

Race to the Bottom?

TDM-Update aus der Informatik

Sebastian Stober

Berlin

10. Nvember 2025



FACULTY OF
COMPUTER SCIENCE



ins^{ight}
coach & consult

Prof. Dr.-Ing. Sebastian Stober

GenAI Training = TDM?

GenAI Training $\neq$ TDM!

Quellendokumentation

Article 53(1)(d) AI Act.

Providers of general-purpose AI models shall [...] draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

Recital 107 AI Act.

In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form.

Template for the Public Summary of Training Content for General-Purpose AI models

This template is provided by the European Commission and required to be filled in by providers of general-purpose AI models prior to their placing on the Union market in order to comply with their obligation under Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act).

For more information and guidance see Commission's [Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models | Shaping Europe's digital future.](#)

Version of the Summary: Version of the summary, with link(s) to previous versions where applicable

Last update: Click or tap to enter a date.

1. General information

1.1. Provider identification

Provider name and contact details: Replace this with your response...

Authorised representative name and contact details: Only applicable if the provider is established outside the Union (see Article 54 AI Act).

1.3 Modalities, overall training data size and other characteristics

This Section requires general information about the overall training data after pre-processing and before the training of the model.

Modality <i>Select the modalities present in the training data, to the extent that they are identifiable</i>	Training data size <i>For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.</i>	Types of content <i>For each selected modality, provide a general description of the type of content that has been included in the training data.</i>
☐ Text	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillions tokens ☐ More than 10 trillions tokens Alternatively, specify the appropriate unit in a different measurement unit <i>Replace this with your response</i>	Urheberrecht = 1 Token Ur-heber-recht = 3 Token Ur-he-ber-recht = 4 Token U-r-h-e-b-e-r-r-e-c-h-t = 12 Token
☐ Image	☐ Less than 1 million images ☐ 1 Million to 1 billion images ☐ More than 1 billion images	<i>Examples of possible types of content include photography, visual art works, infographics, social media images, logos, brands.</i>

2.3 Data crawled and scraped from online sources

Summary of the most relevant domain names crawled:

In so far as any content from internet domains has been crawled or scraped and used for the training of the model, provide a list of those most relevant internet domains names (top and second-level domain, e.g. "example.com") by listing the top 10 % of all domain names determined by the size of the content scraped (in a representative manner across all modalities where applicable). Small and medium-sized enterprises (SMEs), including start-ups, should disclose top 5% of all domain names or 1 000 internet domain names, whichever is lower. You can provide this list for instance as a downloadable file or provide here information on how to access it.

Beispiel: <https://ai.ovgu.de/Research.html>

Minimallösung einer Quellendokumentation

URLs und Download-Zeitstempel:

- für Daten aus Web-Scraping oder Web-Suche sehr leicht umsetzbar (Metadaten werden ohnehin geloggt)
- Speicherbedarf vernachlässigbar (im Vgl. zum Inhalt)
- Nachteil: für Rechteinhaber nicht immer leicht abzugleichen (unbekannte und potentiell nicht autorisierte Kopien)

Umsetzungsstufen der Quellendokumentation

- Datasheets (sehr grobe textuelle Beschreibung)
- URLs und Download-Zeitstempel
- weitere Metadaten (z.B. IPTC bei Fotos, ID3 bei MP3s)
- Eindeutige Identifikatoren
- Fingerprints

weitere Details:

Sebastian Stober. Möglichkeiten der Quellendokumentation und -auskunft bei generativen KI-Systemen, SSRN, 2025. <https://ssrn.com/abstract=5165182>

Umsetzungsstufen der Quellendokumentation

- Datasheets (sehr grobe textuelle Beschreibung)
- URLs und Download-Zeitstempel
- weitere Metadaten (z.B. IPTC bei Fotos, ID3 bei MP3s)
- Eindeutige Identifikatoren
- Fingerprints

weitere Details:

Sebastian Stober. Möglichkeiten der Quellendokumentation und -auskunft bei generativen KI-Systemen, SSRN, 2025. <https://ssrn.com/abstract=5165182>

Umsetzungsstufen der Quellendokumentation

Eindeutige Identifikatoren (IDs), z.B.

- URNs (Uniform Resource Names)
 - ISBNs ((International Standard Book Numbers)
 - ISSNs (International Standard Serial Numbers)
 - ISWC (International Standard Musical Work Code)
 - ISRC (International Standard Recording Code)
 - ISAN (International Standard Audiovisual Number)
 - DOIs (Digital Object Identifiers)
-
- unabhängig vom Ort oder Zugriffsweg

Vollumfängliche Quellendokumentation

Fingerprints:

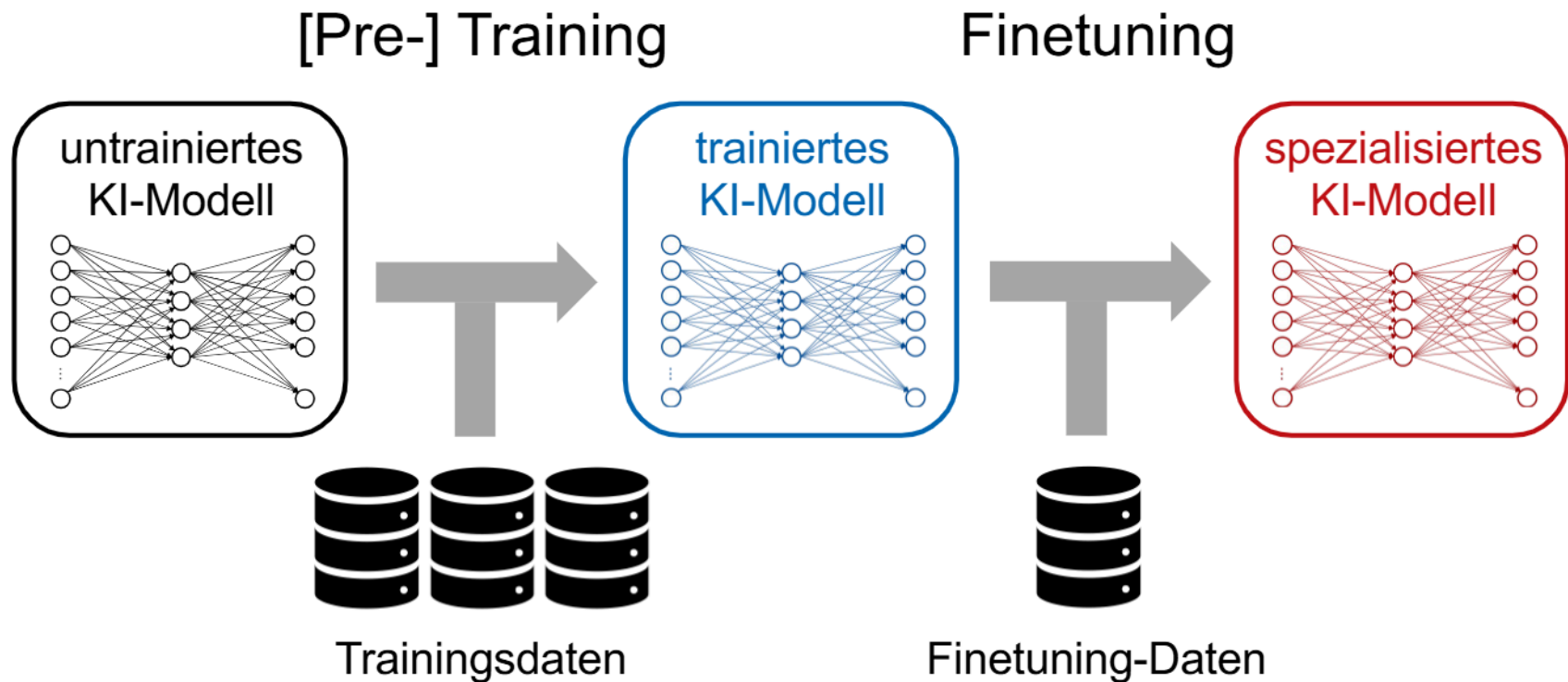
- vom Inhalt abgeleitet
- in der Regel mit eindeutigem Identifikator verknüpft
- abhängig vom Datentyp
- bereits weit verbreitete und robuste Verfahren, z.B.
 - Shazam (seit 2003)
 - Text-Plagiaterkenner
 - Uploadfilter
- auch zur Erkennung authentischer Inhalte geeignet

Verwendung von Quellen

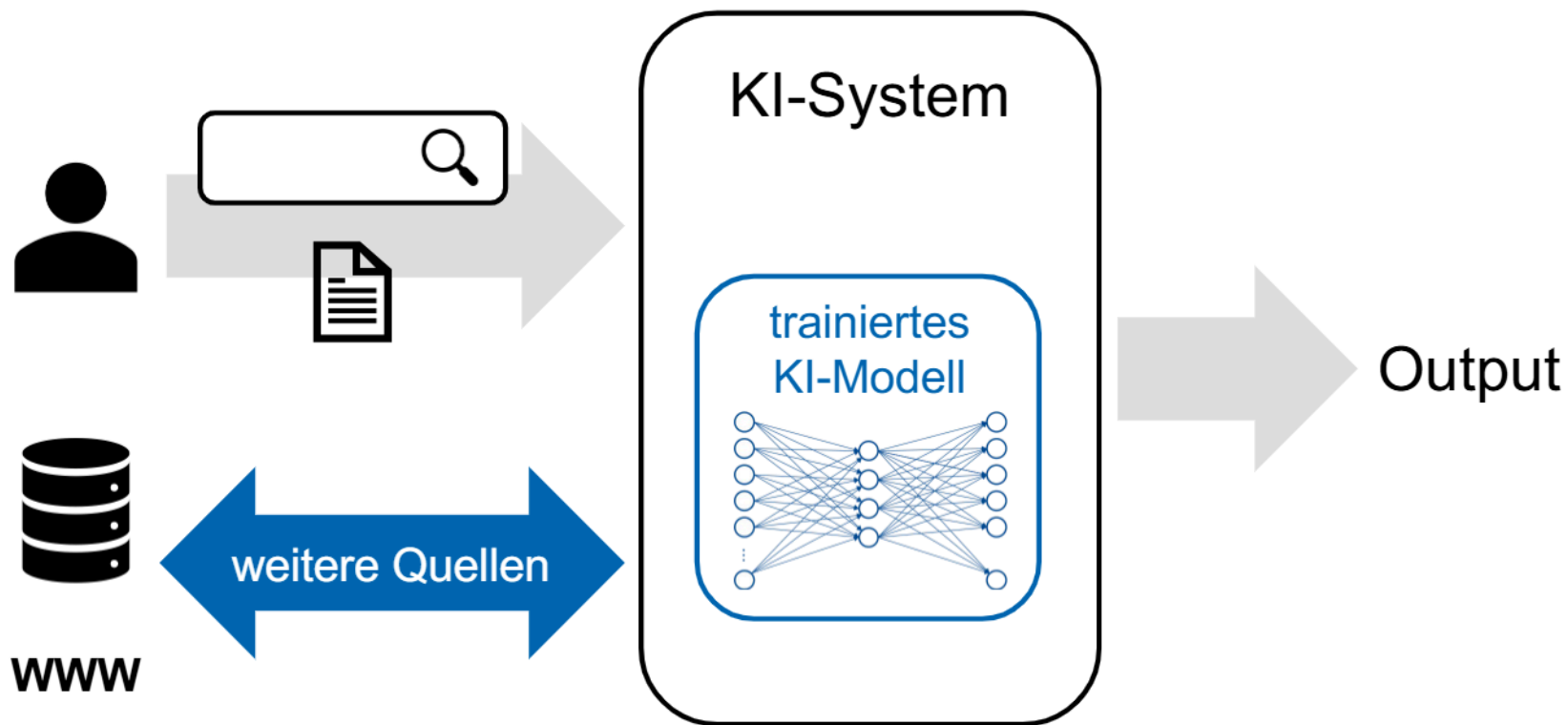
... in generativen KI-Systemen:

1. beim Training von (generativen) KI-Modellen (z.B. LLMs)
2. beim Einsatz
 - als Eingabe durch Benutzer
 - als Suchergebnis aus Datenbank oder Websuche (z.B. Retrieval-Augmented Generation, RAG)

KI Training



KI Einsatz



Mein Fazit zur Quellendokumentation

- Das Template des AI Office ist leider völlig unzureichend.
- “zu schwer”, “zu teuer” oder “zu viel Aufwand” sind nur billige und bequeme Ausreden
- Wer Daten im großen (industriellen) Maßstab für seine Geschäftsmodelle verwenden und verwerten möchte, sollte auch problemlos in der Lage sein, seine Quellendokumentation entsprechend zu skalieren.

Race to the bottom?

Welcome to the 26th International Society
for Music Information Retrieval Conference

ISMIR 2025

Sep 21-25, Daejeon, Korea

ISMIR 2025 Sponsors





UNIVERSAL MUSIC GROUP AND UDIO ANNOUNCE UDIO'S FIRST STRATEGIC AGREEMENTS FOR NEW LICENSED AI MUSIC CREATION PLATFORM

*Companies settle copyright infringement litigation and collaborate on groundbreaking new creative
product suite*

SANTA MONICA, October 29, 2025 – Universal Music Group (UMG), the world leader in music-based entertainment, and Udio, an AI-powered music creation platform, today announced industry-first strategic agreements, under which the companies settled copyright infringement litigation and will collaborate on an innovative, new commercial music creation, consumption and streaming experience.

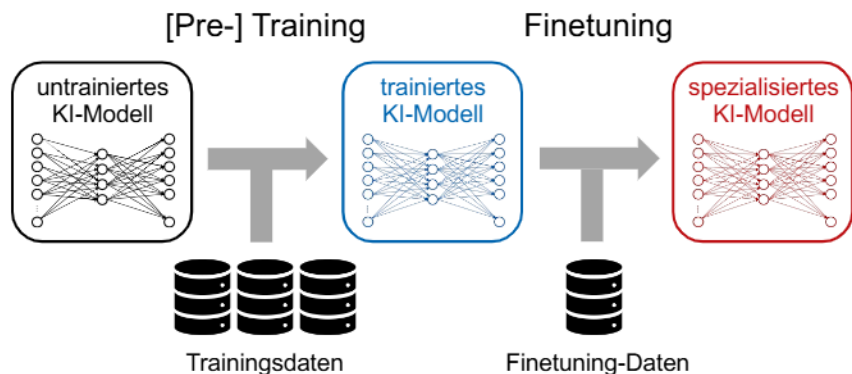
In addition to the compensatory legal settlement, the new license agreements for recorded music and publishing will provide further revenue opportunities for UMG artists and songwriters.

<https://www.universalmusic.com/universal-music-group-and-udio-announce-udios-first-strategic-agreements-for-new-licensed-ai-music-creation-platform/>

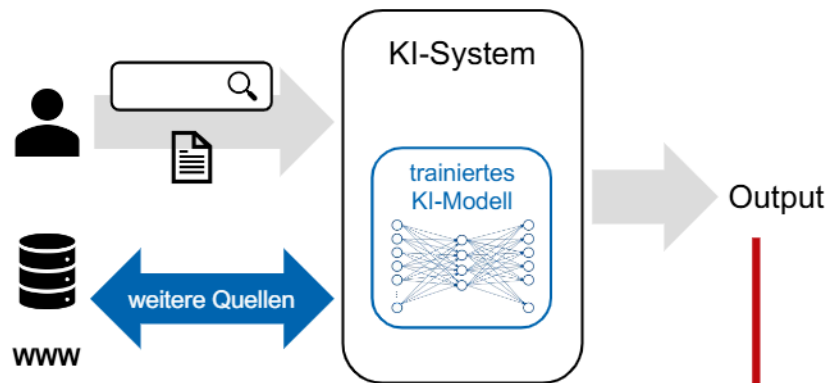
Source Attribution

... zur Verteilung von Ausschüttungen

Training



Einsatz (Inferenz)



Welche Quellen tragen zum Output bei und wie hoch ist der jeweilige Anteil?

Bonus:

Weitere Neuigkeiten aus der Forschung

Weitere Neuigkeiten aus der Forschung (1)

Deezer:

- > 30% der wöchentlichen Uploads sind KI-Müll
- Verfahren zur Erkennung der Signaturen von verschiedenen GenAI-Modellen:
D. Afchar, G. Meseguer-Brocal, K. Akesbi & R. Hennequin. *“A Fourier Explanation of AI-music Artifacts”*



ISMIR 2025 Best Paper Award

Weitere Neuigkeiten aus der Forschung (2)

LLM sind gar nicht so deterministisch wie erwartet!

Question:

"Let $A, B, C,$ and D be point on the hyperbola:
Find the greatest real number that is less than BD^2 for all such rhombi."



Greedy, Seed=42, BS=32, #GPU=4

Okay, so I have this problem ... perpendicular, but in a square,
... for all such rhombi is $\boxed{480}$.



BF16



Okay, so I have this problem ... perpendicular. Wait, no, hold on,
... for all rhombi is 960



Greedy, Seed=42, BS=8, #GPU=4

Hintergrund: begrenzte Genauigkeit führt z.B. zu $(a+b)+c \neq a+(b+c)$

Vielen Dank!

Kontakt: stober@ovgu.de

Webseite: ai.ovgu.de | insaight.de