



Community of Practice KIPerWeb

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen



KIPerWEB



Forschungsinstitut
Betriebliche Bildung

- **Update**
 - News & Leaderboard-Update
- **Input**
 - „KI in der Softwareentwicklung“
- **Diskussion**

Leaderboard-Update (31.10.2025)



Rechts das aktuelle LMArena-Leaderboard für die Kategorie German mit Blick auf die besten open-weights-Modelle (sowie ausgewählte proprietäre Modelle).

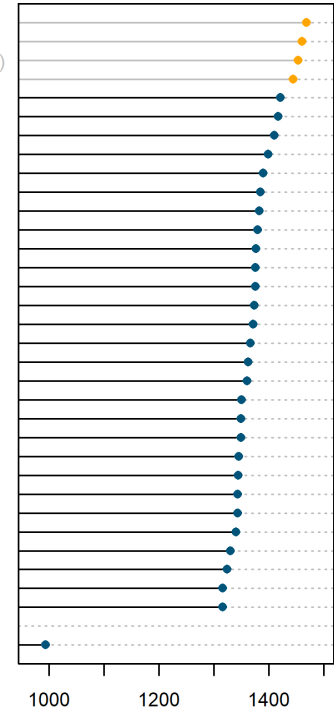
Bemerkenswert:

- Mit **GPT-4.5-preview** bleibt ein proprietäres Modell an der Spitze und auch sonst diesmal kaum nennenswerte Änderungen
- Neuzugang **step-3** (321B total parameters, 38B active) von stepfun-ai
- Fliegengewicht **Gemma-3n-e4b-it** gleichauf mit OpenAI's **gpt-oss-120b**
- Bestes Modell für den Hausgebrauch bleibt m.E. **Gemma-3-12b-it**
- Ausgewiesenes Schlusslicht bleibt **qwen1.5-4b-chat** (994)

Arena Score German

based on lmarena.ai on Oct 31, 2025

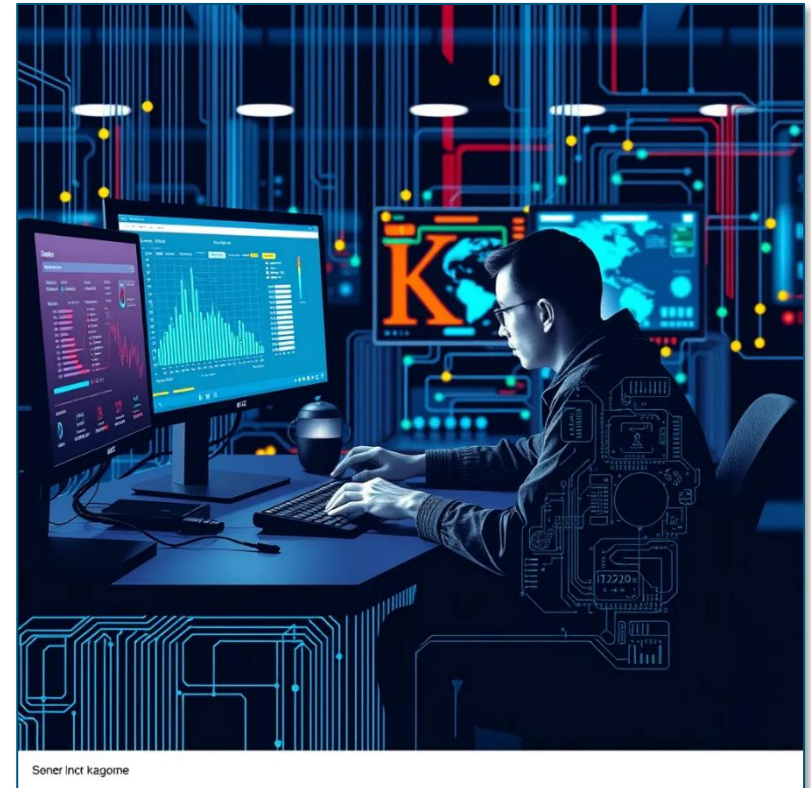
gpt-4.5-preview-2025-02-27 (Proprietary)
Gemini-2.5-Pro (Proprietary)
claude-opus-4-1-20250805-thinking-16k (Proprietary)
gpt-5-high (Proprietary)
kimi-k2-0905-preview (Modified MIT)
Qwen3-235B-A22B-instruct-2507 (Apache 2.0)
kimi-k2-0711-preview (Modified MIT)
Deepseek-V3.1 (MIT)
Deepseek-R1 (MIT)
Deepseek-V3-0324 (MIT)
longcat-flash-chat (MIT)
Qwen3-235B-A22B-thinking-2507 (Apache 2.0)
Qwen3-235B-A22B (Apache 2.0)
Qwen3-235B-A22B-no-thinking (Apache 2.0)
Deepseek-R1-0528 (MIT)
Qwen3-next-80b-a3b-instruct (Apache 2.0)
Gemma-3-12B-it (Gemma)
Gemma-3-27B-it (Gemma)
Qwen3-next-80b-a3b-thinking (Apache 2.0)
glm-4.5 (MIT)
Qwen3-Coder-480b-a45b-instruct (Apache 2.0)
mistral-small-2506 (Apache 2.0)
Command-a-03-2025 (CC-BY-NC)
Qwen3-30b-a3b-instruct-2507 (Apache 2.0)
Minimax-m1 (Apache 2.0)
glm-4.5-air (MIT)
Deepseek-V3 (Deepseek)
step-3 (MIT)
Llama-4-Maverick-17B-128E-Instruct (LLama 4)
Qwen3-32b (Apache 2.0)
gpt-oss-120b
Gemma-3n-e4b-it (Gemma)
...
qwen1.5-4b-chat (Qianwen Licence)



Fokusthema: KI in der Softwareentwicklung



- Prompt „KI in der Softwareentwicklung“
(rechts visualisiert von FLUX.1-schnell, seed 1659051023)














Varianten von KI in der Software-Entwicklung



- General-Purpose Chatbots
(z.B. Claude.ai)
- Coding Agents
(z.B. Claude Code, OpenAI Codex, Gemini CLI)
- Coding Assistants
(z.B. Tabby, Github Copilot)
- ...

P.S. Bzgl. der Modelle ist die Performance in den Kategorien „Coding“ deutlich anders als die eben gezeigte Kategorie „Deutsch“ (s. Abb. rechts)

Coding		Search by model name...	Default			
Rank (UB) ↑	Model ↓	Score ↓	95% CI (±) ↓	Votes ↓	Organization ↓	License ↓
1	 claude-sonnet-4-5-20250929-thinking-32k	1527	±17	1.313	Anthropic	Proprietary
1	 claude-opus-4-1-20250805-thinking-16k	1515	±10	3.849	Anthropic	Proprietary
2	 claude-opus-4-1-20250805	1499	±9	6.171	Anthropic	Proprietary
3	 claude-opus-4-20250514-thinking-16k	1489	±8	6.048	Anthropic	Proprietary
3	 qwen3-max-2025-09-23	1476	±16	1.388	Alibaba	Proprietary
4	 qwen3-max-preview	1477	±11	3.187	Alibaba	Proprietary
4	 longcat-flash-chat	1473	±13	2.260	Meituan	MIT
4	 claude-sonnet-4-5-20250929	1472	±17	1.275	Anthropic	Proprietary
4	 qwen3-v1-235b-a22b-instruct	1468	±17	1.213	Alibaba	Apache 2.0
4	 claude-haiku-4-5-20251001	1461	±27	459	Anthropic	Proprietary
5	 qwen3-235b-a22b-instruct-2507	1470	±9	5.312	Alibaba	Apache 2.0



Andrej Karpathy ✓

@karpathy

...

There's a new kind of coding I call "vibe coding", where you fully give in to the vibes, embrace exponentials, and forget that the code even exists. It's possible because the LLMs (e.g. Cursor Composer w Sonnet) are getting too good. Also I just talk to Composer with SuperWhisper so I barely even touch the keyboard. I ask for the dumbest things like "decrease the padding on the sidebar by half" because I'm too lazy to find it. I "Accept All" always, I don't read the diffs anymore. When I get error messages I just copy paste them in with no comment, usually that fixes it. The code grows beyond my usual comprehension, I'd have to really read through it for a while. Sometimes the LLMs can't fix a bug so I just work around it or ask for random changes until it goes away. It's not too bad for throwaway weekend projects, but still quite amusing. I'm building a project or webapp, but it's not really coding - I just see stuff, say stuff, run stuff, and copy paste stuff, and it mostly works.

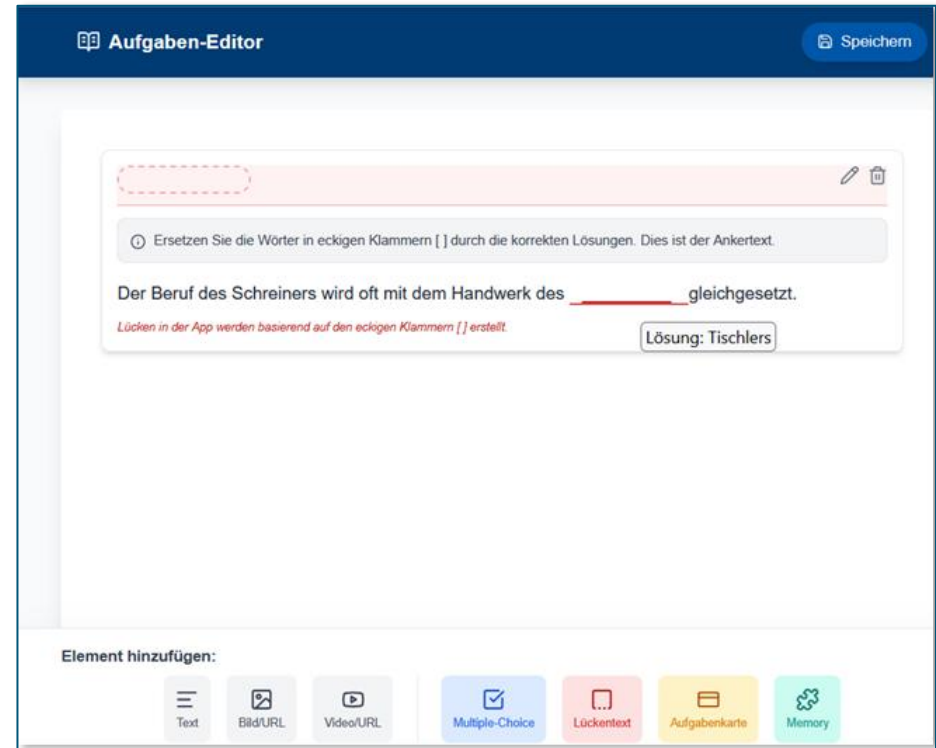
Live-Demo mit Google's Gemini

Beispiel-Prompt:

„Ich brauche einen Aufgaben-Editor, über den man Aufgaben für eine Berufswahl-App erstellen kann. Bitte erstelle eine HTML-Datei, die alles enthält was man dafür braucht.

In der Mitte soll eine Arbeitsfläche sein, wie in Word. Hier soll man die erstellten Elemente untereinander sehen, bearbeiten und löschen können.

Unten sollten Buttons für das Hinzufügen von Elementen sein. Man soll z.B. unterschiedliche Fragentypen ergänzen können (z.B. Lückentext, Multiple-choice-Frage, Aufgabenkarten etc.). Auch sollte man Textfelder, Bilder oder Videos (hochladen und) ergänzen können.“



(Die Abbildung zeigt das erzielte Ergebnis von Gemini <https://gemini.google.com/> - nach Anpassung des Farbschemas und Ergänzung eines Memory-Aufgabenformats)

Exkurs: MyQDA – Open Text Annotation Tool



Save Project

Load Project

Dokumente

+ Hinzufügen

Löschen

Text_A_Protokoll

Text_B_Interview

Codes

+ Hinzufügen

Löschen

Positive Äußerung

Negative Äußerung

Neutrale Beobachtung

Aktueller Code: Positive Äußerung

✓ Annotieren

✗ Unannotieren

🔍 Annotationen

📊 Zahlen

🤖 KI

Der erste Satz ist unproblematisch. Hier beginnt der spannende Teil, der annotiert werden sollte. Das ist der dritte Satz.

- Fragen?
- Anregungen?
- Erfahrungen?