



Community of Practice KIPerWeb

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen



KIPerWEB



Forschungsinstitut
Betriebliche Bildung

- **Update**
 - News & Leaderboard-Update
- **Input**
 - „KI-basierte Transkription und Beschreibung von Audio und Video“
- **Diskussion**

Leaderboard-Update (26.06.2025)

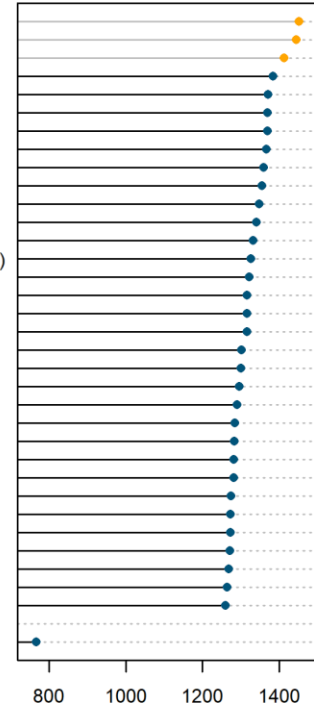


Arena Score German

based on Imarena.ai on May 26, 2025

- Mit Blick auf die Performanz in der Kategorie „German“ löst **Gemini-2.5-Pro** das ältere **Gemini-2.5-pro-preview-05-06** als Spitzenreiter ab.
- Arena-Scores von *nicht*-proprietären Modelle sind rechts ausgewiesen sofern sie mindestens das Niveau von **Gemma-2-9b-it-SimPo** erreichen:
 - **Minimax-m1** (456b parameters total, 45.9b active) löst Deepseek als bestes Open-Weights Modell ab
 - **Gemma-3-12B-it** bleibt das beste Modell für den Hausgebrauch (noch vor **Gemma-3-27B-it**)
 - **Gemma-3n-e4b-it** liegt vor **Gemma-3-4B-it** bei geringerem Ressourcenverbrauch
- Schlusslicht auf dem Leaderboard bleibt Chatglm2-6b

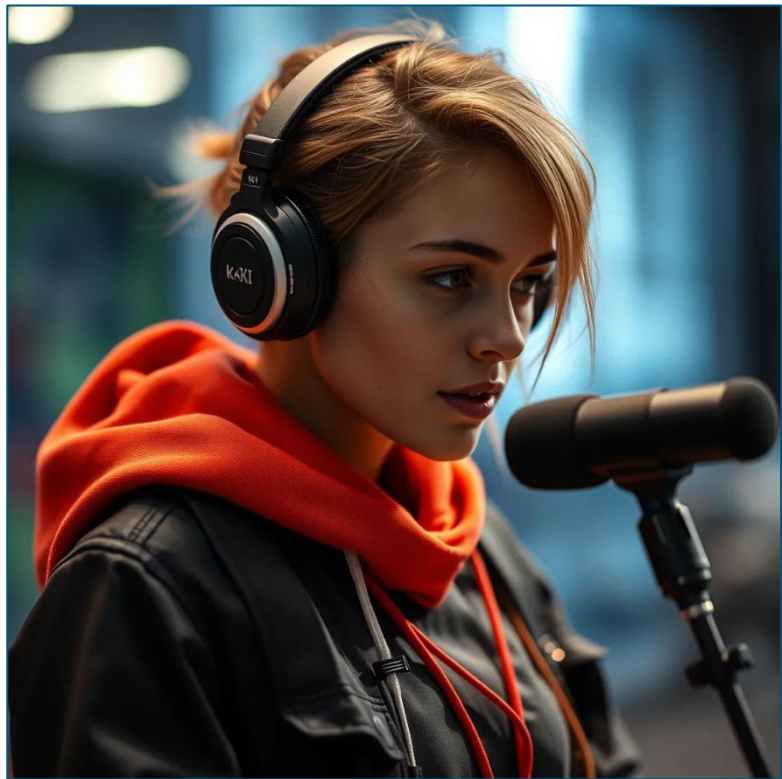
Gemini-2.5-Pro (Proprietary)
o3-2025-04-16 (Proprietary)
ChatGPT-4o-latest 20250326 (Proprietary)
Minimax-m1 (Apache 2.0)
Deepseek-R1 (MIT)
Qwen3-235B-A22B-no-thinking (Apache 2.0)
Deepseek-V3-0324 (MIT)
Gemma-3-12B-it (Gemma)
Deepseek-R1-0528 (MIT)
Gemma-3-27b-it (Gemma)
Qwen3-235B-A22B (Apache 2.0)
Deepseek-V3 (DeepSeek)
Command-a-03-2025 (CC-BY-NC)
llama-3.1-nemotron-ultra-253-v1 (Nvidia Open Licence)
Llama-4-Maverick-17B-128E-Instruct (LLama 4)
Gemma-3n-e4b-it (Gemma)
Qwen3-32b (Apache 2.0)
Llama-4-Scout-17b-16e-instruct (LLama 4)
Llama-3.1-Nemotron-70b-instruct (Llama 3.1)
Meta-Llama-3.1-405b-Instruct-bf16 (Llama 3.1)
Meta-Llama-3.1-405b-Instruct-fp8 (Llama 3.1)
Mistral-Large-2407 (Mistral Research)
QwQ-32B (Apache 2.0)
Qwen3-30b-a3b (Apache 2.0)
Gemma-3-4B-it
Llama-3.3-70B-Instruct (Llama-3.3)
Deepseek-v2.5-1210 (DeepSeek)
Qwen-max-0919 (Qwen)
Athene-70b-0725 (CC-BY-NC-4.0)
Mistral-Small-3.1-24b-instruct-2503 (Apache 2.0)
Mistral-Large-2411 (Mistral Research)
Athene-v2-chat (NexusFlow)
Gemma-2-9b-it-SimPO (MIT)
...
Chatglm2-6b (Apache 2.0)



- Prompt:
„KI-basierte Transkription und Beschreibung
von Audio und Video“

(rechts visualisiert von FLUX.1-schnell)

...



- **Transkription:** „Die Definition von Transkription ist die Verschriftlichung von Audio- und Videodateien. Wortwörtlich bedeutet der aus dem Lateinischen stammende Begriff so viel wie „hinüber Geschriebenes“ und bezeichnet das Übertragen eines gesprochenen Textes in Schrift.“ (Abtipper.de)
- **Bildbeschreibung:** „Beschreibung (besonders in Form eines Schulaufsatzes) von etwas, was auf einem Bild zu sehen ist“ (Duden)
- **Videobeschreibung:** „Eine (YouTube)-Videobeschreibung ist der zugehörige Text unter einem Video. Dieser sollte Zuschauer:innen vermitteln, worum es in dem entsprechenden Video inhaltlich geht. Es sollte also eine kurze Inhaltsbeschreibung sein, die die wichtigsten Keywords und den Titel des Videos enthält.“ (More-fire.com)
- **Audiodeskription:** „Audiodeskription, auch als akustische Bildbeschreibung oder seltener als Audiokommentierung bezeichnet, ist ein Verfahren, das blinden und sehbehinderten Menschen ermöglichen soll, visuelle Vorgänge besser wahrnehmen zu können.“ (Wikipedia)

Quellen: Abtipper.de: <https://www.abtipper.de/transkription/was-bedeutet-transkription/>; More-Fire.com: <https://www.more-fire.com/online-marketing-wiki/video-beschreibung/>; Duden: <https://www.duden.de/rechtschreibung/Bildbeschreibung>; Wikipedia: <https://de.wikipedia.org/wiki/Audiodeskription>;

Kennenswerte KI-Modelle

- openai/whisper-large-v3 (freie Apache 2.0-Lizenz)
- openai/whisper-large-v3-turbo (freie MIT-Lizenz)
- pyannote/speaker-diarization-3.1 (freie MIT-Lizenz)

Kennenswerte KI-Systeme, Varianten und Interfaces

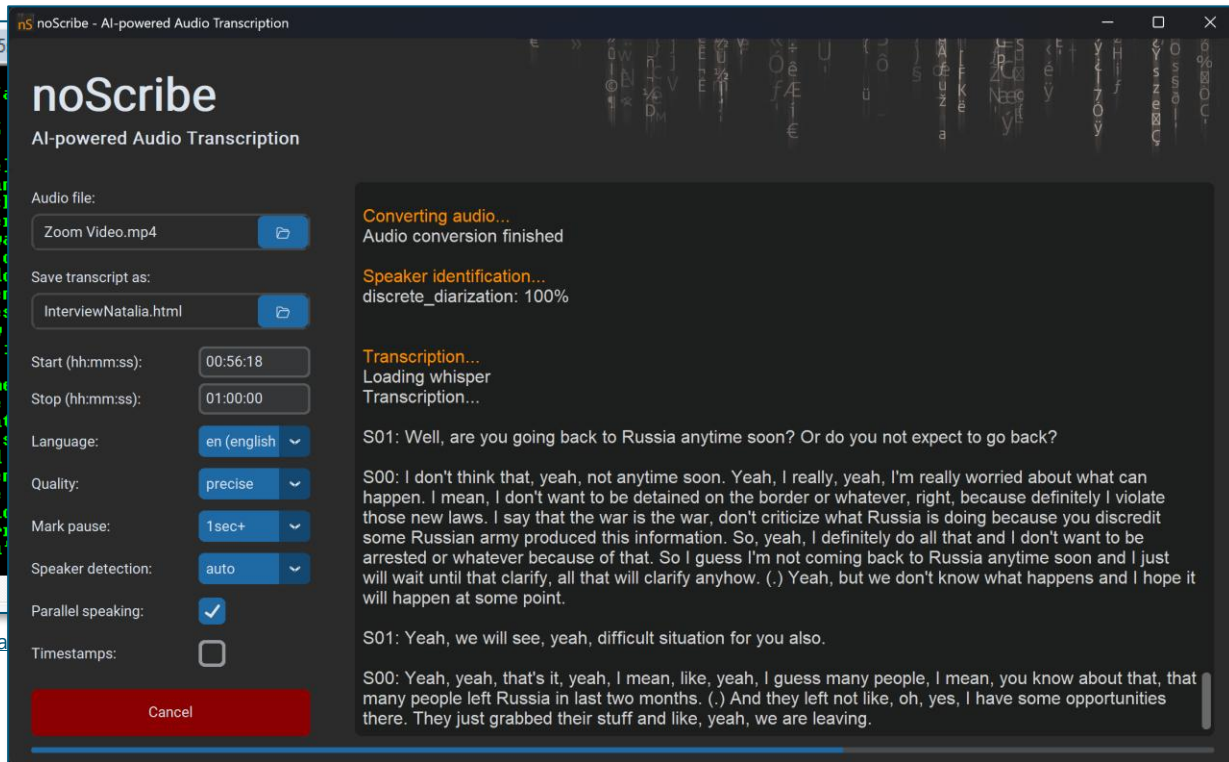
- Faster-Whisper (auch als Standalone Executables für Windows-Nutzende)
<https://github.com/Purfview/whisper-standalone-win>)
- noScribe <https://github.com/kaixxx/noScribe>
- ...

Beispiel Video-/Audio-Transkription



```
Administrator: 41% | 124/306 | 03:48 < 05:00
d:\whisper-fast\dist>whisper-fast
Standalone Faster-Whisper r125
[00:00.760 --> 00:02.760] Fee
[00:07.890 --> 00:08.890] Tha
[00:08.890 --> 00:10.890] I t
[00:12.890 --> 00:17.890] The
[00:17.890 --> 00:22.890] I w
[00:23.890 --> 00:25.890] We
[00:25.890 --> 00:28.890] I d
[00:29.890 --> 00:31.890] Ever
[00:31.890 --> 00:33.890] Does
[00:36.890 --> 00:38.890] Why
[00:43.280 --> 00:46.280] Is
[00:47.280 --> 00:49.280] No,
[01:22.270 --> 01:24.270] Some
[01:24.270 --> 01:26.270] One
[01:29.270 --> 01:31.270] What
[01:31.270 --> 01:35.270] He's
[01:35.270 --> 01:40.270] And
[01:42.270 --> 01:44.270] See
[01:46.270 --> 01:48.270] The
[01:54.740 --> 01:56.740] I lo
[01:58.740 --> 01:59.740] Sor
[02:00.740 --> 02:03.740] You
```

<https://github.com/Purview/whisper-sta>



https://github.com/kaixxx/noScribe/blob/main/img/noScribe_main_window.png

Beispiel Bildbeschreibung

- Hier die Antwort des Leichtgewichts Gemma-3-4b-it zu „Bitte erstelle eine Bildbeschreibung, die das Bild als Alternativtext für blinde und sehbehinderte Menschen detailliert beschreibt.“

- „Hier ist eine detaillierte Bildbeschreibung, die als Alternativtext für blinde und sehbehinderte Personen dienen kann:

Bildbeschreibung:

Das Bild zeigt ein halbes Porträt einer jungen Frau, die vor einem urbanen Hintergrund steht. Sie trägt Kopfhörer mit dem Logo "KONI" auf den Ohren und blickt leicht zur Seite, in Richtung eines Mikrofons. Ihre Frisur ist blond und offen, mit einigen widerschnittlichen Strähnen. Ein auffälliger, leuchtend orangefarbener Schal umgibt ihren Hals und ihre Schultern. Sie trägt eine schwarze Jacke mit Kapuze, die offen liegt.

Das Gesicht der Frau ist im Fokus. Ihre Augen sind dunkel und aufmerksam, ihr Mund leicht geöffnet, als ob sie spricht oder liest. Sie hat einen hellen Teint und wirkt konzentriert. Ein schwarzes Mikrofonstativ steht neben ihr, das Mikrofon selbst ist schwarz und mit einem Kabel verbunden. Der Hintergrund ist verschwommen und zeigt Elemente einer Stadtlandschaft – vermutlich eine Straße oder ein Gebäude mit Glasfassaden, die das Licht reflektieren. Das Licht scheint von links zu kommen und erzeugt einen sanften Schatten auf der rechten Seite des Gesichts.

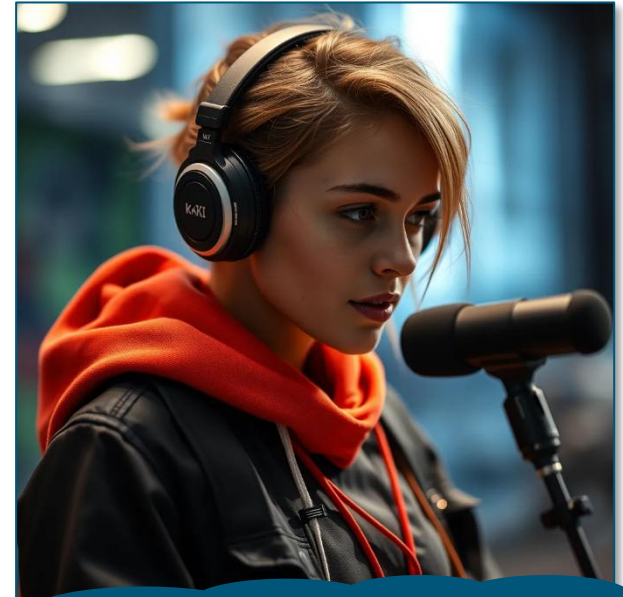
Das Bild hat eine warme Farbtemperatur und wirkt professionell, möglicherweise aufgenommen während einer Podcast-Aufnahme oder eines ähnlichen Audio-Produktionsprozesses“

- Und hier eine Bildbeschreibung „in einem Satz“ statt „detailliert“:

- "Eine junge Frau mit blonden Haaren trägt Kopfhörer und steht vor einem Mikrophon, umgeben von einer urbanen Umgebung."

- Für noch akkuratere Ergebnisse empfehlen sich größere Modelle(mit Blick auf Gemma-3 insb. 12B, 27B)

...seit gestern auch Gemma-3n (bislang nur Preview) bzw. andere/aktuellere Modelle



Mein Fazit: dichte Beschreibung, die einen guten Eindruck vom Bild vermittelt, mit vereinzelten Fehlern bei Vieldeutigem & Nebensächlichem (KONI statt K 🔥 KI; Kapuze vs. Schal)

- Fragen?
- Anregungen?
- Erfahrungen?