Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS

06.03.2026 Symposion des IUM und EMR zum Thema "KI-Nutzung am Beispiel von Audio - Ein Blick in die Praxis"

Herausforderungen und Grenzen generativer KI für Audio und Musik

Hanna Lukashevich - hanna.lukashevich@idmt.fraunhofer.de

Christian Dittmar - christian.dittmar@iis.fraunhofer.de

Übersicht

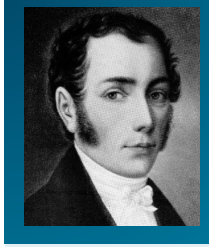
1. Einleitung
2. Generative KI für Audio: Sprachsynthese
3. Generative KI für Audio: Musiksynthese

Kapitel 1

Einleitung

Die Fraunhofer-Gesellschaft

Joseph von Fraunhofer (1787–1826) – Unser Namensgeber



Joseph von Fraunhofer

Entdecker der Fraunhofer-Linien im Sonnenspektrum

Neue Methoden zum Schleifen von Linsen

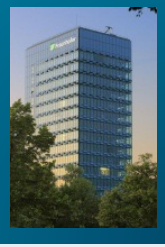
Direktor und Teilhaber einer Glasfabrik

Forscher

Erfinder

Unternehmer

Die Fraunhofer-Gesellschaft



Forschung und Entwicklung im Auftrag von Industrie und Regierung

MP3-Musikformat, weiße LEDs, hochauflösende Wärmebildkamera

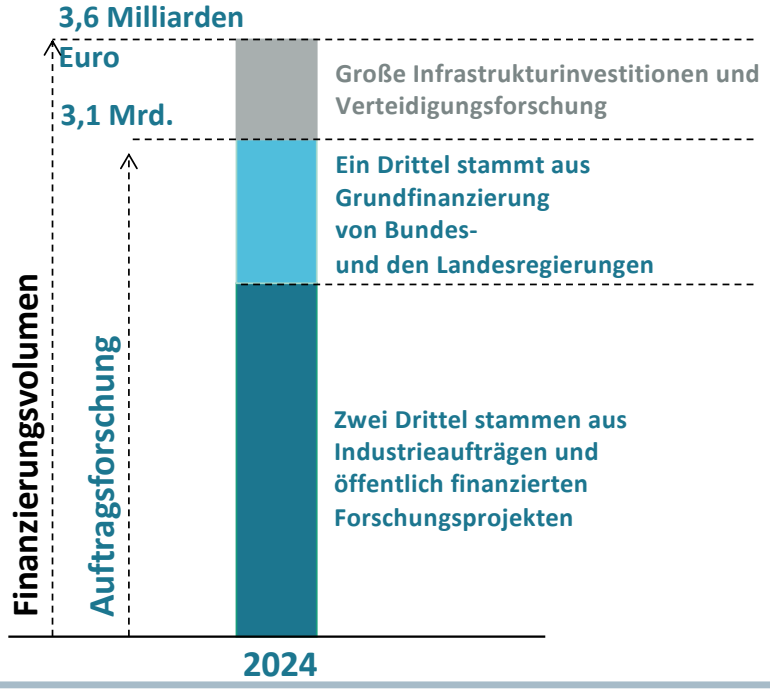
Forschungsbudget:
ca. 2,9 Milliarden Euro jährlich

Die Fraunhofer-Gesellschaft – Auf einen Blick

Angewandte Forschung mit Schwerpunkt auf zukunftsrelevanten Schlüsseltechnologien und der Kommerzialisierung von Forschungsergebnissen in Wirtschaft und Industrie. Vorreiter und Trendsetter bei innovativen Entwicklungen.

> 30.000 Mitarbeiter

75 Institute und Forschungseinrichtungen



Fraunhofer IDMT
Fraunhofer IIS

Fraunhofer-Institut für Digitale Medientechnologie IDMT

- Hauptsitz auf dem Campus der Technischen Universität Ilmenau
- Gegründet im Januar 2004
- Direktor: Prof. Dr.-Ing. Joachim Bös, M.S./SUNY
- Insgesamt 200 Mitarbeiter, Auszubildende, Doktoranden, Praktikanten, wissenschaftliche und studentische Hilfskräfte

Signalanalyse und vertrauenswürdige KI mit Schwerpunkt auf Akustik und Audio

- Industrielle Geräuschanalyse
- Umgebungsgeräuschanalyse
- Audio- und visuelle Inhaltsanalyse
- Medienforensik
- Datenschutz und vertrauenswürdige KI
- Kontrollierte Schallfelder
- Akustische Simulation für die Validierung und das Training von KI
- Intelligente akustische Sensoren

Fraunhofer IDMT Fraunhofer IIS

Über mich

Hanna Lukashevich

Forschungsinteressen

- Musikanalyse (Ähnlichkeit, Tagging, Klassifizierung, Transkription)
- **GenAI für Musik, einschließlich Erkennung und Zuordnung**
- Audiosignalverarbeitung und maschinelles Lernen
  - Unsicherheitsschätzung im Deep Learning

@ Fraunhofer IDMT

- Seit 2006 im Institut, am Hauptsitz in Ilmenau
- Leiterin der Forschungsgruppe Semantic Media Technologies
- Tägliche Aufgaben
  - Akquisition/Management von F&E-Projekten und Personal
- Team
  - Senior Scientist, PostDocs, Doktoranden, wissenschaftliche Mitarbeitende
  - 5–10 Studierende

Fraunhofer IDMT Fraunhofer IIS

Einleitung

Fraunhofer-Institut für Integrierte Schaltungen IIS

- Das größte Institut der Fraunhofer-Gesellschaft
- Gegründet 1985
- Standorte in 12 Städten
- Hauptsitz in Erlangen
- Über 1200 Mitarbeiter und 430 studentische Hilfskräfte
- „Audio- und Medien" als zentraler Schwerpunkt (mp3, AAC, MPEG-H…).
- Konsequente Innovationen, beispielsweise durch die Integration künstlicher Intelligenz.

Über mich

Christian Dittmar

- Studium der Elektrotechnik in Jena

- Fraunhofer IDMT (2003–2014)
  - Leiter Semantische Musiktechnologien

- Spin-off-Unternehmen Songquito (2012 – heute)
  - www.songs2see.com

- AudioLabs @ FAU (2014–2018)
  - Promotion in der Gruppe von Prof. Meinard Müller
  - Dissertation: „ Source separation and restoration of drum sounds in music recordings”
  - Dissertationspreis der Staedtler-Stiftung 2019

- Fraunhofer IIS (2018 – heute)
  - Leiter Spoken Language Processing
  - Forschungskoordinator der AudioLabs
  - Allinga TTS Cloud Service: https://allinga.lze-innovation.de/

Grenzen generativer KI, über die wir (heute) nicht reden …

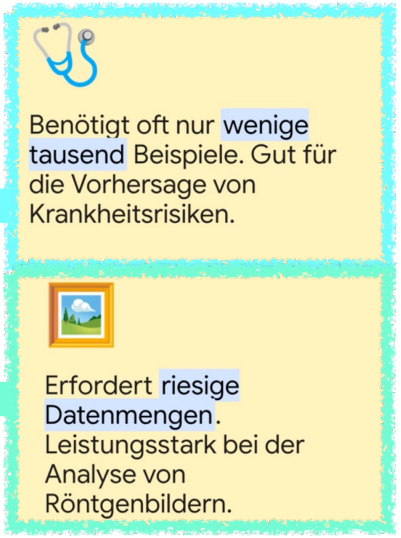
- *Fehleranfälligkeit & Halluzinationen im Allgemeinen*
- *Ressourcenbedarf, speziell Energie & Wasser*
- *Desinformation, Deep-Fakes, Verantwortung & Haftung*
- *Fehlanzeige bzgl. kombinatorischer Kreativität & Verständnis*
- *Dominanz der BigTech Firmen des Silicon Valley*

Ab-Grenzen: (Allgemeine) KI, maschinelles Lernen, weitere Buzzwords …

KI - jede Technik, die Computern ermöglicht, menschliche Intelligenz nachzuahmen

ML - Maschinelles Lernen sind Algorithmen, die Muster (in meist strukturierten Daten) lernen und deren Leistungen sich mit zunehmender Erfahrung verbessern

DL - Deep Learning ist eine Teilmenge des ML, bei der mehrschichtige, neuronale Netzwerke aus riesigen (unstrukturierten) Datenmengen lernen



Large Language Models (LLMs)

= nicht-denkende, nicht-verstehende „Wortvorhersage-Maschinen“
(Wahrscheinlichkeitsrechner)

Generative KI

Ein Teilbereich der KI, der völlig neue Inhalte wie Texte, Bilder oder Musik erzeugen kann.

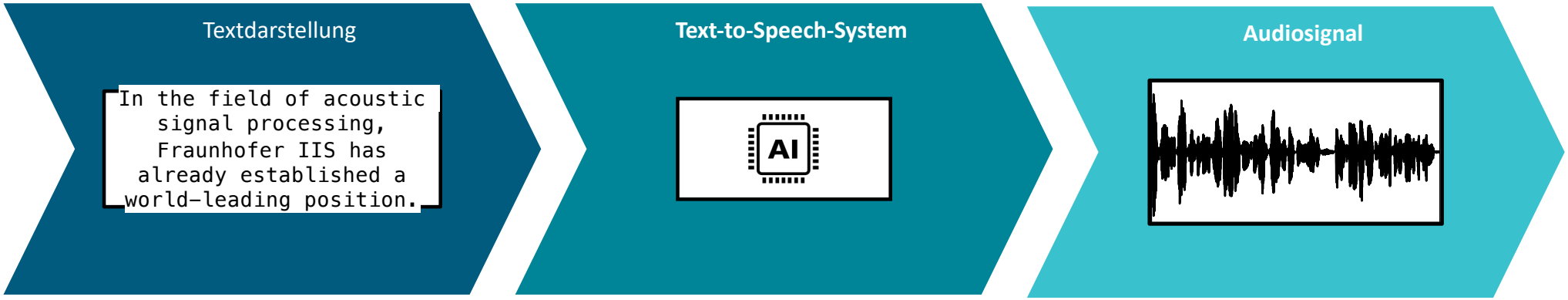
GAN: Generator und Diskriminator trainieren sich gegenseitig in einem Wettbewerb.

Autoregressiv: Berechnet Wahrscheinlichkeiten für das nächste Wort. Basis für ChatGPT.

Diffusion: Approximation von hochdimensionalen, komplexen Wahrscheinlichkeitsdichteverteilungen. Basis für Dall-E.

Generative KI für Audio

Sprachsynthese



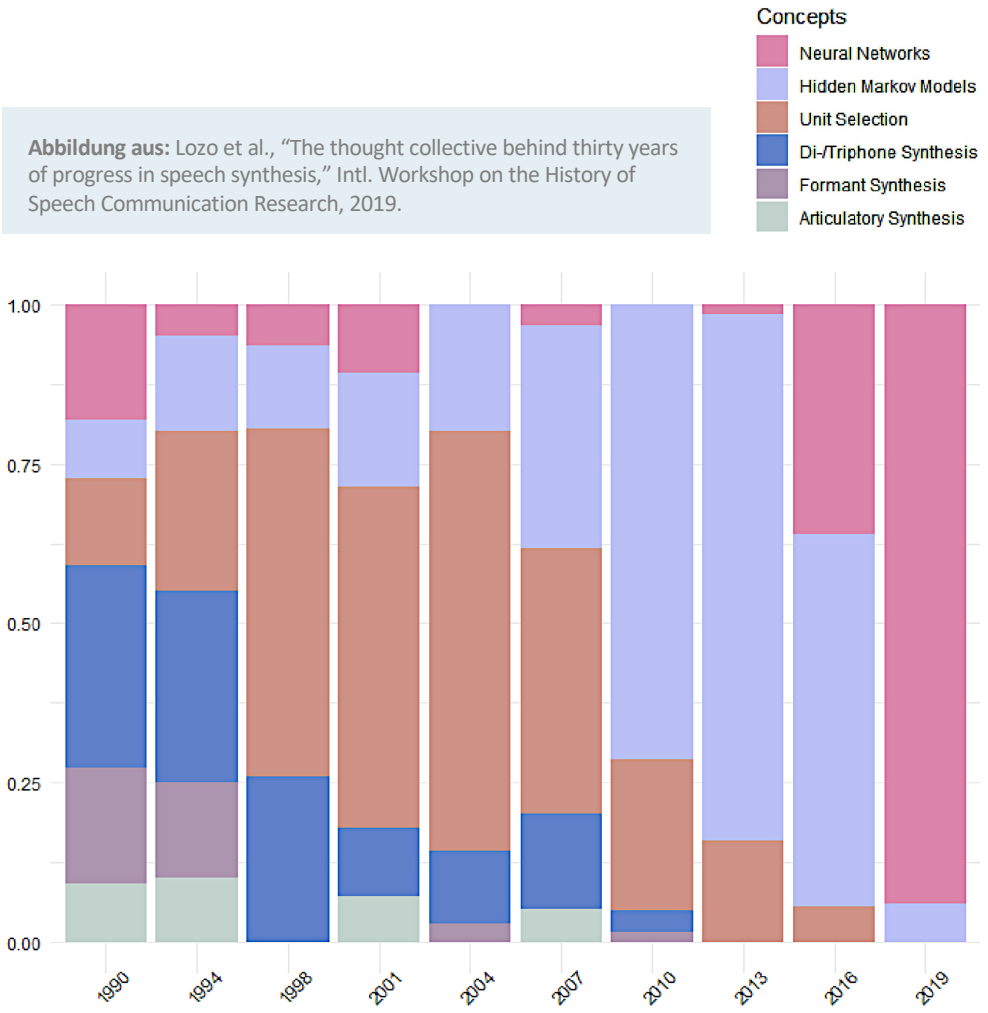
- Text-to-Speech-Synthese (TTS) wandelt Text in ein anhörbares Sprachsignal um
- Grundlegende Anforderungen:
  - Die synthetisierte Sprache muss verständlich sein
  - Die synthetisierte Sprache sollte natürlich klingen

Generative AI für Audio

Kurze Geschichte der Sprachsynthese

TTS-Methode	Wann	Kategorie
Formant Synthesis	1940s–1980s	Physical modeling
Articulatory Synthesis	1980s–today	
Di-/Triphone Synthesis	1980s–1990s	Concatenative Synthesis
Unit Selection	1990s–2000s	
Hidden Markov Models (SPSS)	2000–2015	Machine Learning
Neural Networks (Deep Learning)	2016–2023	
Generative AI (LLM / Diffusion)	2023–today	

Abbildung aus: Lozo et al., "The thought collective behind thirty years of progress in speech synthesis," Intl. Workshop on the History of Speech Communication Research, 2019.



Fraunhofer IDMT Fraunhofer IIS

Generative KI für Audio

Fähigkeiten moderner TTS-Systeme

- Moderne, KI-basierte TTS-Systeme können:
  - natürliche Sprache erzeugen und **menschliche Zuhörer täuschen**
  - Echte Stimmen klonen (**instant voice cloning, professional voice cloning**)
  - **Prosodie, Betonung und Emotionen** entsprechend dem Textinhalt anpassen
  - Sprachaufnahmen in die Stimmen anderer Personen umwandeln (**Speech-to-Speech, Voice Conversion**)
  - **Mehrsprachige** Stimmen in vielen verschiedenen Sprachen generieren, nahtloser Sprachwechsel innerhalb eines Satzes

Die **Human Fooling Rate (HFR)** ist definiert als der Prozentsatz der Fälle, in denen synthetische Sprache in einem **binary forced choice** Hörtest mit menschlicher Sprache verwechselt wird.

TTS System	Typ	HFR	MUSHRA
PlayHT	**Kommerziell**	**72**	**85**
Human	**Referenz**	**70**	**74**
ElevenLabs	**Kommerziell**	**69**	**80**
F5-TTS	Open Source	50	70
GPT-SoVITS	Open Source	44	68
XTTS	Open Source	41	76
StyleTTS2	Open Source	38	71
VoiceCraft	Open Source	30	49

Tabelle aus: Varadhan et al., "The State Of TTS: A Case Study with Human Fooling Rates," Interspeech, 2025.

Generative AI für Audio

Limitations of modern TTS systems

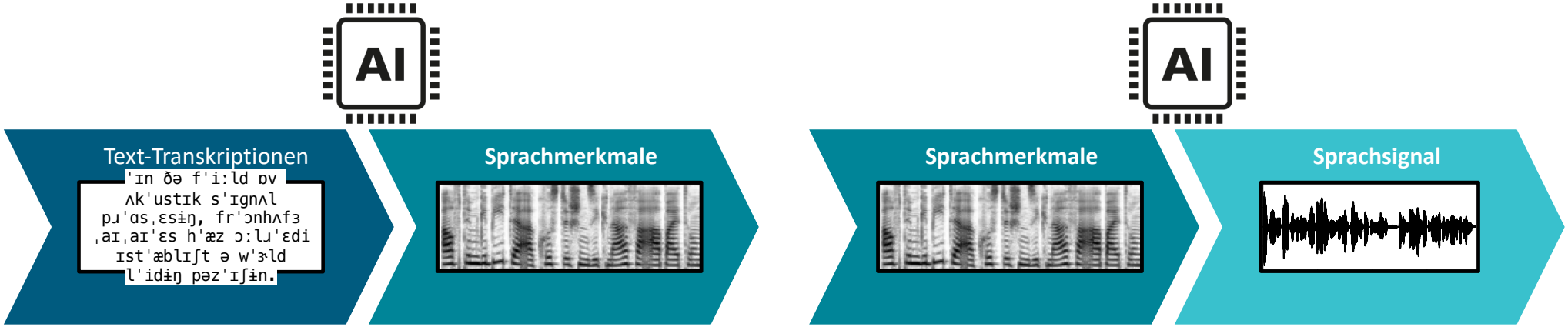
- Moderne, KI-basierte TTS-Systeme sind (noch) nicht in der Lage:
  - Die Art und Weise, wie ein Mensch spricht, in jeder Hinsicht nachzuahmen (z. B. Akzent, Spontanität, Paralinguistik)
  - Die Prosodie bei der Speech-to-Speech Voice Conversion perfekt übertragen
  - In Echtzeit, mit Wort-für-Wort Texteingabe ausgeführt werden
  - Auf kleinen Hardwaregeräten (z.B. Smartwatch) ohne Cloud-Verbindung ausgeführt zu werden
  - Andere KI-Systeme zu täuschen, die auf die Erkennung synthetischer Sprache spezialisiert sind

Voice Cloning		
Human reference recoding	ElevenLabs Instant Voice Clone 3 min training data	Allinga Text-to-Speech Professional Voice Clone 7.5 hrs training data
		
	Qwen3-TTS (Open Source) 20 sec training data	CosyVoice3 (Open Source) 20 sec training data
		

Voice Conversion		
Human source recording	ElevenLabs Voice Changer	Allinga Prosody-aware Voice Conversion
	 	 

Generative KI für Audio

Allgemeines Prinzip für KI-basierte TTS



Akustisches Modell / Großes Sprachmodell

- Umwandlung von Text in Sprachmerkmale/Tokens
- Zuordnung von Textsymbolen zu kurzen Lauteinheiten
- Nichtlineare zeitliche Alignierung
- Vorhersage von Prosodie und Betonung

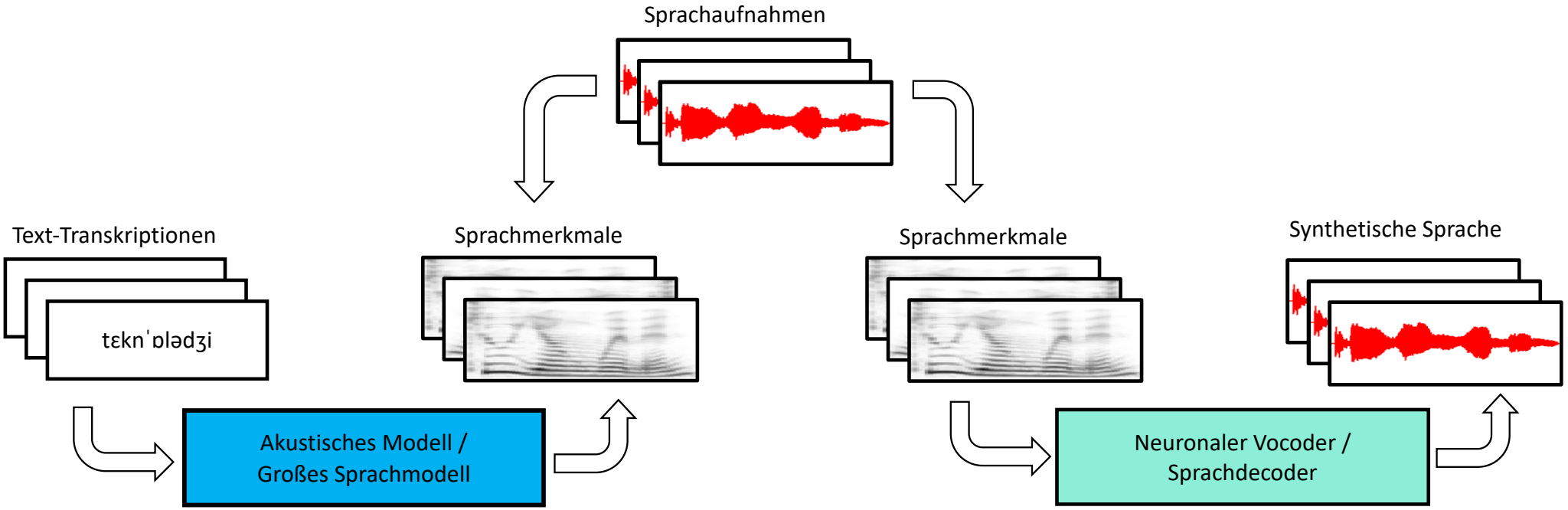
Neuronaler Vocoder / Decoder

- Umwandlung von Sprachmerkmalen/Tokens in hörbare Sprache
- Rekonstruktion eines hochqualitativen Audiosignals aus stark komprimierter Datenströmen

Fraunhofer IDMT Fraunhofer IIS

Generative KI für Audio

Erforderliche Daten für das Training von KI-basierter TTS



General Stage (S1): During the initial pre-training phase, we leverage over 5 million hours of multilingual speech data to train Qwen3-TTS. This stage establishes a monotonic mapping from multilingual text representations to speech and builds general capabilities for Qwen3-TTS.

Hu et al., „Qwen3-TTS Technical Report“, ArXiV, 2026.

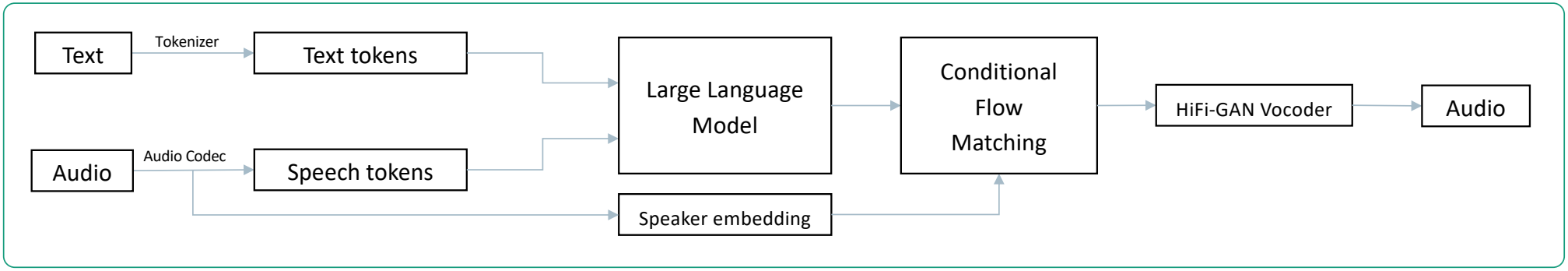
Fraunhofer IDMT Fraunhofer IIS

CosyVoice3 Text-to-Speech

Beispiel für ein aktuelles generatives TTS-System (Open Source)

Hauptfunktionen

- **LLM-gesteuerte TTS**: Sprachliche Struktur, Phonem-Alignierung und Sprach-Rhythmus
- **Conditional Flow Matching**: Voice Cloning, Sprachstil und Emotionen
- **Zero-Shot Voice Cloning**: Lernt die Klangfarbe einer Stimme aus einer kurzen Referenz-Audioaufnahme (wenige Sekunden)



Xiang et al., „Build LLM-Based Zero-Shot Streaming TTS System with Cosyvoice“, IEEE International Conference on Acoustics, Speech and Signal Processing, 2025.

Allinga Text-to-Speech

Systemübersicht

Predicts language-specific, optimal pronunciation

- DeepPhonemizer
- Pronunciation Dictionaries

17 MB
#params
787,968

Predicts speech features

- ForwardTacotron
- Conformer

55 MB
#params
28,233,449

Enhances naturalness, lowers data requirements

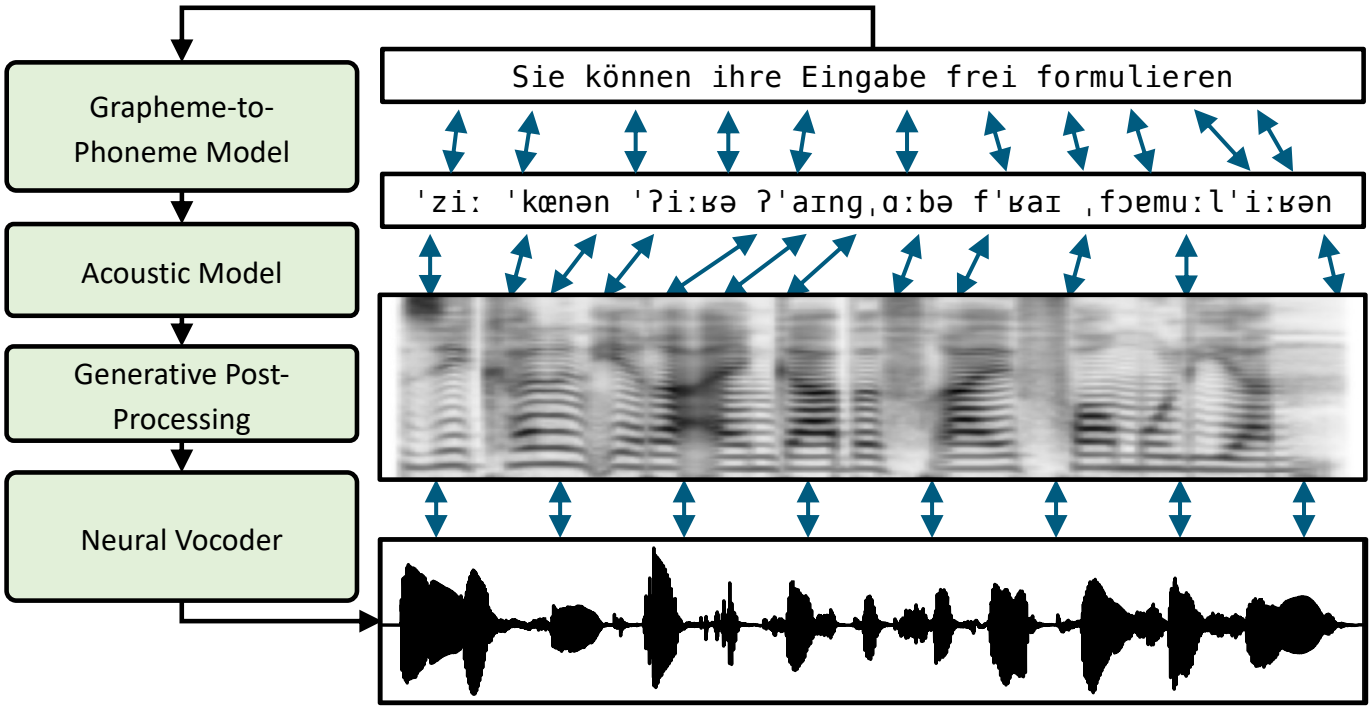
- PostProCFM
- PostProGAN

42 MB
#params
21,872,976

Reconstructs audio signal from speech features

- StyleMelGAN
- BigVGAN2
- Vocos

9 MB
#params
3,852,992



Allinga Voice Conversion

Systemübersicht

Phoneme Activations, Prosody Features
- Wav2Vec2
- TorchCrepe
- DSP-based Features

1.5 GB
#params 315,536,095

Predicts speech features
- PAD-VC

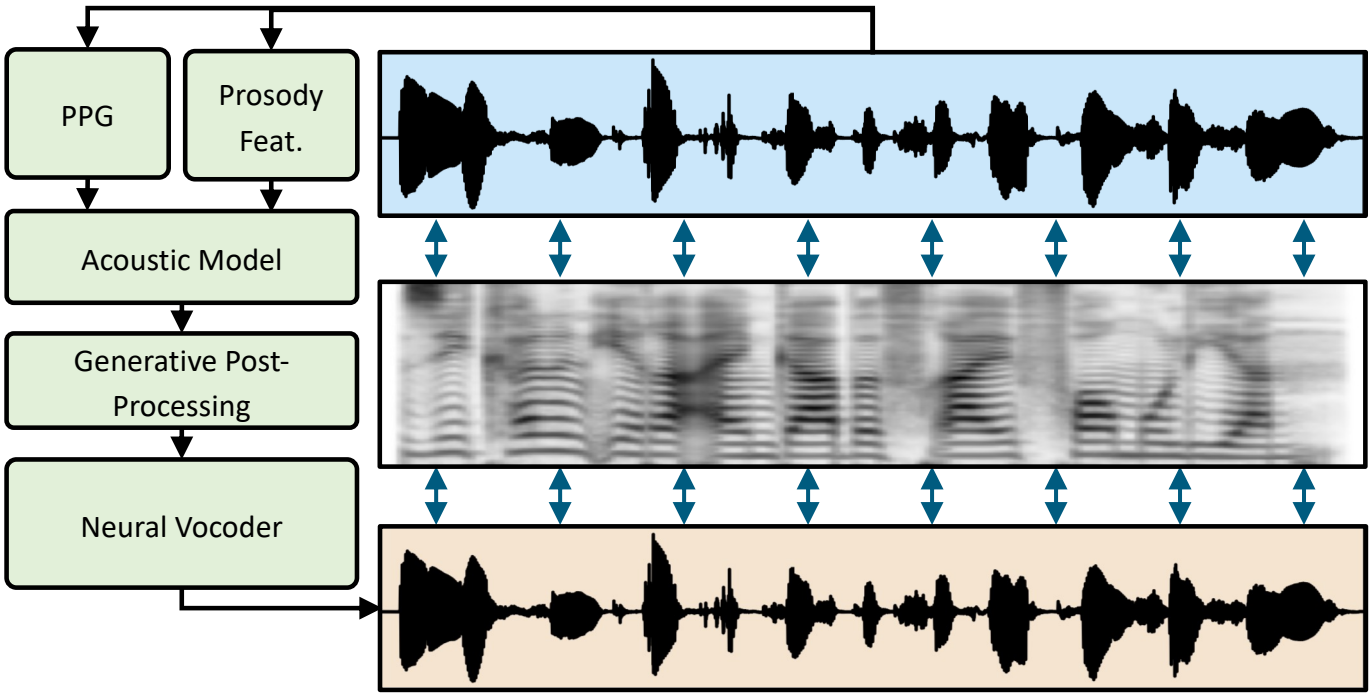
14 MB
#params 7,293,296

Enhances naturalness, lowers data requirements
- PostProCFM
- PostProGAN

42 MB
#params 21,872,976

Reconstructs audio signal from speech features
- StyleMelGAN
- BigVGAN2
- Vocos

9 MB
#params 3,852,992



„Disentanglement“ der Sprache für TTS

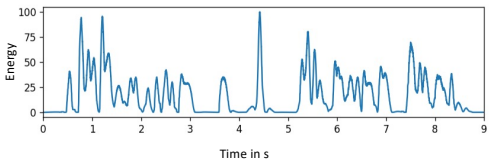
Fähigkeit, verschiedene Aspekte der Sprache unabhängig voneinander zu verarbeiten

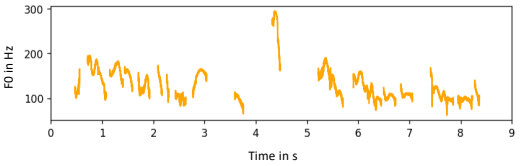
- Es stellen sich interessante rechtliche Fragen:
  - Können Prosodie und Sprachstil urheberrechtlich geschützt werden?
  - Werden die Persönlichkeitsrechte verletzt, wenn die geklonte Stimme nicht wiedererkennbar ist?
  - Wie verhält es sich mit Urheberrechten an Texten, wenn keine explizite textuelle Darstellung verwendet wird?

Sprachlicher Inhalt:

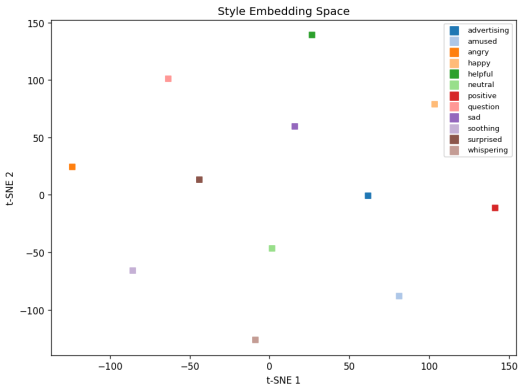
mɪt ˈɔfənɜ dˈoʊzənt ʃlˈaɪçt ˈzɪç dɛ ˈalbrt ˈɔn ˈʔʊnt ʃvˈʊp, hˈaʊldə dɛs kfˈeɪs aʊf ən bˈamʃtɔm dˈaʊbodɛ ʃvˈarmzˈɪtst.

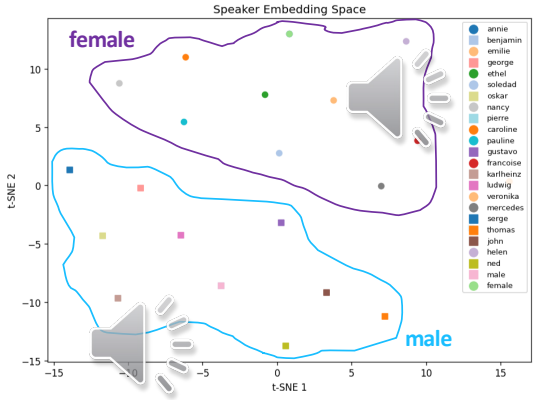
Prosodie-Merkmale:





Sprachstil, Klangfarbe:





Ganz aktuelle Grenzen

Rechtsgutachten: Klauseln sind "rechtswidrig"

Streit um Netflix-Vertrag für Synchronsprecher:innen

Jetzt teilen

Im Streit um Netflix-Vertrag für Synchronsprecher:innen hat der Verband Deutscher Sprecher:innen (VDS) ein Rechtsgutachten vorgelegt.

Es geht um die AOR-Vereinbarung (Assignment of Rights Agreement), die Netflix über Synchronstudios allen deutschsprachigen Synchronsprecher:innen zur Unterzeichnung vorlegt. Das Gutachten der Rechtsanwaltssozietät Spirit Legal kommt zu dem Ergebnis: „Zentrale Klauseln des Vertrags sind unwirksam oder rechtswidrig“, heißt es in einer VDS-Pressemitteilung vom 9. Februar 2026. Von einer Unterzeichnung sei abzuraten.

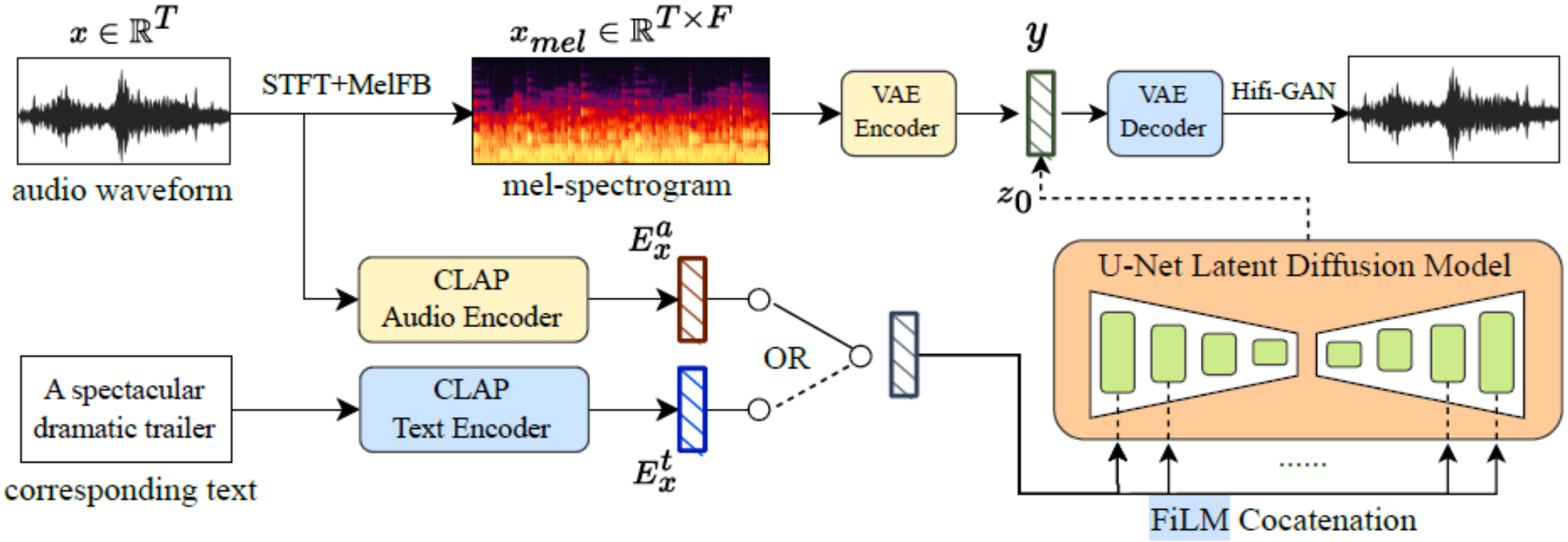
Grenzen des Einfachen

Grenzen durch KI-generierte Trainingsdaten

KI: Kollaps droht wegen KI-generierter Trainingsdaten | heise online

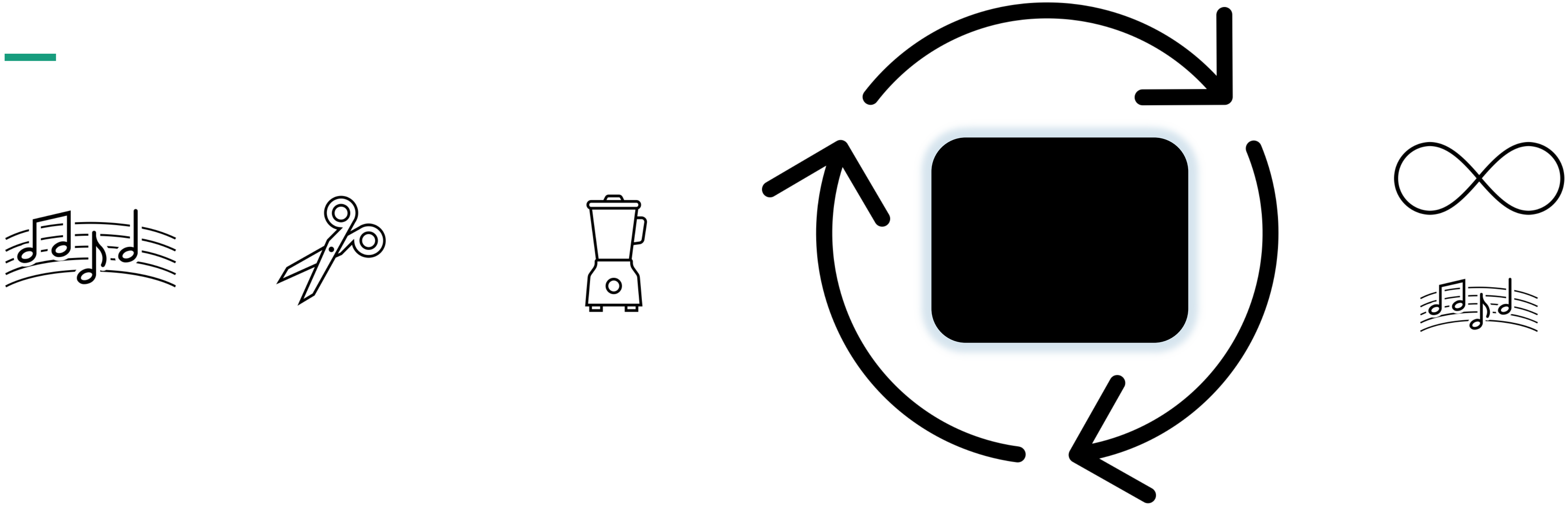
KI-Modelle sabotieren sich selbst laut Studien, indem sie zunehmend oder ausschließlich KI-generierte Daten für das Training verwenden, was dazu führt, dass sie immer mehr Müll produzieren.

Ohne Grenzen …



Source: https://arxiv.org/pdf/2308.01546

Grenzen des Trainingsmaterials



1 Millionen Songs

in (z.B. 20-100) Teile zerlegt *

zufällig in Gruppen kombiniert

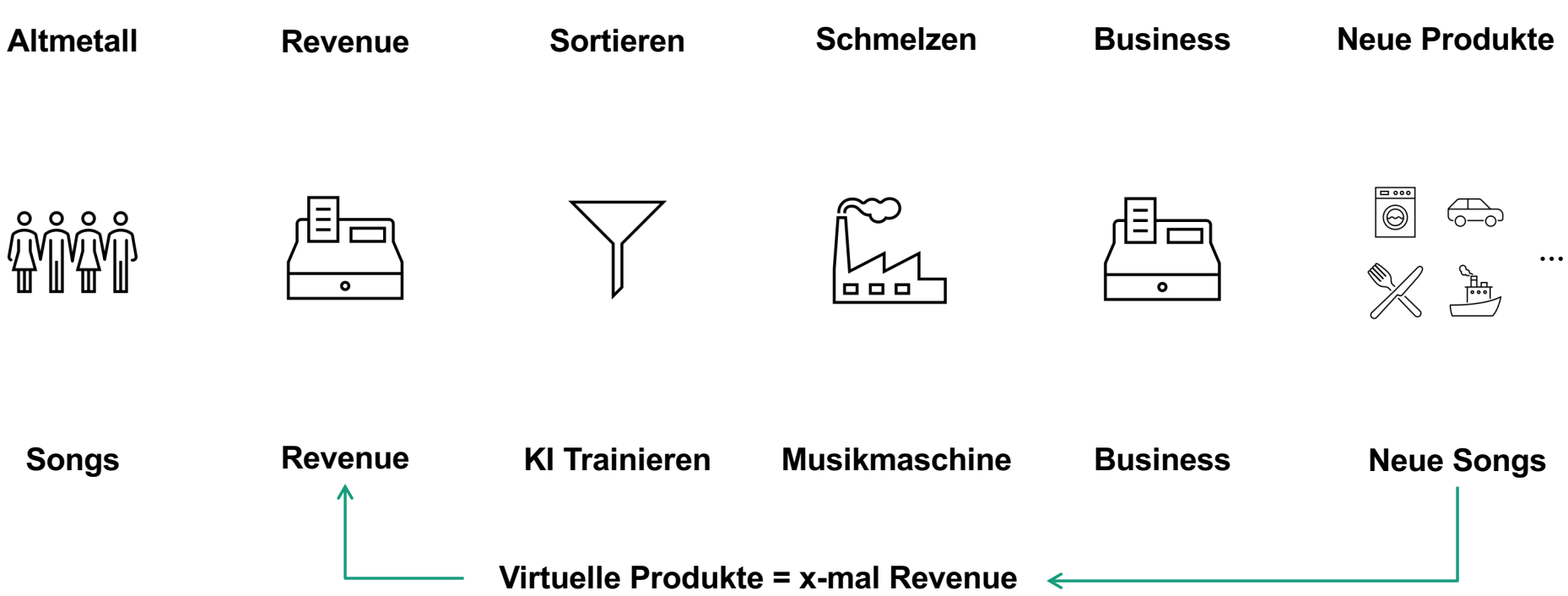
Millionen Male durch verschiedene Schichten in mehreren neuronalen Netzen iteriert

um eine unbegrenzte Anzahl neuer Musikstücke auszugeben

** ap. 43.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000 variations*

Bilder: Dall-E 3

Grenzen im Training: Vom Wertstoffhof zu Messer, Gabel, Löffel …



Fraunhofer IDMT
Fraunhofer IIS

Grenzen der Zuordnung des KI-generierten Output

Nur Versprechen
ohne Nachweis, bisher …

Nur Theorie: Wenn das **Trainingsmaterial** bekannt ist **UND** das **GenAI-Modell** einschließlich aller Gewichte usw. **bekannt** ist, besteht die **Möglichkeit**, eine **Verbindung** zwischen der generierten **KI-Ausgabe** und der Wirkung bestimmter Teile des **Trainingsmaterials** herzustellen ... allerdings hängt dies noch von vielen (unbekannten) Anforderungen ab. *

** TRAK: Attributing Model Behavior at Scale - https://arxiv.org/abs/2303.14186*

Grenzen des Einflusses

Bild: Dall-E 3

Auch für unser Hirn ist es unmöglich festzustellen, welche (wann/wie/wo) gehörten Lieder unseren Musikgeschmack geprägt haben.

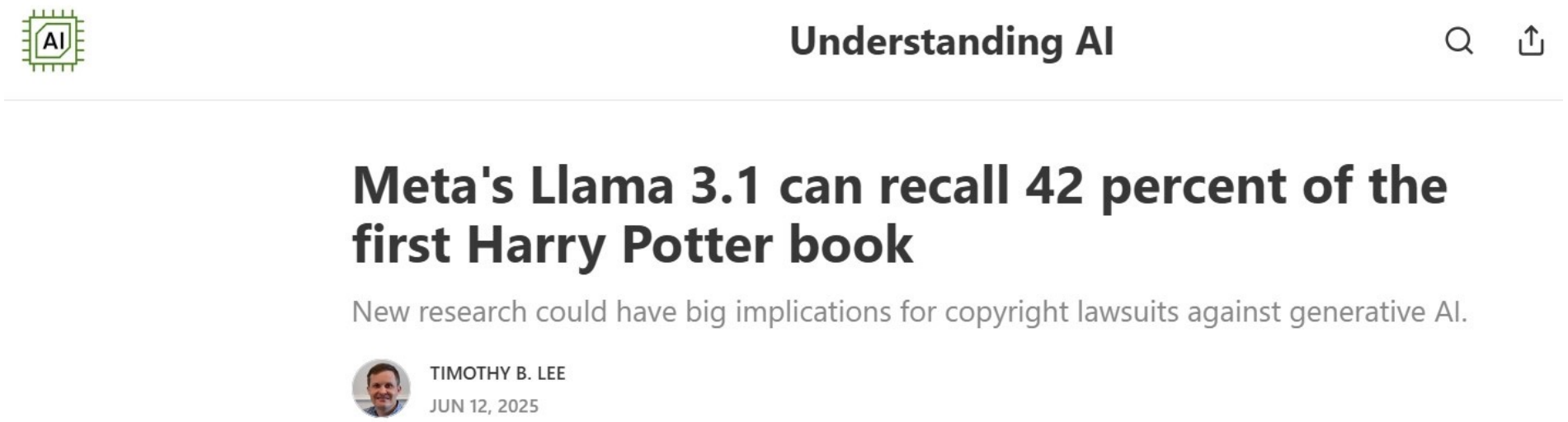
Und hat das Kinderlied vom Kindergarten inzwischen seine Wirkung verloren?

Grenzen der Ähnlichkeit – sie ist KEIN Herkunftsnachweis

Aber keine Kennzeichnung wie Watermark o.ä. „überlebt“ den Trainingsprozess

So ist z.B. ein Musikstück immer irgendwie ähnlich zu allen anderen Musikstücken auf einer Skala zwischen 0 und 100 in %.

Grenzen der Auswendiglernens



Nur möglich bei Open-Source-Modellen, deren Gewichte veröffentlicht werden – wobei zu vermuten ist, **dass Modelle in Zukunft überhaupt keine Gewichte mehr veröffentlichen** werden, wenn sie dadurch anfällig für den Nachweis von Urheberrechtsverletzungen werden.

https://www.understandingai.org/p/metas-llama-31-can-recall-42-percent

Investment Grenzen

Ein paar dutzend Millionen für Ethik & Herausforderungen an Forschung & Startups

Source: Dall-E 3

Unzählige Milliarden Investitionen in generative KI & multimodale LLMs

Erkenntnisgrenzen - kognitive Auswirkungen durch LLM-Nutzung

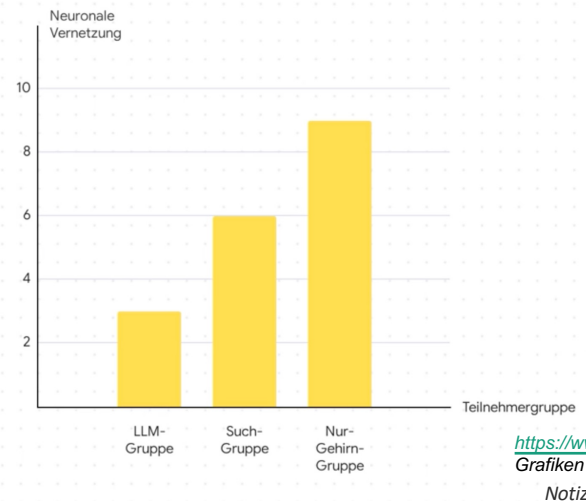
LLM-Gruppe
Nutzt GPT-4o zum Schreiben von Essays.

→

Such-Gruppe
Durfte nur traditionelle Suchmaschinen verwenden.

→

Nur-Gehirn-Gruppe
Musste Essays ohne jegliche Hilfsmittel verfassen.



Je mehr externe Hilfe, desto weniger vernetzt Arbeit das Gehirn (EEG).

https://www.media.mit.edu/publications/your-brain-on-chatgpt/
Grafiken mit Hilfe von Notebook LM (Google)

Notiz der Autoren: begrenzte Teilnehmerzahl und Beschränkung auf ChatGPT.

>80% der LLM-Gruppe = keine Erinnerung an einen Text.

Kognitive Schulden

Ausgelagerte Denkarbeit ersetzt Prozesse für unabhängiges Denken und beeinträchtigt kritisches Denken.

⏱️ Kurzfristiger Gewinn: Schnellere Ergebnisse, weniger Aufwand.

💸 Langfristige Kosten: Schwächeres Gedächtnis, weniger Kreativität.

Musikquiz: Mensch oder Machine?

#1		Suno, Pop
#2	Lukas Graham - Love Someone (real country)	
#3	Mona Lisa - You Said (real rock/pop)	
#4		Suno, Rock
#5		Suno, Country
#6	Mandy Moore - Crush (real pop)	

Grenzen des Publikums

Deezer: **täglich** werden ca. **60.000 vollständig KI-generierte Titel** hochgeladen
= 39 % aller täglich angelieferten Songs

https://newsroom-deezer.com/2025/11/deezer-ipsos-survey-ai-music/

97 % konnten in einem **Blindtest** mit zwei KI-Songs und einem echten Song **keinen Unterschied** zwischen vollständig KI-generierter Musik und von Menschen gemachter Musik **feststellen**

52 % empfanden es als **unangenehm**, dass sie den Unterschied **nicht feststellen konnten**


DEEZER NEWSROOM

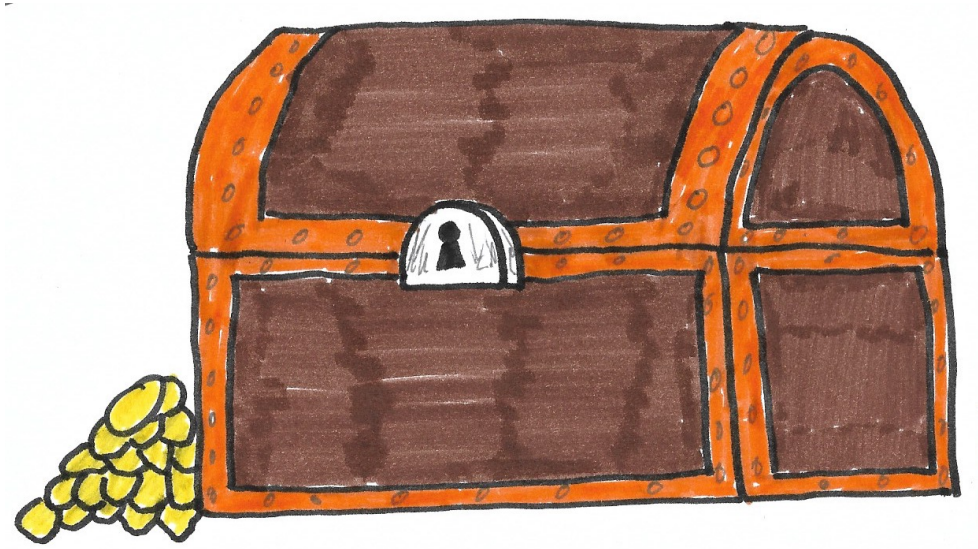
Deezer/Ipsos survey: 97% of people can't tell the difference between fully AI-generated and human made music – clear desire for transparency and fairness for artists

Deezer - Nov 12, 2025 • 9 min read

Gegen die Grenzen: Was kann generative KI für Kreative bewirken?

KI, die Routinearbeiten durch Nachahmung übernimmt, ist eine gute Sache.

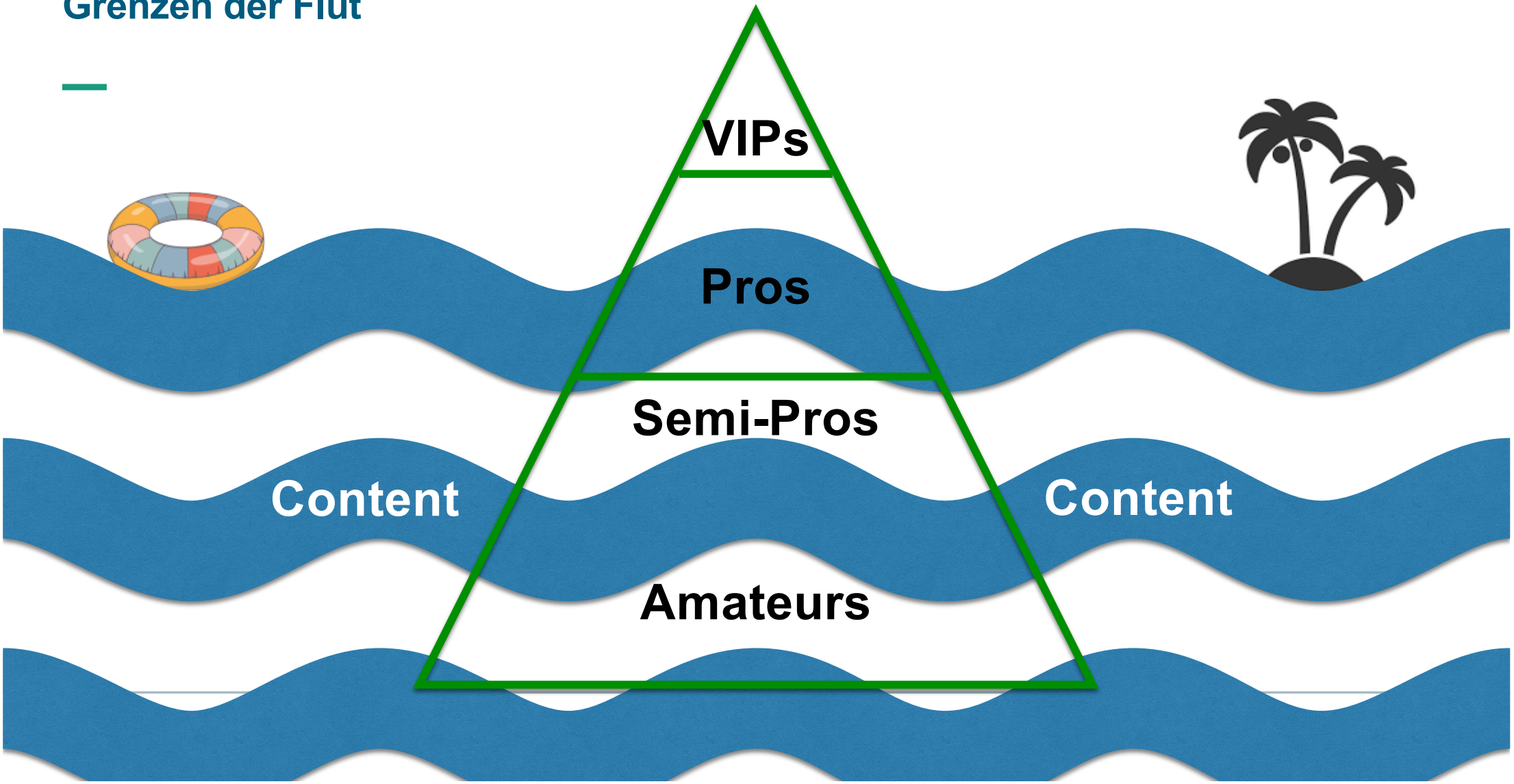
“Be unique, create something worth imitating, something the machine doesn’t know yet.” *



How AI might make a lot of musicians irrelevant

(The Venus Theory YouTube Channel)

Grenzen der Flut



Erkenntnis gegen Grenzen

“It is always people who decide how and what data is collected and what it means.”

(… and how it is going to be used)

2023 Meredith Whittaker, Präsidentin Signal

"Our risk is not the advent of superintelligent computers, but of subintelligent human beings."

1972 Hubert Dreyfus

Problem: KI Unternehmen brauchen enorme Datenmengen zum Training

… und viele weitere

Wir werden derzeit Zeugen der größten Urheberrechtsverletzung der Menschheit

… und viele weitere

IMAGE CREDITS: JAKUB PORZYCKI/NURPHOTO / GETTY IMAGES

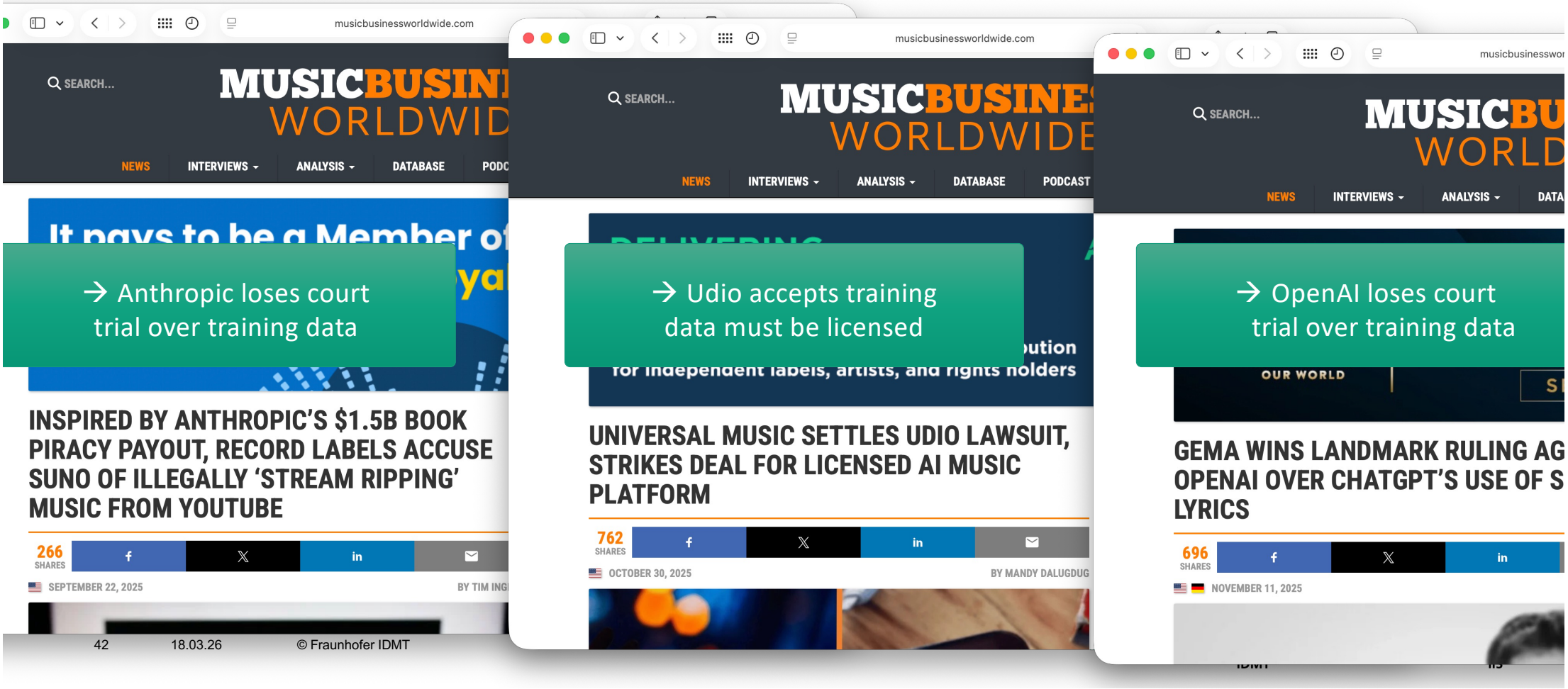
AI

OpenAI failed to deliver the opt-out tool it promised by 2025

Gen-KI Service

Kyle Wiggers — 7:00 AM PST · January 1, 2025

GenAI Unternehmen akzeptieren langsam, dass Datennutzung lizensiert werden muss



Fazit

Sprachsynthese und Musikgenerierungen haben ein Qualitätsniveau erreicht, das für viele Verwertungs-Szenarien in der Kreativwirtschaft ausreichend ist

Menschliche Zuhörer können oft nicht mehr real von synthetisch unterscheiden (nur im Detail)

Wir beobachten derzeit den Beginn einer „Legalize“ Phase der großen KI Unternehmen

Dies wird in je nach Weltregion sehr unterschiedlich gehandhabt (USA vs. Asia vs. EU)

Wie ist Ihre Einschätzung zu den genannten Grenzen und Herausforderungen?

Contact

Hanna Lukashevich
Head of Semantic Media Technologies

+49 3677 467-224
hanna.lukashevich@idmt.Fraunhofer.de

Fraunhofer IDMT
Ehrenbergstrasse 31
98693 Ilmenau, Germany
www.idmt.Fraunhofer.de

Dr.-Ing. Christian Dittmar
Group Manager Spoken Language Processing
Audio and Media Technologies
+49 9131 776-6315
christian.dittmar@iis.fraunhofer.de

Fraunhofer IIS
Am Wolfsmantel 33
91058 Erlangen
https://www.allinga.fraunhofer.de/

Fraunhofer IDMT

Fraunhofer Institute for Digital Media Technology IDMT

Fraunhofer IIS

Fraunhofer Institute for Integrated Circuits IIS