Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

[Andreas Liesenfeld](https://orcid.org/)
andreas.liesenfeld@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Alianda Lopez
ada.lopez@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Mark Dingemanse
mark.dingemanse@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Project	Availability					Documentation						Access methods	
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Data sheet	Package	API
chatGPT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	~
OpenAI	LLM base: GPT3.5, GPT4			RLHF base: Instruct-GPT									https://chat.openai.com
StableVicuna-13B	✓	✓	~	~	~	~	~	✓	✓	✗	✗	~	✗
CarperAI	LLM base: LLaMA			RLHF base: oasst1, anthropic									https://huggingface.co/CarperAI/stable-vicuna-13b-delta
text-generation-webu	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
oobabooga	LLM base: various			RLHF base: various									https://github.com/Akegarasu/ChatGLM-webui
MPT-7B-Instruct	✓	✗	✓	~	✗	✓	✓	~	✗	✗	✗	✓	✗
MosaicML	LLM base: MosaicML			RLHF base: dolly, anthropic									https://github.com/mosaicml/llm-foundry#mpt
Falcon-40B-Instruct	✓	~	✓	~	✓	✓	~	~	~	✗	~	~	✗
TII	LLM base: Falcon 40B			RLHF base: Baize (synthetic)									https://huggingface.co/tiiuae/falcon-40b-instruct
minChatGPT	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	✓
ethanyanjiali	LLM base: GPT2			RLHF base: anthropic									https://github.com/ethanyanjiali/minChatGPT
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	~	✓
carperai	LLM base: various (pythia, flan, OPT)			RLHF base: various									https://github.com/carperai/trlx
stanford_alpaca	✓	✓	~	~	✗	~	✓	✓	✗	✗	~	✗	✗
Tatsu labs	LLM base: LLaMA			RLHF base: Self-Instruct (synthetic)									https://github.com/tatsu-lab/stanford_alpaca
Cerebras-GPT-111M	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✗
Cerebras, Schramm	LLM base: not open			RLHF base: alpaca (synthetic)									https://huggingface.co/SebastianSchramm/Cerebras-GPT-111M-instruction
OpenChatKit	✓	✓	✓	✓	✓	✓	✓	✗	~	✗	✗	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RLHF base: OIG									https://github.com/togethercomputer/OpenChatKit
dolly	✓	✓	✓	~	✗	✓	✓	✓	~	✗	✗	✓	✗
databrickslabs	LLM base: EleutherAI pythia			RLHF base: databricks-dolly-15k									https://github.com/databrickslabs/dolly
CharRWKV	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
BlinkDL	LLM base: RWKV-LM (own)			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/BlinkDL/ChatRWKV
BELLE	✓	✓	~	✓	✓	✓	✓	✓	✓	✗	~	✗	✗
LianjiaTech	LLM base: LLaMA, BLOOMZ			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/LianjiaTech/BELLE
Open-Assistant	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
LAION-AI	LLM base: oasst1 (own)			RLHF base: OIG									https://github.com/LAION-AI/Open-Assistant
xmtf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RLHF base: xP3									https://github.com/bigscience-workshop/xmtf

Figure 1: How open and transparent are current instruction-following text generators? Snapshot as of June 2023 — updates at osf.io/d6fsr

ABSTRACT

Large language models that exhibit instruction-following behaviour represent one of the biggest recent upheavals in conversational interfaces, a trend in large part fuelled by the release of OpenAI's ChatGPT, a proprietary large language model for text generation fine-tuned through reinforcement learning from human feedback (LLM+RLHF). We review the risks of relying on proprietary software and survey the first crop of open-source projects of comparable architecture and functionality. The main contribution of this paper is to show that openness is differentiated, and to offer scientific documentation of degrees of openness in this fast-moving field. We evaluate projects in terms of openness of code, training data, model weights, RLHF data, licensing, scientific documentation, and access methods. We find that while there is a fast-growing list of projects billing themselves as 'open source', many inherit undocumented data of dubious legality, few share the all-important instruction-tuning (a key site where human annotation labour is involved), and careful scientific documentation is exceedingly rare. Degrees of openness are relevant to fairness and accountability at all points, from data collection and curation to model architecture, and from training and fine-tuning to release and deployment.

CCS CONCEPTS

• **Natural language generation**; • **Emerging technologies**; • **Surveys and overview**; • **Open-source software**; • **Evaluation**;

KEYWORDS

open source, survey, chatGPT, large language models, RLHF

ACM Reference Format:
Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3571884.3604316

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3604316

1 INTRODUCTION

Open research is the lifeblood of cumulative progress in science and engineering. In today's technological landscape, it is hard to find any research finding or technology that does not rely to a significant extent on the fruits of open research, often publicly funded. For instance, AlexNet [25], the deep neural net kickstarting the deep learning revolution a decade ago, derived its strength from a human-annotated dataset of 3.2 million images created by Princeton computer scientists [10, 14]. And the striking progress in protein folding in recent years (with the AlphaFold deep learning system predicting the structure of nearly all known proteins [53], where decades of prior work had reached a comparatively meagre 17%) has only been possible thanks to openly deposited structural data in the Protein Data Bank that goes back half a century [3].

The talk of the town in conversational interfaces today is undoubtedly ChatGPT, an instruction-tuned text generator that impresses many because of its fluid prose. Yet striking new capabilities should not detract us from the risks of proprietary systems. Only three months after OpenAI rolled out ChatGPT, it abruptly discontinued API support for its widely used Codex model that had been available as a "free limited beta" since 2021 [44] — surprising users with only three days' notice and undercutting at one blow the reproducibility of at least 100 research papers.[1] This is a stark reminder that proprietary systems are designed to offer smooth onboarding and convenience but come at the price of user lock-in and a lack of reliability.

Proprietary systems come with considerable further risks and harms [2, 9]. They tend to be developed without transparent ethical oversight, and are typically rolled out with profit motives that incentivise generating hype over enabling careful scientific work. They allow companies to mask exploitative labour practices, privacy implications [27] and murky copyright situations [49]. Today there is a growing division between global academia and the handful of firms who wield the computational resources required for training large language models. This "Compute Divide" [1] contributes to the growing de-democratisation of AI. Against this, working scientists call for avoiding the lure of proprietary models [51], for decolonizing the computational sciences [5], and for regulatory efforts to counteract harmful impacts [17].

1.1 Why openness matters

Open data is only one aspect of open research; open code, open models, open documentation, and open licenses are other crucial elements [8, 18]. Openness promotes transparency, reproducibility, and quality control; all features that are prequisites for supporting robust scientific inference [33] and building trustworthy AI [30]. Openness also allows critical use in research and teaching. For instance, it enables the painstaking labour of documenting ethical problems in existing datasets [7, 49], important work that can sometimes result in the retraction of such datasets [6]. In teaching, it can help foster critical computational literacy [29].

Despite strong evidence of the scientific and engineering benefits of open research practices, openness is not a given in machine learning and AI research [18, 20, 30]. Gundersen and Kjensmo, in one of the most detailed examinations of reproducibility in AI to date [19], systematically surveyed 400 papers for a range of open science practices. They found that only about a third of papers share test datasets, only 8% share source code, and only a single paper shared training, validation and test sets along with results. We are not aware of more recent systematic surveys of this kind (nor do we attempt this here), but the increasing trend of corporate releases with glossy blog posts replacing peer-reviewed scientific documentation provides little reason for optimism.

Openness is perhaps especially important for today's breed of instruction-following text generators, of which ChatGPT is the best known example. The persuasiveness of these language models is due in large part to an additional reinforcement learning component in which text generator output is pruned according to a reward function that is based on human feedback [12, 43, 59], using insights from early work on evaluative reinforcement [24, 26, 55]. Human users appear to be highly susceptible to the combination of interactivity and fluid text generation offered by this technology. The ubiquity of ChatGPT interfaces makes it easy for anyone today to try out some prompt engineering (while freely providing further training data to OpenAI) — but it does not allow one to gain a critical and holistic understanding of the constraints and capabilities of such systems, nor of their risks and harms. For true progress in this domain, we will need open alternatives.

In this paper, we survey alternatives to ChatGPT and assess them in terms of openness of data, models, documentation and access methods. The aim of our survey is threefold: to sketch some of the major dimensions along which it is useful to assess openness and transparency of large language models; to provide a view of the state of the art in open source instruction-tuned text generation; and to contribute towards a platform for tracking openness, transparency and accountability in this domain.

1.2 Previous work

Existing work reviewing and comparing large language models falls into two categories: informal lists and structured surveys. Informal lists are crowd-sourced pointers to available resources, from open RLHF datasets[2] to open examples of instruction-tuned text generators.[3] Systematic surveys of instruction-tuned language models are still rare and mostly focus on comparing model capabilities and performance, e.g., of "augmented language models" [37] and language models for writing code [58] (not our focus here). Complementary to our focus on degrees of openness in instruction-tuned models, a recent survey of generative AI systems more broadly focuses on gradience in release methods, from closed to staged to fully open [50].

An important development in this domain the introduction of data statements [34] and model cards [38]. These are structured documents that help creators document the process of curating, distributing and maintaining a dataset or model, and that help users to critically judge underlying assumptions, potential risks and harms, and potential for broader use. These resources have seen considerable uptake in the scientific community, though their adoption by for-profit entities lags behind.

[1] See aclanthology.org/search/?q=openai-davinci-002 (the same search term yields >150 arXiv preprints and >800 entries on Google Scholar)

[2] github.com/yaodongC/awesome-instruction-dataset

[3] github.com/nichtdax/awesome-totally-open-chatgpt

The risks of relying on proprietary solutions has spurred the development of several more open alternatives. For instance, the Bloom collaboration [56] is a team science project of unprecedented magnitude. It has trained and open-sourced a large language model based on a collection of almost 500 HuggingFace datasets amounting to 1.6TB of text and code in 46 spoken languages and 13 programming languages. [28, 39]. A related initiative is The Pile [16], a 800GB dataset of English text that serves as pre-training data for language models by EleutherAI [46]. Meta AI's LLaMA [52] provides researchers with access to a series of base models trained on data claimed to be 'publicly available'. It should be noted that none of these initiatives have undergone rigorous peer-review or data auditing at this point, and that claims of openness do not cancel out problems, legal or otherwise.

In recent years, the private company HuggingFace has emerged as an important hub in the open source community, bringing together developers and users of projects in machine learning and natural language processing. It offers infrastructure for hosting code, data, model cards, and demos [35]. It also provides a widely used setup for automated evaluation, generating leaderboards and allowing quick comparison on a number of automated metrics, making it somewhat of a balancing act between offering incentives for documentation and for SOTA-chasing [11]. Our focus here is not performance evaluation of the kind offered by leaderboards; instead it is to survey degrees of openness in the fast-evolving landscape of text generators.

2 METHOD

We survey open-source instruction-tuned text generators and evaluate them with regard to openness, scientific documentation, and access methods. Since any survey in this fast-growing field deals with moving targets, we focus here mainly on dimensions of enduring relevance for transparency and accountability. An up to date list of all models surveyed can be found at osf.io/d6fsr.

2.1 Requirements

The target breed of models in focus here is characterized by the following two features: its architecture is at base a large language model with reinforcement learning from human feedback (LLM + RLHF) and it aims for openness and transparency (along degrees we quantify). Projects are not included if they are as proprietary and undocumented as ChatGPT (like Google's Bard), or if they merely provide a front-end that calls some version of ChatGPT through an OpenAI API (like Microsoft's Bing). We explicitly include small-scale projects and projects that are in early stage development if they are open, sufficiently documented, and released under an open source license. Querying academic search engines and open code repositories, we find at least 15 projects that have sprung up in the last six months alone.

Availability	Documentation	Access methods
Open code	License	Package
LLM data	Code	API
LLM weights	Architecture	
RLHF data	Preprint	
RLHF weights	Paper	
	Data sheet	

Table 1: Overview of the 13 assessment features.

2.2 Survey elements

We assess projects on 13 features divided over three areas (Table 1): availability, documentation, and access methods. For each feature, we document openness along a scale from maximum to partial to no openness and transparency. For licenses, only systems that are fully covered by a true open-source licence count as maximally open, less permissive or partial licensing counts as partially open, and non-open or unclear licensing situations count as closed. Figure 1 shows a snapshot of 15 projects assessed for all features, with degrees of openness colour-coded (✓, ∼ , ×). Please refer to the data repository for more information about how each feature is evaluated, and for a more up to date listing.

3 RESULTS

Projects roughly fall into two categories. First, small, relatively bare bones projects that only provide source code and build on existing large language models. These projects often cannot share information on architecture, training data, and documentation because they inherit closed-source data from the LLMs they build on. They usually also do not provide APIs or other user interfaces. However, some of such small projects do come with high-quality documentation and some build only on explicitly open LLMs. What such small projects lack in performance, they make up in utility for the open source community as they can provide useful entry points to learning about LLM+RLHF tools.

We also identify a handful of projects backed by larger organisations, which aim to offer similar features to proprietary tools such as ChatGPT but are open-sourced and well documented. Two such initiatives top our list of open-source alternatives to ChatGPT: bigscience-workshop's xmtf tool building on the BLOOMZ and mT0 models (sponsored by HuggingFace) and LAION-AI's OpenAssistant based on an open, crowd-sourced RLHF training dataset (oasst1). Open Assistant also features a text-based and graphical user interface as well as a web resources for crowd-sourcing training data. We also found that several projects are not as open as they initially seemed to be, with many of them merely wrappers of closed models.

We observe three recurring issues in the area of availability and documentation. *Inheritance of undocumented data.* Many tools build on existing large language models (which we here call base models) and inherit the undocumented datasets (often web-scraped and often of dubious legality) these base models are trained on.

Training data of RLHF component is not shared. Building RLHF training datasets requires labour-intensive work by human annotators. The lack of RLHF training data is a major performance bottleneck for smaller research teams and organisations, and hampers reproducible research into the use of instruction-tuned text generators for conversational user interfaces.

Papers are rare, peer-review even rarer. Most projects reviewed here follow the corporate 'release by blog post' model. While there are some preprints, none of the systems we review is currently documented in a peer-reviewed paper. Habitually bypassing this important (albeit sometimes flawed) quality assurance mechanism allows systems to escape critical scrutiny and risks undermining scientific and ethical standards.

Some other patterns are worth noting. One is the rise of synthetic data especially for the instruction component. Prominent examples are Self-Instruct (derived from GPT3) [54], and Baize, a corpus generated by having ChatGPT engage in interaction with itself, seeded by human-generated questions scraped from online knowledge bases [57]. This stretches the definition of LLM + RLHF architectures because the reinforcement learning is no longer directly from human feedback but has a synthetic component, in effect parasitizing on the human labour encoded in source models. The consequences of using synthetic reinforcement learning data at scale are unknown and in need of close scrutiny.

The derivative nature of synthetic datasets is probably one reason they are released specifically "for research purposes only" [57], with commercial use strictly prohibited. This leads to an important wrinkle. Baize models and data are incorporated in several popular instruction-tuned text generators, including the Falcon family of models which bills itself as ready for "research and commercial utilization"[4] in direct violation of Baize's prohibition against commercial use. This is merely one example of the complex dependencies embedded in these tools, and the legal quagmires obscured by simple claims of 'openness'.

4 DISCUSSION

The goal of this short paper has been to provide a critical review of degrees of openness in the fast-moving field of instruction-tuned large language models. We have found projects at varying stages of implementation, documentation, and useability. Most of them offer access to source code and some aspects of pre-training data, sometimes in legally ambiguous ways. Data from the reinforcement learning step, crucial to the simulation of instruction-following in these interfaces, is more elusive, provided by at best half of the initiatives. Strikingly, only a handful of projects are underpinned by a scientific write-up and none of them have as yet undergone scientific peer review.

There are many shades of openness [50], yet all of the projects surveyed here are significantly more open than ChatGPT. ChatGPT was announced in a company blog post and rolled out to the public with an interface designed to capture as much free human labour as possible, but without any technical documentation. (The RLHF component, arguably the biggest differentiator for the instruction-following behavior, was sketched in [43], though without data.) Its follow-up GPT-4 continues OpenAI's tradition of openness in name only: it comes with an evaluation framework that primarily benefits the company yet contains the absolute minimum of technical documentation. In particular, an unreviewed preprint distributed by OpenAI and billed as a "technical report" [42] mostly provides cherry-picked examples and spends more space on crediting company workers for blog post content, communications, revenue, and legal advice than on actual technical details. (Companies like OpenAI sometimes give "AI safety" as a pretext for closedness; this is hard to take seriously when their own public-facing proprietary models provide clear and present harms [17].)

How can we foster more openness and accountability? First, incentives need changing. In high-stakes AI research, data work is often seen as low-level grunt work [48] and incentive structures generally encourage a 'move fast and break things' mentality over careful scientific work [47]. But work that documents data provenance and traces harmful impacts [4, 49] deserves major scholarly and societal credit. Here, AI and NLP might benefit from work in software engineering and infrastructure, where strong frameworks already exist to foster accountability for datasets [22, 31, 45]. Interactive model cards [13] offer a promising step towards a human-centered approach to documentation.

Second, corporate capture and user lock-in are well-known strategies by which companies exercise control over scientific results and research infrastructure. In the age of large language models, this is amplified by the possibility to extract human labour and repackage it in amiable conversational formats. Openness not only aligns with principles of sound and ethical scholarship [51]; it also safeguards transparent and reproducible research [40, 41]. Recent work on legal datasets offers an example in responsible data curation with insights that may be more broadly applicable [21].

Third, technology is never a *fait accompli* unless we make it so. It is one of the achievements of publicly funded science that it can afford to not jump on the bandwagon and instead make room for reflection [2, 5]. Today's language technology landscape offers ample opportunities for what philosopher Ivan Illich has called *counterfoil research*: "Counterfoil research must clarify and dramatize the relationship of people to their tools. It ought to hold constantly before the public the resources that are available and the consequences of their use in various ways. It should impress on people the existence of any trend that threatens one of the major balances of which life depends" [23]. Among the consequences of unleashing proprietary LLM + RLHF models are untold harms to workers exploited in labeling data; energy demands of computational resources [32]; and tidal waves of plausible-looking text generated without regard for truth value (technically, bullshit [15]).

One possible outcome of the kind of deeper understanding fostered by openness is a call for responsibly limited technology [23, 36]. The spectre of regulation (a key way to keep corporate powers in check) is a powerful incentive for companies to keep things proprietary and so shield them from scrutiny. The systems we have surveyed here provide elements of a solution. Open to various degrees, they provide ways to build reproducible workflows, chart resource costs, and lessen reliance on corporate whims.

5 CONCLUSION

Openness is not the full solution to the scientific and ethical challenges of conversational text generators. Open data will not mitigate the harmful consequences of thoughtless deployment of large language models, nor the questionable copyright implications of scraping all publicly available data from the internet. However, openness does make original research possible, including efforts to build reproducible workflows and understand the fundamentals of LLM + RLHF architectures. Openness also enables checks and balances, fostering a culture of accountability for data and its curation, and for models and their deployment. We hope that our work provides a small step in this direction.

[4]Technology Innovation Institute, https://falconllm.tii.ae/, June 7, 2023

ACKNOWLEDGMENTS

This research is funded by Dutch Research Council (NWO) grant 016.vidi.185.205 to MD. For the purpose of Open Access the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] Nur Ahmed, Muntasir Wahed, and Neil C. Thompson. 2023. The growing influence of industry in AI research. *Science* 379, 6635 (March 2023), 884–886. https://doi.org/10.1126/science.ade2420

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Frances C. Bernstein, Thomas F. Koetzle, Graheme J. B. Williams, Edgar F. Meyer, Michael D. Brice, John R. Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112, 3 (May 1977), 535–542. https://doi.org/10.1016/S0022-2836(77)80200-3

[4] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (Feb. 2021), 100205. https://doi.org/10.1016/j.patter.2021.100205

[5] Abeba Birhane and Olivia Guest. 2021. Towards Decolonising Computational Sciences. *Kvinder, Køn & Forskning* 2 (2021), 60–73. https://doi.org/10.7146/kkf.v29i2.124899

[6] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158 ISSN: 2642-9381.

[7] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. https://doi.org/10.48550/arXiv.2110.01963 arXiv:2110.01963 [cs].

[8] Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data* 2 (2019). https://www.frontiersin.org/articles/10.3389/fdata.2019.00043

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. https://doi.org/10.48550/arXiv.2012.07805 arXiv:2012.07805 [cs].

[10] Sanjay Chawla, Preslav Nakov, Ahmed Ali, Wendy Hall, Issa Khalil, Xiaosong Ma, Husrev Taha Sencar, Ingmar Weber, Michael Wooldridge, and Ting Yu. 2023. Ten years after ImageNet: a 360° perspective on artificial intelligence. *Royal Society Open Science* 10, 3 (March 2023), 221414. https://doi.org/10.1098/rsos.221414 Publisher: Royal Society.

[11] Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering* 28, 2 (March 2022), 249–269. https://doi.org/10.1017/S1351324922000043 Publisher: Cambridge University Press.

[12] Deborah Cohen, Moonkyung Ryu, Yinlam Chow, Orgad Keller, Ido Greenberg, Avinatan Hassidim, Michael Fink, Yossi Matias, Idan Szpektor, Craig Boutilier, and Gal Elidan. 2022. Dynamic Planning in Open-Ended Dialogue using Reinforcement Learning. https://doi.org/10.48550/arXiv.2208.02294 arXiv:2208.02294 [cs].

[13] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. https://doi.org/10.1145/3531146.3533108

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[15] Harry G. Frankfurt. 2009. *On Bullshit*. Princeton University Press. https://doi.org/10.1515/9781400826537

[16] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. https://doi.org/10.48550/arXiv.2101.00027 arXiv:2101.00027 [cs].

[17] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. 2023. Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. https://www.dair-institute.org/blog/letter-statement-March2023

[18] Odd Erik Gundersen, Yolanda Gil, and David W. Aha. 2018. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine* 39, 3 (Sept. 2018), 56–68. https://doi.org/10.1609/aimag.v39i3.2816 Number: 3.

[19] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11503 Number: 1.

[20] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J. W. L. Aerts. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (Oct. 2020), E14–E16. https://doi.org/10.1038/s41586-020-2766-y Number: 7829 Publisher: Nature Publishing Group.

[21] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 29217–29234. https://proceedings.neurips.cc/paper_files/paper/2022/hash/bc218a0c656e49d4b086975a9c785f47-Abstract-Datasets_and_Benchmarks.html

[22] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[23] Ivan Illich. 1973. *Tools for conviviality.* Harper & Row, New York.

[24] W. Bradley Knox and Peter Stone. 2008. TAMER: Training an Agent Manually via Evaluative Reinforcement. In *2008 7th IEEE International Conference on Development and Learning*. 292–297. https://doi.org/10.1109/DEVLRN.2008.4640845 ISSN: 2161-9476.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90. https://doi.org/10.1145/3065386

[26] Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face Blog* (2022).

[27] Martha Larson, Nelleke Oostdijk, and Frederik Zuiderveen Borgesius. 2021. Not Directly Stated, Not Explicitly Stored: Conversational Agents and the Privacy Threat of Implicit Information. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 388–391. https://doi.org/10.1145/3450614.3463601

[28] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. https://openreview.net/forum?id=UoEw6KigkUn

[29] Clifford H. Lee and Elisabeth Soep. 2016. None But Ourselves Can Free Our Minds: Critical Computational Literacy as a Pedagogy of Resistance. *Equity & Excellence in Education* 49, 4 (Oct. 2016), 480–492. https://doi.org/10.1080/10665684.2016.1227157

[30] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *Comput. Surveys* 55, 9 (Jan. 2023), 177:1–177:46. https://doi.org/10.1145/3555803

[31] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (Aug. 2022), 669–677. https://doi.org/10.1038/s42256-022-00516-1 Number: 8 Publisher: Nature Publishing Group.

[32] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. http://arxiv.org/abs/2211.02001 arXiv:2211.02001 [cs].

[33] Erin C. McKiernan, Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A. Nosek, Karthik Ram, Courtney K. Soderberg, Jeffrey R. Spies, Kaitlin Thaney, Andrew Updegrove, Kara H. Woo, and Tal Yarkoni. 2016. How open science helps researchers succeed. *eLife* 5 (July 2016), e16800. https://doi.org/10.7554/eLife.16800

[34] Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing* (2023). https://doi.org/10.1145/3594737 Just Accepted.

[35] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics, Online, 121–135. https://doi.org/10.18653/v1/2021.gem-1.11

[36] Dan McQuillan. 2022. *Resisting AI: an anti-fascist approach to artificial intelligence.* Bristol University Press, Bristol, UK. OCLC: on1328026349.

[37] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. https://doi.org/10.48550/arXiv.2302.07842 arXiv:2302.07842 [cs].

[38] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[39] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual Generalization through Multitask Finetuning. https://doi.org/10.48550/arXiv.2211.01786 arXiv:2211.01786 [cs].

[40] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3517644

[41] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering (ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 413–425. https://doi.org/10.1145/3510003.3510209

[42] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[43] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. (2022), 68.

[44] Mohit Pandey. 2023. OpenAI Might Invite Legal Trouble. *Analytics India Magazine* (March 2023). https://analyticsindiamag.com/openai-might-invite-legal-trouble/

[45] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[46] Jason Phang, Herbie Bradley, Leo Gao, Louis Castricato, and Stella Biderman. 2022. EleutherAI: Going Beyond "Open Science" to "Science in the Open". https://doi.org/10.48550/arXiv.2210.06413 arXiv:2210.06413 [cs].

[47] Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2182–2194. https://doi.org/10.18653/v1/2021.acl-long.170

[48] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

[49] Kevin Schaul, Szu Yu Chen, and Nitasha Tiku. 2023. Inside the secret list of websites that make AI like ChatGPT sound smart. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[50] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/3593013.3593981

[51] Arthur Spirling. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature* 616, 7957 (April 2023), 413–413. https://doi.org/10.1038/d41586-023-01295-4 Bandiera_abtest: a Cg_type: World View Number: 7957 Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community.

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971 arXiv:2302.13971 [cs].

[53] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596, 7873 (Aug. 2021), 590–596. https://doi.org/10.1038/s41586-021-03828-1 Number: 7873 Publisher: Nature Publishing Group.

[54] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. https://doi.org/10.48550/arXiv.2212.10560 arXiv:2212.10560 [cs].

[55] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11485 Number: 1.

[56] BigScience Workshop. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. https://doi.org/10.48550/arXiv.2211.05100

[57] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. https://doi.org/10.48550/arXiv.2304.01196 arXiv:2304.01196 [cs].

[58] Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS 2022)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3520312.3534862

[59] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. https://doi.org/10.48550/arXiv.1909.08593 arXiv:1909.08593 [cs, stat].

Received 20 April 2023; accepted 26 May 2023