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

Andreas Liesenfeld  
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

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 all the 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
1 INTRODUCTION
Open research is the lifeline 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 [33], 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 Kjemsnmo, 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; 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. 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 arXiv.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/nichtdac/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.

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 barebones 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 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 xmfn tool building on the BLOOMZ and mT0 models (sponsored by HuggingFace) and LAION-ALI’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" 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.