



29.8.2023

# Whitebox statt Blackbox

## Mindeststandards und offene Sprachmodelle im Schreibprozess und in der Schreibberatung

*Christopher Klamm*

*University of Mannheim*

AI

# OpenAI built a text generator so good, it's considered too dangerous to release

Zack Whittaker @zackwhittaker /

*[...] OpenAI said its new natural language model [...] was trained to predict the next word in a sample of [...] internet text. The end result was the system generating text that “adapts to the style and content of the conditioning text,” allowing the user to “generate realistic and coherent continuations about a topic of their choosing.” The model is a vast improvement on the first version by producing longer text with greater coherence. [...]*

AI

# OpenAI built a text generator so good, it's considered too dangerous to release

**Zack Whittaker** @zackwhittaker / 6:17 PM GMT+1 • February 17, 2019

*[...] OpenAI said its new natural language model [...] was trained to predict the next word in a sample of [...] internet text. The end result was the system generating text that “adapts to the style and content of the conditioning text,” allowing the user to “generate realistic and coherent continuations about a topic of their choosing.” The model is a vast improvement on the first version by producing longer text with greater coherence. [...]*

**Anwendungen** von  
Sprachmodellen  
im Schreibprozess  
und der Schreibberatung?

# [Sprachmodelle] und *Anwendungen beim Schreiben*

*“Current LLMs should be used as writing aids, not much more.” (LeCun 2023)*

- **Frage-Antwort-Szenarien mit KI als “Schreib-Copilot”**, um Fragen im Schreibprozess unmittelbar (24/7) zu beantworten
- Studierende **aktivierend unterstützen**, um bspw. beim Brainstorming, **Schreibblockaden** oder bei der Überarbeitung zu helfen
- **gezielte Schreibberatungen**, bei denen die SchreibberaterInnen sich fokussierter auf individuelle Fragestellungen konzentrieren können
- **Argumentationsfähigkeiten** verbessern
- **Verständlichkeit** und **adressatInnen-gerechnetes Schreiben** verbessern

*... und viele mehr!*

Nicht nur *für was können [Sprachmodelle] verwendet werden*,  
sondern auch *was sind unsere Anforderungen an*  
*[Sprachmodelle]*, wenn wir diese im wissenschaftlichen Schreiben  
und in der Schreibberatung einsetzen wollen?

*“One of the key questions to ask is whether a **demonstrated capability is a cherry-picked example** that a model produces 40% of the time, or if it points to **robust and reliable model behavior**.”* (Alammar

Was sind **offene**  
Sprachmodelle?

# Whitebox (offene) vs. Blackbox (closed) Sprachmodelle

Considerations	internal research only high risk control low auditability limited perspectives	community research low risk control high auditability broader perspectives
Level of Access	fully closed	fully open
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	BLOOM (BigScience) GPT-J (EleutherAI)



# Whitebox (offene) vs. Blackbox (closed) Sprachmodelle

Considerations	internal research only high risk control low auditability limited perspectives	community research low risk control high auditability broader perspectives
Level of Access	fully closed	fully open
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	BLOOM (BigScience) GPT-J (EleutherAI)

“When all **aspects and components** of a system are **inaccessible outside the developer organization**, or even closed outside a specific subsection of an organization, the system is fully closed.” (Solaiman 2023a)

# Whitebox (offene) vs. Blackbox (closed) Sprachmodelle

Considerations	internal research only high risk control low auditability limited perspectives	community research low risk control high auditability broader perspectives
Level of Access	fully closed	fully open 🦋
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	BLOOM (BigScience) GPT-J (EleutherAI)

“When **all aspects** of the system are **accessible and downloadable**, **including all components**, the system is fully open. These systems cannot be gated and by definition are fully public.” (Solaiman 2023a)



# Whitebox (offene) vs. Blackbox (closed) [Sprachmodelle]

→ keine binäre Trennung  !

Considerations	internal research only high risk control low auditability limited perspectives					community research low risk control high auditability broader perspectives	
Level of Access	fully closed	gradual/staged release	hosted access	gated to public cloud-based/API access	downloadable	fully open	
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	GPT-2 (OpenAI) Stable Diffusion (Stability AI)	DALLE-2 (OpenAI) Midjourney (Midjourney)	GPT-3 (OpenAI)	OPT (Meta) Craiyon (craiyon)	BLOOM (BigScience) GPT-J (EleutherAI)	

# Whitebox (offene) vs. Blackbox (closed) [Sprachmodelle]

→ keine binäre Trennung ●● !



BLOOM 

HuggingChat

<https://huggingface.co/chat/>

*“BLOOM is an autoregressive Large Language Model (LLM), trained to continue text from a prompt on vast amounts of text data using industrial-scale computational resources.” (BigScience 2022)*

- **Open-Source Multilingual Language Model** (46 Sprachen)
- von Mai 2021 bis Mai 2022, haben mehr als 1.000 ForscherInnen aus 60 Ländern und mehr als 250 Unternehmen in der BigScience-Initiative an der Erstellung von BLOOM gearbeitet
- ModelCard für mehr Informationen: [huggingface.co/bigscience/bloom](https://huggingface.co/bigscience/bloom)

weitere offene Sprachemodelle <https://github.com/eugeneyan/open-llms>

# Stärken offener Modelle

- **Transparenz und Überprüfbarkeit** *(alle Komponenten sind öffentlich, wodurch das Verständnis, wie das Model arbeitet, ersichtlich ist und auch mögliche Schwachstellen durch die Community besser identifiziert werden können)*
- **Community und Forschungsgemeinschaften** *(können direkter gemeinsam Fehler, wie Sicherheitsprobleme oder Verletzung der Privatsphäre, beheben)*
- **Anpassungsmöglichkeit** *(durch den direkten Zugang können die Modelle an die Bedürfnisse angepasst werden und Anpassung ist häufig ressourcen-effizienter)*
- **Qualität und Peer Review** *(unabhängige Begutachtung und häufig die Perspektiven von unabhängigen Forschungsgruppen)*
- **Schnellere Entwicklung und Innovation** *(auf der Basis bereits bestehender offener Modelle lassen sich neue schneller entwickeln)*

# Schwächen von offenen Modellen

- **Datenschutzprobleme** (*können persönliche Daten enthalten, da schwer überprüft werden kann, welche Trainingsdaten verwendet wurden*)
- **Schwachstellenerkennung** (*Schwachstellen können leichter gefunden werden, da alles durch die AngreiferInnen leicht analysiert werden kann*)
- **Regulierungs- und Lizenzprobleme** (*verwenden ggf. nicht lizenzierte Trainingsdaten und beachten vorhandene Regularien nicht*)
- **Zero-Day Schwachstellen** (*eingebaute Hintertüren, die bspw. verletzendes Inhalte auf Knopfdruck produzieren → Datenintegrität oft schwer zu überprüfen*)

**Standards** für  
Sprachmodelle?



**“On artificial intelligence, trust is a must, not a nice to have.”**

Margrethe Vestager 2021, Executive Vice-President for a Europe fit for the Digital Age

# Diskurs um Standards für Sprachmodelle

- aktiver Diskurs, um die Frage, wie *Sprachmodelle und deren Anwendung* reguliert, dokumentiert, etc. werden sollen
- ein Framework kann ForscherInnen, AnwenderInnen und NutzerInnen bei einer systematischen Analyse helfen
- offene Frage, **welche Standards** es geben soll (z.B. in Bezug auf *Transparenz, Zugänglichkeit, Folgenabschätzung* usw. Solaiman et al. 2023b)
- Vorschläge z.B. durch die EU (EU AI Act 2023)

*“In the absence of clear standards for deployment and risk mitigation, release decisionmakers must weigh the trade-offs of different options themselves.” (Solaiman 2023b)*

# EU AI Act

## Data

*Data Sources*  
*Data Governance*  
*Copyright data*

## Model

*Capabilities/ Limitations*  
*Risks/ Mitigations*  
*Evaluation*  
*Testing*

## Compute











*Compute*  
*Energy*

## Deployment


*Machine-generated content*  
*Member states*  
*Downstream documentation*

# Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	 OpenAI	 cohere	 stability.ai	 ANTHROPIC	 Google	 BigScience	 Meta	 AI21 labs	 ALEPH ALPHA	 EleutherAI	
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ● ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ●	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ○ ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

# Standards, Guidelines, ...

**“With LLMs it will soon be less about the code than the training data.”** (Socher 2023) We need to start incorporating **“open source training data, human feedback, source weights”** (Socher 2023) and create **more open source models**, such as BLOOM  (Scao et al. 2022)

## Bestehende Überlegungen

- **Ethical guidelines** (Pistilli et al. 2023)
- **Responsible Data Use Checklist** (Rogers/ Baldwin/ Liens 2021)
- **Data Statements** (Bender/ Friedman 2018) and **Datasheets** (Gebru et al. 2021)
- **AI Democratization** (Seeger et al. 2023)
- **Efficient Methods and Models** (Trevisor et al. 2023, Ostendorf/ Rehm 2023, ...)
- **Benchmarks** (Reimers 2022, Degjani et al. 2021, Raji et al. 2021)
- ...

# **Minimalstandards** für Sprachmodelle im **Schreiben?**

# Probleme von *Blackbox*-Modellen für das Schreiben

- **Fehlende Verlässlichkeit** (die Qualität der Vorhersagen kann sich von Version zu Version (unbemerkt) verändern (auch verschlechtern))
- **Kein geschützter Schreibraum** (Gefahr von fehlendem Datenschutz von “privaten” Schreibsessions, die dann wieder zum Training verwendet werden und später identifiziert werden)
- **Fehlende Kontrollmöglichkeit** (ohne Kenntnis über z.B. die Trainingsdaten sind wissenschaftlichen Standards oder Urheberrechtsfragen nicht direkt überprüfbar)
- **Fehlende systematische Bewertung des Schreibwissens** (ohne gezielte Benchmarks für wissenschaftliches Schreiben, verbleibt die Bewertung der Fähigkeiten beim *cherry picking* von Beispielen)
- **Fehlende Anpassbarkeit** (bei nicht offenen Modelle ist es schwer, selbst die Modelle anzupassen, wenn bspw. falsche Informationen in den Modellen codiert sind)

# Mögliche Minimalstandards *(nicht abschließend)* für Sprachmodelle im Schreibprozess und in der Schreibberatung

- **Überprüfbarkeit** von Trainingsdaten und anderen Komponenten des Models
- **Geschützter Raum** mit Hilfe von lokalen, offline-basierten oder abgesicherten Sprachmodelle, die ihre Daten nicht mit Dritten teilen
- **Verlässlichkeit** der Performanz mit Hilfe von ***systematischen Benchmarks*** für das wissenschaftliche Schreiben
- **Integration von Unsicherheit** als aktive Kommunikation darüber, wie sicher sich gerade das Sprachmodell bei der Generierung ist
- **Zugänglichkeit und Anpassbarkeit** - erlaubt es “Experten in the loop” auch fehlerhafte Informationen anzupassen und zu verbessern



## Mögliche Minimalstandards (*nicht abschließend*) für Sprachmodelle im Schreibprozess und in der Schreibberatung

- **Überprüfbarkeit** von Trainingsdaten und anderen Komponenten des Models
- **Geschützter Raum** mit Hilfe von lokalen, offline-basierten oder abgesicherten

**Für diese Standards braucht es offenere Modelle, wie zum Beispiel BLOOM, die eine *tiefer*e Überprüfbarkeit, Sicherheit, Zugänglichkeit, Anpassbarkeit usw. ermöglichen**

- **Integration von Unsicherheit** als aktive Kommunikation darüber, wie sicher sich gerade das Sprachmodell bei der Generierung ist
- **Zugänglichkeit und Anpassbarkeit** - erlaubt es "Experten in the loop" auch fehlerhafte Informationen anzupassen und zu verbessern

# Mögliche Leitfragen *(nicht abschließend)* für ein Modell im Einsatz für das wissenschaftliche Schreiben

- Können wir unser [Sprachmodell] auch **ohne Teilen von Informationen privat verwenden?**
- Können wir die **Verlässlichkeit** unseres [Sprachmodell] **in der Beantwortung von Standardfragen\*** im Zuge einer Schreibberatung bzw. Bearbeitung eines wissenschaftlichen Textes überprüfen?
- Können wir die Komponenten unseres [Sprachmodell], besonders die **Trainingsdaten, nach wissenschaftlichen Standards analysieren?**
- Können wir unser [Sprachmodell] **bei fehlerhaftem Verhalten anpassen?**

\* offener Fragen-Antwort-Katalog, den wir für eine systematische Bewertung benötigen

# Benchmarking-Initiative 🚦 zur Bewertung von Sprachmodellen für das wissenschaftliche Schreiben

## Benchmarks für Sprachmodelle






*“Progress in NLP has traditionally been measured through a **selection of task-level datasets** that gradually became accepted benchmarks [...]” (Kielbaso et al. 2021)*

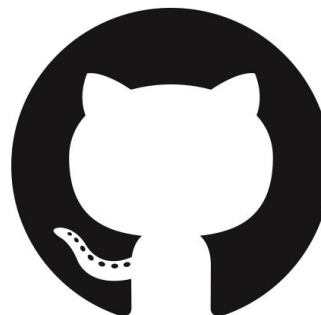
👤 **Gemeinsame Benchmarks 🚦 für die Nutzung von [Sprachmodellen] in der Schreibberatung und im wissenschaftlichen Schreiben entwickeln**

→ [christopher.klamm@uni-mannheim.de](mailto:christopher.klamm@uni-mannheim.de)

# Fragen und Feedback



 klamm.ai  
 github/chkla  
 twitter/chklamm  
 huggingface.co/chkla  
 christopher@klamm.ai



[github.com/chkla/openLLMs-LearningAID](https://github.com/chkla/openLLMs-LearningAID)