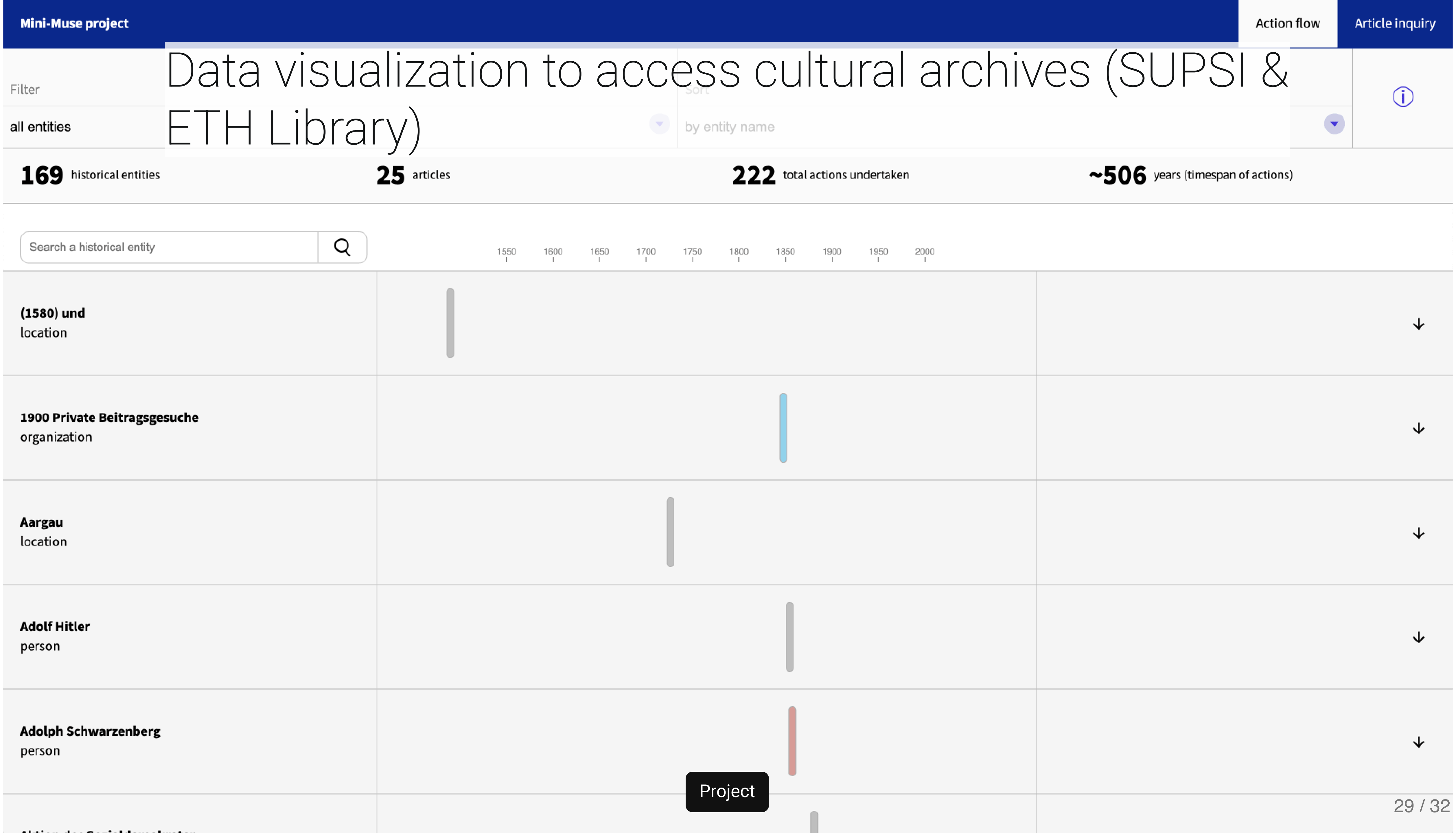
1. Understand the technology and its history
2. Understand the limitations and problems of AI
3. Make better use of AI tools

DH in Action: Swiss Projects Using LLMs (Tools & Platforms)

Re-Experiencing History with AI (UZH)



Data visualization to access cultural archives (SUPSI & ETH Library)



169 historical entities **25** articles **222** total actions undertaken **~506** years (timespan of actions)

(1580) und
location

1900 Private Beitragsgesuche
organization

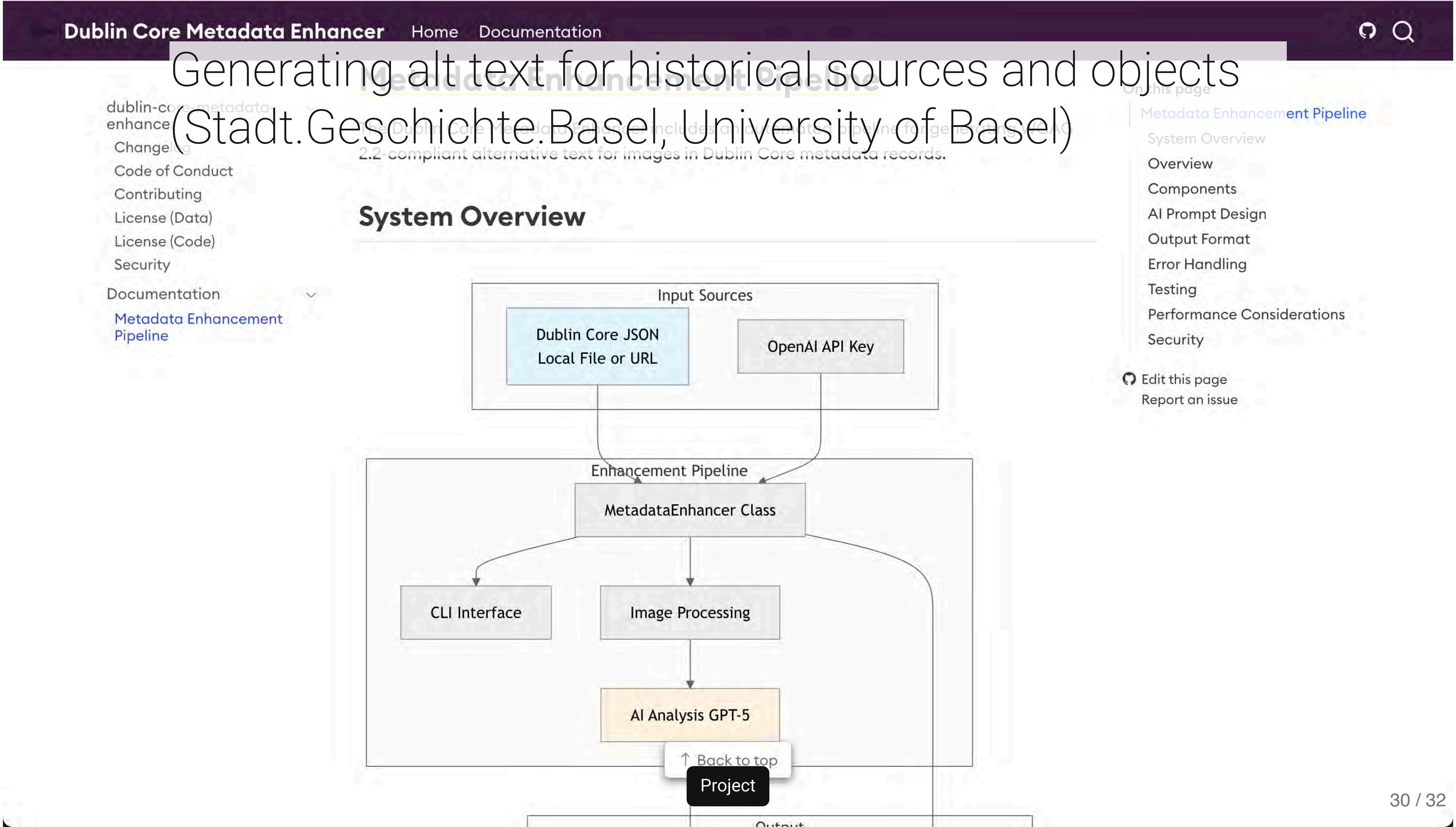
Aargau
location

Adolf Hitler
person

Adolph Schwarzenberg
person

Project

Generating alt text for historical sources and objects (Stadt.Geschichte.Basel, University of Basel)



Dublin Core Metadata Enhancer Home Documentation

dublin-core-metadata-enhancer
- Changelog
- Code of Conduct
- Contributing
- License (Data)
- License (Code)
- Security

Documentation
- Metadata Enhancement Pipeline

Metadata Enhancement Pipeline

The Dublin Core Metadata Enhancer includes an automated pipeline for generating WCAG 2.2-compliant alternative text for images in Dublin Core metadata records.

System Overview

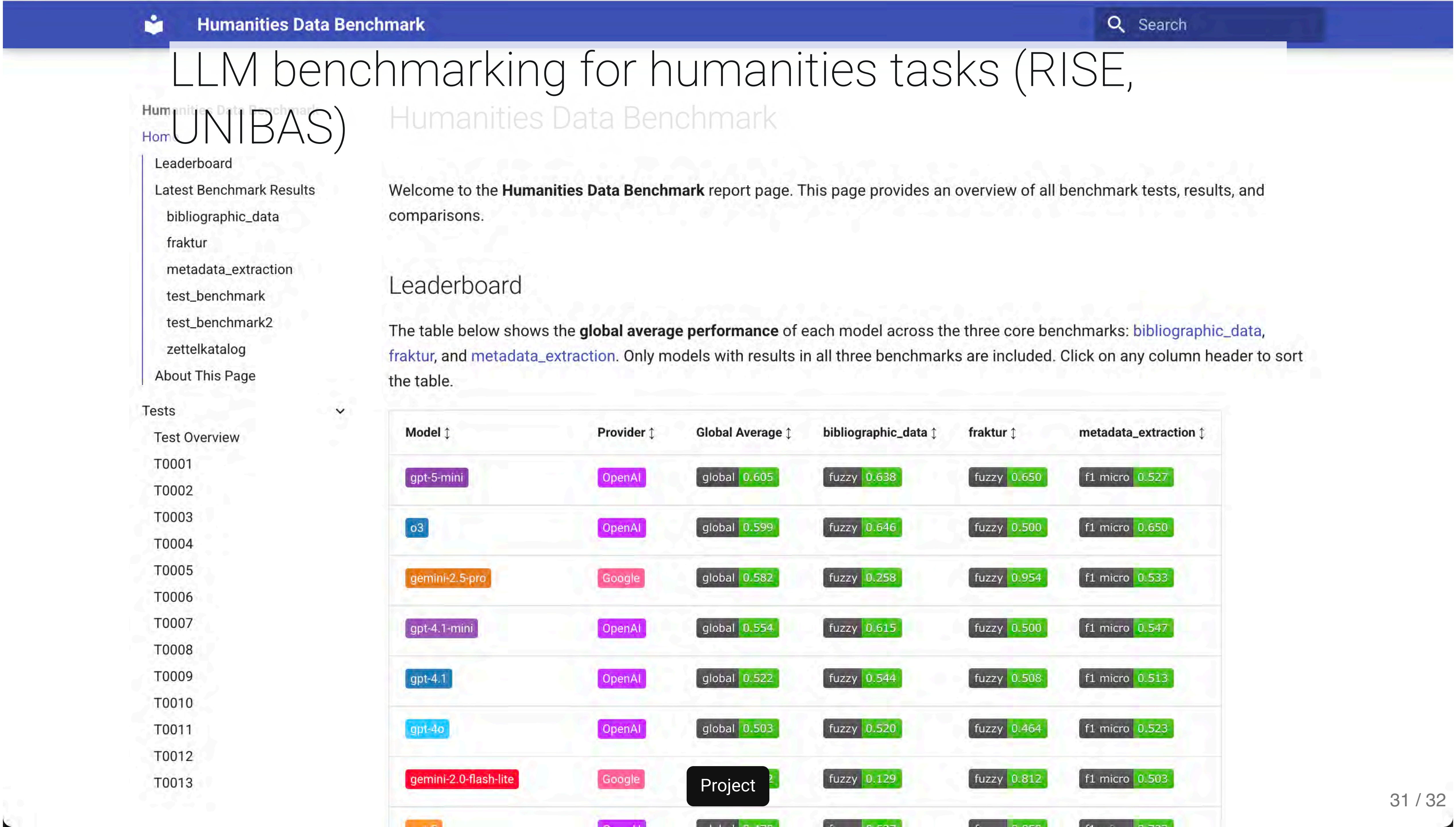
↑ Back to top

On this page
- Metadata Enhancement Pipeline
- System Overview
- Overview
- Components
- AI Prompt Design
- Output Format
- Error Handling
- Testing
- Performance Considerations
- Security

Edit this page
Report an issue

Project

LLM benchmarking for humanities tasks (RISE, UNIBAS)



Humanities Data Benchmark

Search

Humanities Data Benchmark

Home

- Leaderboard
- Latest Benchmark Results
  - bibliographic_data
  - fraktur
  - metadata_extraction
  - test_benchmark
  - test_benchmark2
  - zettelkatalog
- About This Page

Tests

- Test Overview
- T0001
- T0002
- T0003
- T0004
- T0005
- T0006
- T0007
- T0008
- T0009
- T0010
- T0011
- T0012
- T0013

Humanities Data Benchmark

Welcome to the **Humanities Data Benchmark** report page. This page provides an overview of all benchmark tests, results, and comparisons.

Leaderboard

The table below shows the **global average performance** of each model across the three core benchmarks: bibliographic_data, fraktur, and metadata_extraction. Only models with results in all three benchmarks are included. Click on any column header to sort the table.

Model ↕	Provider ↕	Global Average ↕	bibliographic_data ↕	fraktur ↕	metadata_extraction ↕
gpt-5-mini	OpenAI	global 0.605	fuzzy 0.638	fuzzy 0.650	f1 micro 0.527
o3	OpenAI	global 0.599	fuzzy 0.646	fuzzy 0.500	f1 micro 0.650
gemini-2.5-pro	Google	global 0.582	fuzzy 0.258	fuzzy 0.954	f1 micro 0.533
gpt-4.1-mini	OpenAI	global 0.554	fuzzy 0.615	fuzzy 0.500	f1 micro 0.547
gpt-4.1	OpenAI	global 0.522	fuzzy 0.544	fuzzy 0.508	f1 micro 0.513
gpt-4o	OpenAI	global 0.503	fuzzy 0.520	fuzzy 0.464	f1 micro 0.523
gemini-2.0-flash-lite	Google	[illegible]	fuzzy 0.129	fuzzy 0.812	f1 micro 0.503

Project

Bibliography

- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜». In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.
- Long, Duri, and Brian Magerko. «What Is AI Literacy? Competencies and Design Considerations». In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3313831.3376727.
- Loukissas, Yanni A. *All Data Are Local: Thinking Critically in a Data-Driven Society*. Cambridge, Massachusetts: The MIT Press, 2019. https://doi.org/10.7551/mitpress/11543.001.0001.
- Mueller, Milton L. «It's Just Distributed Computing: Rethinking AI Governance». *Telecommunications Policy*, Februar 2025, 102917. https://doi.org/10.1016/j.telpol.2025.102917.
- O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown Publishing Group, 2016.
- Offert, Fabian, and Ranjodh Singh Dhaliwal. «The Method of Critical AI Studies, A Propaedeutic», 10. Dezember 2024. https://doi.org/10.48550/arXiv.2411.18833.