



COPYRIGHT CLAIM

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Copyright der Trainingsdaten

Johannes Hötter
Christian Warmuth

Wo kommen Trainingsdaten her?

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement



An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model 'without permission ... or compensation.'

By [James Vincent](#), a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Feb 6, 2023, 5:56 PM GMT+1 | [16 Comments](#) / [16 New](#)

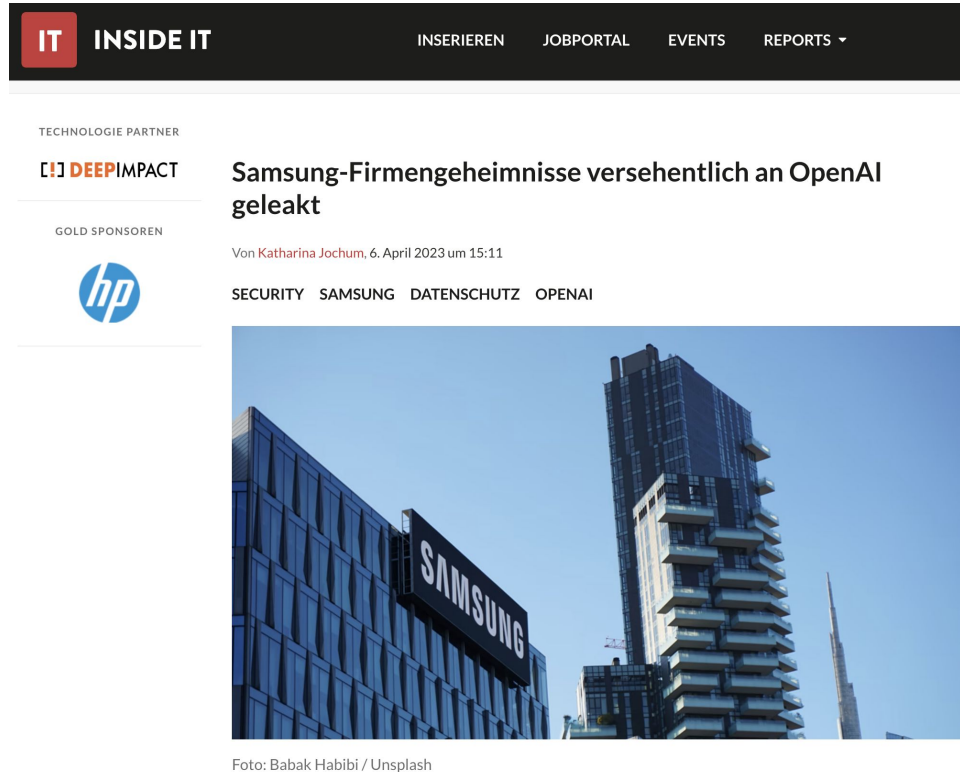


Copyright der Trainingsdaten

Johannes Hötter,
Christian Warmuth

openHPI

Wo kommen Trainingsdaten her?



**Copyright der
Trainingsdaten**

Johannes Hötter,
Christian Warmuth

openHPI

Folie 3

Wie gehen die Provider mit den Daten um?



Research ▾ Product ▾ Developers ▾ Safety Company ▾

3. Content

(a) **Your Content.** You may provide input to the Services ("Input"), and receive output generated and returned by the Services based on the Input ("Output"). Input and Output are collectively "Content." As between the parties and to the extent permitted by applicable law, you own all Input. Subject to your compliance with these Terms, OpenAI hereby assigns to you all its right, title and interest in and to Output. This means you can use Content for any purpose, including commercial purposes such as sale or publication, if you comply with these Terms. OpenAI may use Content to provide and maintain the Services, comply with applicable law, and enforce our policies. You are responsible for Content, including for ensuring that it does not violate any applicable law or these Terms.

(b) **Similarity of Content.** Due to the nature of machine learning, Output may not be unique across users and the Services may generate the same or similar output for OpenAI or a third party. For example, you may provide input to a model such as "What color is the sky?" and receive output such as "The sky is blue." Other users may also ask similar questions and receive the same response. Responses that are requested by and generated for other users are not considered your Content.


(c) **Use of Content to Improve Services.** We do not use Content that you provide to or receive from our API ("API Content") to develop or improve our Services. We may use Content from Services other than our API ("Non-API Content") to help develop and improve our Services. You can read more here about [how Non-API Content may be used to improve model performance](#). If you do not want your Non-API Content used to improve Services, you can opt out by filling out [this form](#). Please note that in some cases this may limit the ability of our Services to better address your specific use case.

Copyright der Trainingsdaten


Johannes Hötter,
Christian Warmuth

openHPI

Erneutes Thema: Open-Source für Self-Hosting

 Product Solutions Open Source Pricing

Search / Sign in Sign up

 nomic-ai/gpt4all Public

Notifications Fork 4k Star 37.6k

<> Code Issues 289 Pull requests 12 Actions Projects Security Insights

main 13 branches 0 tags

Go to file Code

zanussbaum Merge pull request #472 from berkantay/main 2c8e109 4 days ago 267 commits

chat	Merge branch 'main' into chat-windows-binary	2 months ago
configs	Update finetune.yaml	3 weeks ago
eval_data	started eval script and added eval data	2 months ago
figs	feat: wip training log	3 weeks ago
peft @ 098962f	feat: peft submodule	2 months ago
.gitignore	Merge: main into gptj	3 weeks ago
.gitmodules	chore: remove transformers submodule	3 weeks ago
GPT-J_MAP.md	fix: rename	3 weeks ago
LICENSE.txt	Add MIT license.	last month
README.md	Update README.md	last week
TRAINING_LOG.md	fix: format	3 weeks ago
build_map.py	fix: rename	3 weeks ago
clean.py	fix: clean where prompt is randomly 1 char	last month
create_hostname.sh	feat: multinode setup	last month
data.py	Merge: main into gptj	3 weeks ago
env.yaml	feat: env for conda, pip	2 months ago
eval_figures.py	feat: eval on new onli model	last month

About

gpt4all: an ecosystem of open-source chatbots trained on a massive collections of clean assistant data including code, stories and dialogue

Readme MIT license 37.6k stars 497 watching 4k forks Report repository


Releases

No releases published

Packages

No packages published

Contributors 26

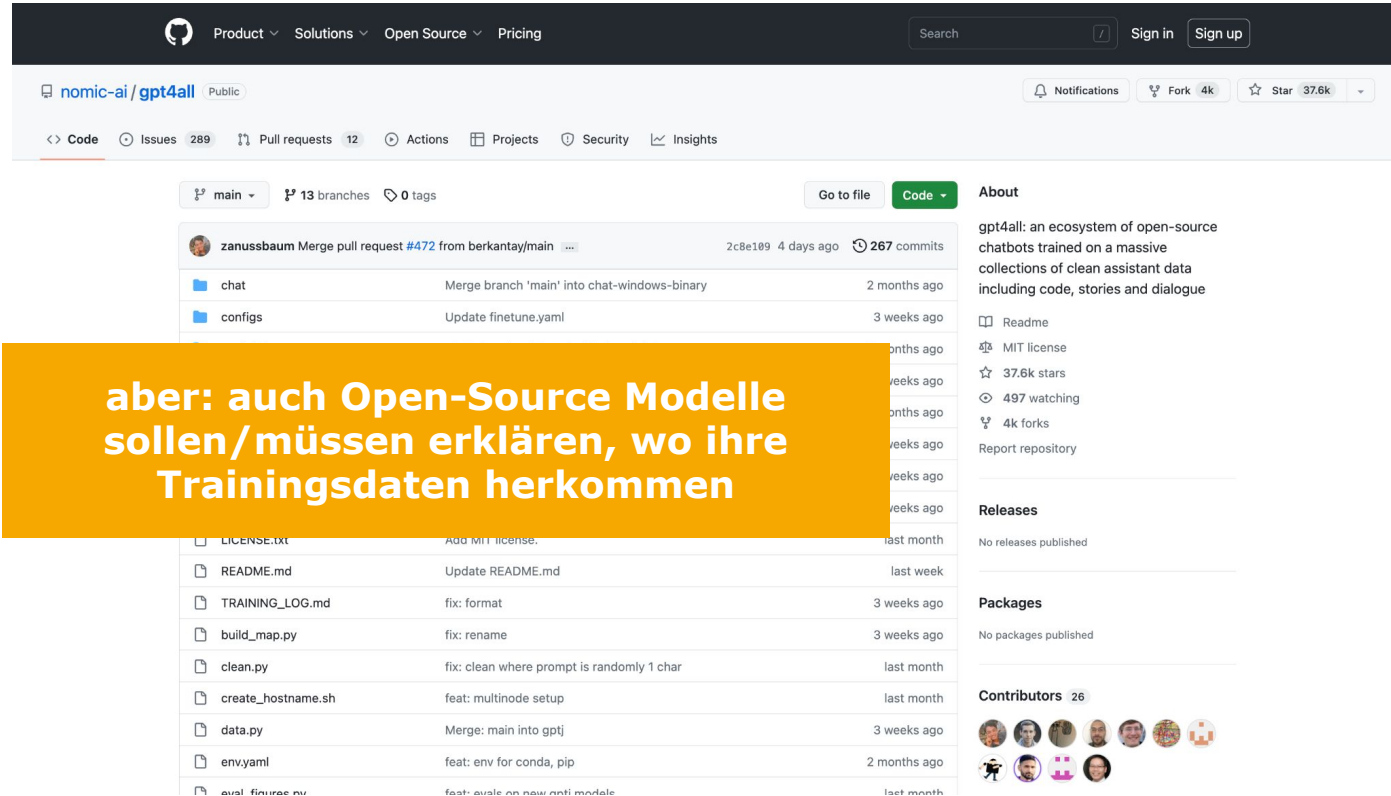


Copyright der Trainingsdaten

Johannes Hötter,
Christian Warmuth

openHPI

Erneutes Thema: Open-Source für Self-Hosting



aber: auch Open-Source Modelle sollen/müssen erklären, wo ihre Trainingsdaten herkommen

File	Commit Message	Time Ago
chat	Merge branch 'main' into chat-windows-binary	2 months ago
configs	Update finetune.yaml	3 weeks ago
LICENSE.txt	Add MIT license.	last month
README.md	Update README.md	last week
TRAINING_LOG.md	fix: format	3 weeks ago
build_map.py	fix: rename	3 weeks ago
clean.py	fix: clean where prompt is randomly 1 char	last month
create_hostname.sh	feat: multinode setup	last month
data.py	Merge: main into gptj	3 weeks ago
env.yaml	feat: env for conda, pip	2 months ago
eval: finetune.py	feat: eval on new gptj models	last month

About
gpt4all: an ecosystem of open-source chatbots trained on a massive collections of clean assistant data including code, stories and dialogue

Releases
No releases published

Packages
No packages published

Contributors 26

**Copyright der
Trainingsdaten**

Johannes Hötter,
Christian Warmuth

openHPI

Zwiespalt: Datensparsamkeit vs. Performance

Art. 5 DSGVO

Grundsätze für die Verarbeitung personenbezogener Daten

- (1) Personenbezogene Daten müssen
- a) auf rechtmäßige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden („Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz“);
 - b) für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; eine Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke gilt gemäß [Artikel 89](#) Absatz 1 nicht als unvereinbar mit den ursprünglichen Zwecken („Zweckbindung“);
 - c) dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein („Datenminimierung“);
 - d) sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein; es sind alle angemessenen Maßnahmen zu treffen, damit personenbezogene Daten, die im Hinblick auf die Zwecke ihrer Verarbeitung unrichtig sind, unverzüglich gelöscht oder berichtigt werden („Richtigkeit“);

mehr Trainingsdaten = bessere Ergebnisse

**Copyright der
Trainingsdaten**

Johannes Hötter,
Christian Warmuth

openHPI

Quelle: <https://dsgvo-gesetz.de/art-5-dsgvo/>

Folie 7



COPYRIGHT CLAIM

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Copyright der Trainingsdaten

Johannes Hötter
Christian Warmuth