05.05.2026

Module 4: AI in research & scientific writing

Michael Bachmann

Poll about AI use

- https://forms.cloud.microsoft/e/c9EVNeTYvv

- Please fill out this poll about your AI use and your interest

  - Results will be collected for all AI courses and will be shared

Overview

- A bit of background
- AI in publishing: the policies and rules
- Use cases for AI in scientific writing and publishing
- Discussion, open questions, etc

Program

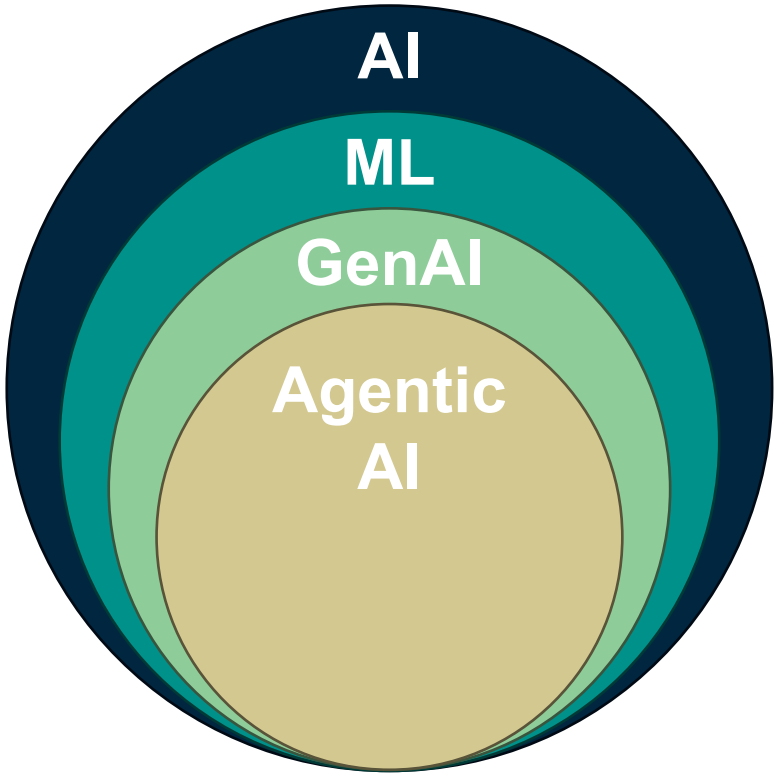
Time	Topic
13:00 – 13:40	**Part 01**
	Break 5`
13:45 – 14:30	**Part 02**
	Break 15`
14:45 – 15:15	**Part 03**
	Break 5`
15:20 – 16:00	**Part 04**

→ Slides ?
Lib4RI website: https://www.lib4ri.ch/trainings

→ Feedback !
Direct, email, social media, …

A bit of background

A bit of background: terminology



AI is the broad field of technology concerning computer systems capable of performing tasks that typically require human intelligence, such as perception, language understanding, learning and reasoning.

Machine Learning (ML) is the field of study in AI concerned with the development and study of statistical algorithms that learn patterns from data to make predictions or decisions without being explicitly programmed for each task.

Generative AI is a type of machine learning that uses statistical models to learn patterns in data and then create new data, such as text, images or music, that resembles the training examples.

Agentic AI is an AI system that can make decisions and act independently within the scope of its objectives. It can therefore carry out research tasks autonomously, such as literature searches, data collection, or experimental design.

A bit of background: terminology

- RAG LLM (retrieval augmented generation LLM):
  - Documents provided by the users are the knowledge source for the answer created by the LLM (in contrast to the knowledge base of the LLM itself or a web search).
  - Documents are cut in chunks of certain size. A certain number of chunks are used to create the answer if they are ranked to be helpful for this question.

AI is the broad field of technology concerning computer systems capable of performing tasks that typically require human intelligence, such as perception, language understanding, learning and reasoning.

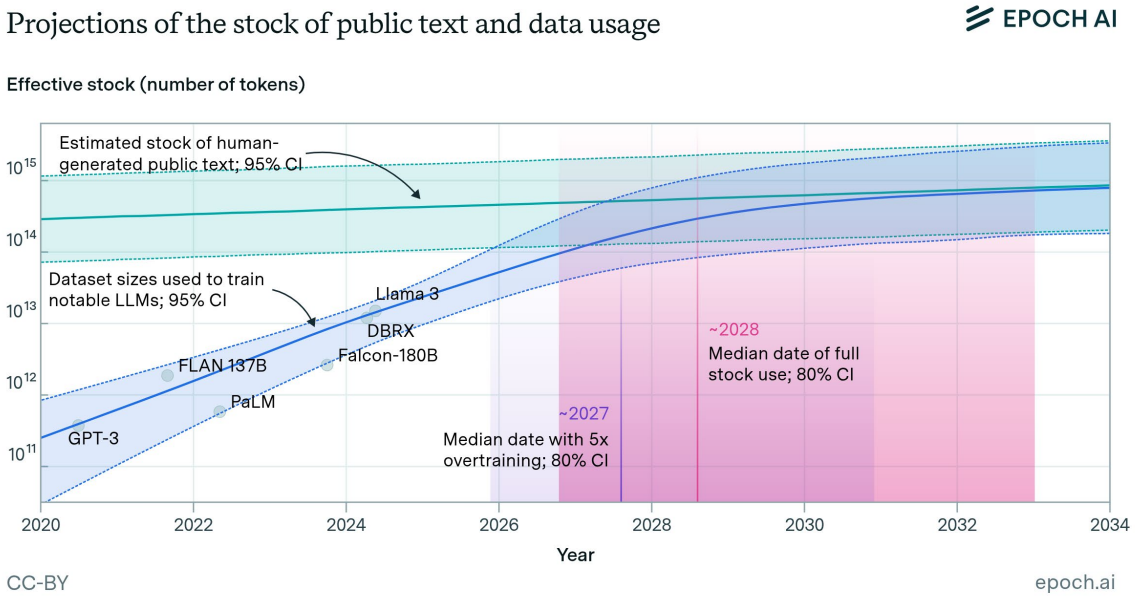
Machine Learning (ML) is the field of study in AI concerned with the development and study of statistical algorithms that learn patterns from data to make predictions or decisions without being explicitly programmed for each task.

Generative AI is a type of machine learning that uses statistical models to learn patterns in data and then create new data, such as text, images or music, that resembles the training examples.

Agentic AI is an AI system that can make decisions and act independently within the scope of its objectives. It can therefore carry out research tasks autonomously, such as literature searches, data collection, or experimental design.

A bit of background: LLM training

- Training sets for LLMs: The bigger, the better
- Common Crawl data set: “We are pleased to announce the release of the March 2026 crawl, containing 1.97 billion web pages, or 344.64 TiB of uncompressed content.” (https://commoncrawl.org/blog/march-2026-crawl-archive-now-available)
  - Plenty of copyright protected material
  - “Ethical” LLMs: Apertus for example
- Certain data are more important
  - Wikipedia
  - Books
  - Academic papers



Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. ‘Will we run out of data? Limits of LLM scaling based on human-generated data’. 2024. arXiv. https://arxiv.org/abs/2211.04325.

A bit of background: Text data mining (TDM) vs. genAI

- Question: “What is known about forever chemicals in the Swiss ecosystem?”
- TDM:
  - Create a query: (“forever chemicals” OR “PFAs” …) AND (“Swiss” OR “Switzerland” …)
  - Search in large text corpus with this query
  - How often is the query fulfilled?
  - Show complete manuscript, only part, …?
  - Interpret the result
- genAI:
  - Ask LLM: “What is known about forever chemicals in the Swiss ecosystem?”
  - Magic happens inside the LLM
  - LLM gives you an answer

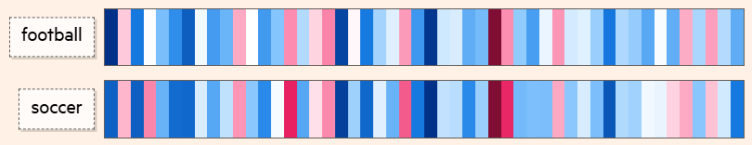
TDM vs. AI in legal context

TDM is decades old and there are laws and court decisions for TDM in the context of copyright. The legal situation around AI will be strongly affected if AI and TDM are (legally) considered to be the same (or if only a certain area of AI is considered to be equal to TDM).

“Generative AI exists because of the transformer.
This is how it works / learns / thinks / hallucinates / writes “

- Read this article: https://ig.ft.com/generative-ai/

- After reading it, discuss with your neighbour
  - What are tokens?
  - Discuss the embedding example for soccer / football
    - Are they close or not? How can you decide this with this figure?
    - What are real life examples to explain differences or similarities between football and soccer?
  - Do you understand self-attention?
    - Phoenix can be a city, a person, a fantasy creature. Make up example sentences that show how an LLM can differentiate.
  - “Hallucination: It’s not a bug, it’s a feature.”
    - Why could hallucination be called an intrinsic element of genAI text?



AI in publishing: the policies and rules

Copyright and AI – AI is never an author

- Copyright law: “In Switzerland, copyright protection arises automatically upon creation of a work, regardless of any formality. Such a work must be an “intellectual creation” and must therefore have a human origin. […]” (https://www.legalis.ch/de/sic/berichte-rapports/copyright-in-artificially-generated-works/)

https://en.wikipedia.org/wiki/Th%C3%A9%C3%A2tre_D%27op%C3%A9ra_Spatial

- Your publication comes with rights and obligations.
  - You get copyrights, not the AI (assuming there is intellectual creation).
  - You are responsible.
  - Rules of good scientific conduct still apply

Rules of scientific integrity
Code of conduct for scientific integrity, published by the Swiss Academy of Sciences:
https://swiss-academies.ch/publications/kodex-fur-wissenschaftliche-integritat

Copyright and AI – You can violate copyright with your AI generation based on your prompts

- Your prompt can lead to a text, an image, … that infringes on the copyright of someone else
  - You are responsible when sharing these results of your prompts
  - Do you have to be afraid? Try to extrapolate from copyright lawsuits in the past.
    - Music, art, pictures, videos, …: Many lawsuits, big companies, individuals affected
    - Manuscripts, articles: individual researchers are not affected, publishers target researchgate, sci-hub, …

LLM training

The copyright situation for LLM training is very different to the one for LLM (or genAI) usage. What kind of material you are allowed to use for what kind of LLM training is under debate.

AI policies of big publishing houses: Check carefully AI output and declare your AI use

- For authors, example Frontiers:
  - “Authors should ***not list a generative AI technology as a co-author*** or author of any submitted manuscript. Generative AI technologies cannot be held accountable for all aspects of a manuscript and consequently do not meet the criteria required for authorship.
  - If the author of a submitted manuscript has used written or visual content produced by or edited using a generative AI technology, this ***use must follow all Frontiers guidelines and policies*****.**
  - Specifically, the author is responsible for ***checking the factual accuracy of any content created*** by the generative AI technology. This includes, but is not limited to, any ***quotes, citations or references***. Figures produced by or edited using a generative AI technology must be checked to ensure they accurately reflect the data presented in the manuscript. Authors must also check that any written or visual content produced by or edited using a generative AI technology is ***free from plagiarism***.
  - If the author of a submitted manuscript has used written or visual content produced by or edited using a generative AI technology, such ***use must be acknowledged in the acknowledgements section of the manuscript and the methods section if applicable***. This explanation must ***list the name, version, model, and source of the generative AI technology***. We encourage authors to ***upload all input prompts provided to a generative AI technology and outputs received from a generative AI technology in the supplementary files for the manuscript***.”

AI policies of big publishing houses: Check carefully AI output and declare your AI use

- Elsevier: https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals#0-about
- Wiley: https://www.wiley.com/en-us/publish/article/ai-guidelines/
  - Very extensive material
- Springer Nature: https://www.springernature.com/gp/policies/editorial-policies (different policies for subcollections, Springer, Nature Portfolio, BMC, etc)
- ACS: https://researcher-resources.acs.org/publish/aipolicy
- Taylor & Francis: https://taylorandfrancis.com/our-policies/ai-policy/
- MDPI: https://www.mdpi.com/ethics#_bookmark3
- Frontiers: https://www.frontiersin.org/journals/artificial-intelligence/for-authors/author-guidelines

AI policies from big publishers are not necessarily the AI policy of a journal published by them

- Society journals, company mergers, journal switched publisher, …
  - Policies can vary for different reasons from the official publisher policies
  - (Scientific) editors can act very independently

- Plan ahead how and where you want to publish (ideally: first three choices)
- Check their AI policies
- In case of doubt: contact editorial office

- Be transparent and declare AI use, try to safe all prompts and responses, carefully check all AI content you use.
  - But also: relax, getting published has always been full of opaque factors out of your control

AI policies: How realistic is the recommended self-declaration?

- Are you working / have you worked on a research project recently for which you have used genAI?
  - How would your self declaration of AI use look like?
  - Do you have the necessary information stored?
  - If a “correct” declaration becomes unreasonable:
    - What would you declare and what not?
    - What should be declared in your opinion?

Frontiers, excerpt of AI policy:

> “If the author of a submitted manuscript has used written or visual content produced by or edited using a generative AI technology, such ***use must be acknowledged in the acknowledgements section of the manuscript and the methods section if applicable***. This explanation must ***list the name, version, model, and source of the generative AI technology***. We encourage authors to ***upload all input prompts provided to a generative AI technology and outputs received from a generative AI technology in the supplementary files for the manuscript***.”
> https://www.frontiersin.org/journals/artificial-intelligence/for-authors/author-guidelines

AI policies for reviewer and for image creation – check carefully: they can vary more and/or contain stronger restrictions

- Reviewers (and editors): The authors' manuscript is not yours and (potentially) contains sensitive data for which you have no rights (data, copyright)
  - Uploading it to external servers is not permitted, even not for spellcheck
  - Journal / publisher might have dedicated services

- Images: Some publishers perceive that there are too many unresolved questions regarding copyright and images
  - Springer Nature: "While legal issues relating to AI-generated images and videos remain broadly unresolved, Springer Nature journals are unable to permit its use for publication.
    - Exceptions:
      - …"

- Both cases: Plan carefully ahead before using AI services and contact editorial offices

Use cases for AI in scientific writing and publishing

Use cases for AI in scientific writing and publishing

- Task: Use the provided material for one option of your choice and work through it (15 min). Explain briefly what it is about (5 min presentation)

  - Detection tools for genAI text: https://www.lib4ri.ch/sites/default/files/media/documents/AI-detection_Summary_MBachmann.pdf
  - Can genAI do research? https://www.anthropic.com/research/vibe-physics
  - AI policies: links provided before and “For Author” section in your journal of choice
  - genAI in scientific writing

Use cases for AI in scientific writing and publishing

- Task: Use the provided material for one option of your choice and work through it (20 min). Explain briefly what it is about (5 min presentation)
  - genAI in scientific writing:
    - Use a publication of your choice (ideally on of yours) and try different things in Apertus and compare results to commercial LLM
    - Take results and ask for an abstract; or take the abstract and ask for a title
      - Try to let the LLM predict manuscript elements and compare prediction to reality. If there are differences, is the LLM wrong or does the real abstract or title not reflect what is in the manuscript?
    - Beyond spell checks: Take a piece of text and ask for
      - Tone of voice
      - Prompt for LLM to take a certain view point (for example different fields for interdisciplinary paper)

Discussion, open questions, etc

Thoughts, opinions, …

- What have we not answered today? What topics are missing?

- Should it have been
  - More technical?
  - More hands on?

Lib4RI – Excellent Services for Excellent Research.

www.lib4ri.ch
info@lib4ri.ch
T: + 41 58 765 57 00

LIB4RI – Excellent Services for Exellent Research.

Lib4RI
Eawag-Empa
Überlandstrasse 133
8600 Dübendorf, Switzerland

Dr. Lothar Nunnenmacher
Head of Lib4RI

T: + 41 58 765 52 21
F: + 41 58 765 50 28
lothar.nunnenmacher@lib4ri.ch

www.lib4ri.ch
info@lib4ri.ch
T: + 41 58 765 57 00