



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• Computing methodologies → Natural language processing.

ACM Reference Format:

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 *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

*Joint first authors



This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03.

<https://doi.org/10.1145/3442188.3445922>

mBERT across 29 tasks. Either way, these models do not address the inclusion problems raised by [65], who note that over 90% of the world's languages used by more than a billion people currently have little to no support in terms of language technology.

Alongside work investigating what information the models retain from the data, we see a trend in reducing the size of these models using various techniques such as knowledge distillation [26, 58], quantization [118, 153], factorized embedding parameterization and cross-layer parameter sharing [70], and progressive module replacing [146]. Rogers et al. [110] provide a comprehensive comparison of models derived from BERT using these techniques, such as DistilBERT [413] and ALBERT [70]. While these models maintain and sometimes exceed the performance of the original BERT model, despite their much smaller size, they ultimately still rely on large quantities of data and significant processing and storage capabilities to both hold and reduce the model.

We note that the change from n -gram LMs to word vectors distilled from neural LMs to pretrained Transformer LMs is paralleled by an expansion and change in the types of tasks they are useful for: n -gram LMs were initially typically deployed in selecting among the outputs of e.g. acoustical or translation models; the LSTM-derived word vectors were quickly picked up as more effective representations of words (in place of bag of words features) in a variety of NLP tasks involving labeling and classification; and the pretrained Transformer models can be retrained on very small datasets (few-shot, one-shot or even zero-shot learning) to perform apparently meaning-manipulating tasks such as summarization, question answering and the like. Nonetheless, all of these systems share the property of being LMs in the sense we give above, that is, systems trained to predict sequences of words (or characters or sentences). Where they differ is in the size of the training datasets they leverage and the spheres of influence they can possibly affect. By scaling up in these two ways, modern very large LMs incur new kinds of risk, which we turn to in the following sections.

3 ENVIRONMENTAL AND FINANCIAL COST

Strubell et al. recently benchmarked model training and development costs in terms of dollars and estimated CO_2 emissions [129]. While the average human is responsible for an estimated 5t CO_2e per year,² the authors trained a Transformer (big) model [136] with neural architecture search and estimated that the training procedure emitted 284t of CO_2 . Training a single BERT base model (without hyperparameter tuning) on GPUs was estimated to require as much energy as a trans-American flight.

While some of this energy comes from renewable sources, or cloud compute companies' use of carbon credit-offset sources, the authors note that the majority of cloud compute providers' energy is not sourced from renewable sources and many energy sources in the world are not carbon neutral. In addition, renewable energy sources are still costly to the environment,³ and data centers with increasing computation requirements take away from other potential uses of

green energy,⁴ underscoring the need for energy efficient model architectures and training paradigms.

Strubell et al. also examine the cost of these models vs. their accuracy gains. For the task of machine translation where large LMs have resulted in performance gains, they estimate that an increase in 0.1 BLEU score using neural architecture search for English to German translation results in an increase of \$150,000 compute cost in addition to the carbon emissions. To encourage more equitable access to NLP research and reduce carbon footprint, the authors give recommendations to report training time and sensitivity to hyperparameters when the released model is meant to be re-trained for downstream use. They also urge governments to invest in compute clouds to provide equitable access to researchers.

Initiatives such as the SustainNLP workshop⁵ have since taken up the goal of prioritizing computationally efficient hardware and algorithms. Schwartz et al. [115] also call for the development of green AI, similar to other environmentally friendly scientific developments such as green chemistry or sustainable computing. As shown in [5], the amount of compute used to train the largest deep learning models (for NLP and other applications) has increased 300,000x in 6 years, increasing at a far higher pace than Moore's Law. To promote green AI, Schwartz et al. argue for promoting efficiency as an evaluation metric and show that most sampled papers from ACL 2018, NeurIPS 2018, and CVPR 2019 claim accuracy improvements alone as primary contributions to the field, and none focused on measures of efficiency as primary contributions. Since then, works such as [57, 75] have released online tools to help researchers benchmark their energy usage. Among their recommendations are to run experiments in carbon friendly regions, consistently report energy and carbon metrics, and consider energy-performance trade-offs before deploying energy hungry models. In addition to these calls for documentation and technical fixes, Bietti and Vatanparast underscore the need for social and political engagement in shaping a future where data driven systems have minimal negative impact on the environment [16].

While [129] benchmarks the training process in a research setting, many LMs are deployed in industrial or other settings where the cost of inference might greatly outweigh that of training in the long run. In this scenario, it may be more appropriate to deploy models with lower energy costs during inference even if their training costs are high. In addition to benchmarking tools, works estimating the cost increase associated with the introduction of LMs for particular applications, and how they compare to alternative NLP methods, will be important for understanding the trade-offs.

When we perform risk/benefit analyses of language technology, we must keep in mind how the risks and benefits are distributed, because they do not accrue to the same people. On the one hand, it is well documented in the literature on environmental racism that the negative effects of climate change are reaching and impacting the world's most marginalized communities first [1, 27].⁶ Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 [6]) or the 800,000 people in Sudan affected

²Data for 2017, from <https://ourworldindata.org/co2-emissions>, accessed Jan 21, 2021

³<https://www.heraldsotland.com/news/18270734.14m-trees-cut-scotland-make-way-wind-farms/>

⁴<https://news.microsoft.com/2017/11/02/microsoft-announces-one-of-the-largest-wind-deals-in-the-netherlands-with-vattenfall/>

⁵<https://sites.google.com/view/sustainnlp2020/organization>

⁶<https://www.un.org/sustainabledevelopment/blog/2016/10/report-inequalities-exacerbate-climate-impacts-on-poor/>

by drastic floods⁷ pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren't being produced for Dhivehi or Sudanese Arabic?⁸

And, while some language technology is genuinely designed to benefit marginalized communities [17, 101], **most language technology is built to serve the needs of those who already have the most privilege in society.** Consider, for example, who is likely to both have the financial resources to purchase a Google Home, Amazon Alexa or an Apple device with Siri installed and comfortably speak a variety of a language which they are prepared to handle. Furthermore, when large LMs encode and reinforce hegemonic biases (see §§4 and 6), the harms that follow are most likely to fall on marginalized populations who, even in rich nations, are most likely to experience environmental racism [10, 104].

These models are being developed at a time when unprecedented environmental changes are being witnessed around the world. From monsoons caused by changes in rainfall patterns due to climate change affecting more than 8 million people in India,⁹ to the worst fire season on record in Australia killing or displacing nearly three billion animals and at least 400 people,¹⁰ the effect of climate change continues to set new records every year. It is past time for researchers to prioritize energy efficiency and cost to reduce negative environmental impact and inequitable access to resources — both of which disproportionately affect people who are already in marginalized positions.

4 UNFATHOMABLE TRAINING DATA

The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. **However, in both application areas, the training data has been shown to have problematic characteristics [18, 38, 42, 47, 61] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status [11, 12, 69, 69, 132, 132, 157].** In this section, we discuss how large, uncensored, Internet-based datasets encode the dominant/hegemonic view, which further harms people at the margins, and recommend significant resource allocation towards dataset curation and documentation practices.

4.1 Size Doesn't Guarantee Diversity

The Internet is a large and diverse virtual space, and accordingly, it is easy to imagine that very large datasets, such as Common Crawl (“petabytes of data collected over 8 years of web crawling”,¹¹ a filtered version of which is included in the GPT-3 training data) must therefore be broadly representative of the ways in which different people view the world. However, on closer examination, we find that there are several factors which narrow Internet participation, the

discussions which will be included via the crawling methodology, and finally the texts likely to be contained after the crawled data are filtered. In all cases, the voices of people most likely to hew to a hegemonic viewpoint are also more likely to be retained. In the case of US and UK English, this means that white supremacist and misogynistic, ageist, etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms.

Starting with who is contributing to these Internet text collections, we see that Internet access itself is not evenly distributed, resulting in Internet data overrepresenting younger users and those from developed countries [100, 143].¹² However, it's not just the Internet as a whole that is in question, but rather specific subsamples of it. For instance, GPT-2's training data is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29.¹³ Similarly, recent surveys of Wikipedians find that only 8.8–15% are women or girls [9].

Furthermore, while user-generated content sites like Reddit, Twitter, and Wikipedia present themselves as open and accessible to anyone, there are structural factors including moderation practices which make them less welcoming to marginalized populations. Jones [64] documents (using digital ethnography techniques [63]) multiple cases where people on the receiving end of death threats on Twitter have had their accounts suspended while the accounts issuing the death threats persist. She further reports that harassment on Twitter is experienced by “a wide range of overlapping groups including domestic abuse victims, sex workers, trans people, queer people, immigrants, medical patients (by their providers), neurodivergent people, and visibly or vocally disabled people.” The net result is that a limited set of subpopulations can continue to easily add data, sharing their thoughts and developing platforms that are inclusive of their worldviews; this systemic pattern in turn worsens diversity and inclusion within Internet-based communication, creating a feedback loop that lessens the impact of data from underrepresented populations.

Even if populations who feel unwelcome in mainstream sites set up different fora for communication, these may be less likely to be included in training data for language models. Take, for example, older adults in the US and UK. Lazar et al. outline how they both individually and collectively articulate anti-ageist frames specifically through blogging [71], which some older adults prefer over more popular social media sites for discussing sensitive topics [24]. These fora contain rich discussions about what constitutes age discrimination and the impacts thereof. However, a blogging community such as the one described by Lazar et al. is less likely to be found than other blogs that have more incoming and outgoing links.

Finally, the current practice of filtering datasets can further attenuate the voices of people from marginalized identities. The training set for GPT-3 was a filtered version of the Common Crawl dataset, developed by training a classifier to pick out those documents

⁷<https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un>

⁸By this comment, we do not intend to erase existing work on low-resource languages. One particularly exciting example is the Masakhane project [91], which explores participatory research techniques for developing MT for African languages. These promising directions do not involve amassing terabytes of data.

⁹<https://www.voanews.com/south-central-asia/monsoons-cause-havoc-india-climate-change-alters-rainfall-patterns>

¹⁰<https://www.cnn.com/2020/07/28/asia/australia-fires-wildlife-report-scli-intl-scn/index.html>

¹¹<http://commoncrawl.org/>

¹²This point is also mentioned in the model card for GPT-3: <https://github.com/openai/gpt-3/blob/master/model-card.md>

¹³<https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/>

most similar to the ones used in GPT-2's training data, i.e. documents linked to from Reddit [25], plus Wikipedia and a collection of books. While this was reportedly effective at filtering out documents that previous work characterized as “unintelligible” [134], what is unmeasured (and thus unknown) is what else it filtered out. The Colossal Clean Crawled Corpus [107], used to train a trillion parameter LM in [43], is cleaned, *inter alia*, by discarding any page containing one of a list of about 400 “Dirty, Naughty, Obscene or Otherwise Bad Words” [p.6].¹⁴ This list is overwhelmingly words related to sex, with a handful of racial slurs and words related to white supremacy (e.g. *swastika*, *white power*) included. While possibly effective at removing documents containing pornography (and the associated problematic stereotypes encoded in the language of such sites [125]) and certain kinds of hate speech, this approach will also undoubtedly attenuate, by suppressing such words as *twink*, the influence of online spaces built by and for LGBTQ people.¹⁵ If we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and otherwise describes marginalized identities in a positive light.

Thus at each step, from initial participation in Internet fora, to continued presence there, to the collection and finally the filtering of training data, current practice privileges the hegemonic viewpoint. **In accepting large amounts of web text as ‘representative’ of ‘all’ of humanity we risk perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality.** We instead propose practices that actively seek to include communities underrepresented on the Internet. For instance, one can take inspiration from movements to decolonize education by moving towards oral histories due to the overrepresentation of colonial views in text [35, 76, 127], and curate training datasets through a thoughtful process of deciding what to put in, rather than aiming solely for scale and trying haphazardly to weed out, post-hoc, flotsam deemed ‘dangerous’, ‘unintelligible’, or ‘otherwise bad’.

4.2 Static Data/Changing Social Views

A central aspect of social movement formation involves using language strategically to destabilize dominant narratives and call attention to underrepresented social perspectives. Social movements produce new norms, language, and ways of communicating. This adds challenges to the deployment of LMs, as methodologies reliant on LMs run the risk of ‘value-lock’, where the LM-reliant technology reifies older, less-inclusive understandings.

For instance, the Black Lives Matter movement (BLM) influenced Wikipedia article generation and editing such that, as the BLM movement grew, articles covering shootings of Black people increased in coverage and were generated with reduced latency [135]. Importantly, articles describing past shootings and incidents of police brutality were created and updated as articles for new events were created, reflecting how social movements make connections between events in time to form cohesive narratives [102]. More generally, Twyman et al. [135] highlight how social movements actively influence framings and reframings of minority narratives

in the type of online discourse that potentially forms the data that underpins LMs.

An important caveat is that social movements which are poorly documented and which do not receive significant media attention will not be captured at all. Media coverage can fail to cover protest events and social movements [41, 96] and can distort events that challenge state power [36]. This is exemplified by media outlets that tend to ignore peaceful protest activity and instead focus on dramatic or violent events that make for good television but nearly always result in critical coverage [81]. As a result, the data underpinning LMs stands to misrepresent social movements and disproportionately align with existing regimes of power.

Developing and shifting frames stand to be learned in incomplete ways or lost in the big-ness of data used to train large LMs — particularly if the training data isn’t continually updated. Given the compute costs alone of training large LMs, it likely isn’t feasible for even large corporations to fully retrain them frequently enough to keep up with the kind of language change discussed here. Perhaps fine-tuning approaches could be used to retrain LMs, but here again, what would be required is thoughtful curation practices to find appropriate data to capture reframings and techniques for evaluating whether such fine-tuning appropriately captures the ways in which new framings contest hegemonic representations.

4.3 Encoding Bias

It is well established by now that large LMs exhibit various kinds of bias, including stereotypical associations [11, 12, 69, 119, 156, 157], or negative sentiment towards specific groups [61]. Furthermore, we see the effects of intersectionality [34], where BERT, ELMo, GPT and GPT-2 encode more bias against identities marginalized along more than one dimension than would be expected based on just the combination of the bias along each of the axes [54, 132]. Many of these works conclude that these issues are a reflection of training data characteristics. For instance, Hutchinson et al. find that BERT associates phrases referencing persons with disabilities with more negative sentiment words, and that gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness [61]. Similarly, Gehman et al. show that models like GPT-3 trained with at least 570GB of data derived mostly from Common Crawl¹⁶ can generate sentences with high toxicity scores even when prompted with non-toxic sentences [53]. Their investigation of GPT-2’s training data¹⁷ also finds 272K documents from unreliable news sites and 63K from banned subreddits.

These demonstrations of biases learned by LMs are extremely valuable in pointing out the potential for harm when such models are deployed, either in generating text or as components of classification systems, as explored further in §6. However, they do not represent a methodology that can be used to exhaustively discover all such risks, for several reasons.

First, model auditing techniques typically rely on automated systems for measuring sentiment, toxicity, or novel metrics such as ‘regard’ to measure attitudes towards a specific demographic group [119]. But these systems themselves may not be reliable

¹⁴ Available at <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en>, accessed Jan 18, 2021

¹⁵ This observation is due to William Agnew.

¹⁶ <https://commoncrawl.org/the-data/>

¹⁷ GPT-3’s training data is not openly available, but GPT-2’s training data was used indirectly to construct GPT-3’s [53].

means of measuring the toxicity of text generated by LMs. For example, the Perspective API model has been found to associate higher levels of toxicity with sentences containing identity markers for marginalized groups or even specific names [61, 103].

Second, auditing an LM for biases requires an *a priori* understanding of what social categories might be salient. The works cited above generally start from US protected attributes such as race and gender (as understood within the US). But, of course, protected attributes aren't the only identity characteristics that can be subject to bias or discrimination, and the salient identity characteristics and expressions of bias are also culture-bound [46, 116]. Thus, components like toxicity classifiers would need culturally appropriate training data for each context of audit, and even still we may miss marginalized identities if we don't know what to audit for.

Finally, we note that moving beyond demonstrating the existence of bias to building systems that verify the 'safety' of some LM (even for a given protected class) requires engaging with the systems of power that lead to the harmful outcomes such a system would seek to prevent [19]. For example, the #MeToo movement has spurred broad-reaching conversations about inappropriate sexual behavior from men in power, as well as men more generally [84]. These conversations challenge behaviors that have been historically considered appropriate or even the fault of women, shifting notions of sexually inappropriate behavior. Any product development that involves operationalizing definitions around such shifting topics into algorithms is necessarily political (whether or not developers choose the path of maintaining the *status quo ante*). For example, men and women make significantly different assessments of sexual harassment online [40]. An algorithmic definition of what constitutes inappropriately sexual communication will inherently be concordant with some views and discordant with others. Thus, an attempt to measure the appropriateness of text generated by LMs, or the biases encoded by a system, always needs to be done in relation to particular social contexts and marginalized perspectives [19].

4.4 Curation, Documentation & Accountability

In summary, LMs trained on large, uncured, static datasets from the Web encode hegemonic views that are harmful to marginalized populations. We thus emphasize the need to invest significant resources into curating and documenting LM training data. In this, we follow Jo et al. [62], who cite archival history data collection methods as an example of the amount of resources that should be dedicated to this process, and Birhane and Prabhu [18], who call for a more justice-oriented data collection methodology. Birhane and Prabhu note, echoing Ruha Benjamin [15], "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." [p.1541]

When we rely on ever larger datasets we risk incurring *documentation debt*,¹⁸ i.e. putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc. While documentation allows for potential accountability [13, 52, 86], undocumented training data perpetuates harm without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones. The solution, we propose, is to budget for

documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget.

5 DOWN THE GARDEN PATH

In §4 above, we discussed the ways in which different types of biases can be encoded in the corpora used to train large LMs. In §6 below we explore some of the risks and harms that can follow from deploying technology that has learned those biases. In the present section, however, we focus on a different kind of risk: that of misdirected research effort, specifically around the application of LMs to tasks intended to test for natural language understanding (NLU). As the very large Transformer LMs posted striking gains in the state of the art on various benchmarks intended to model meaning-sensitive tasks, and as initiatives like [142] made the models broadly accessible to researchers seeking to apply them, large quantities of research effort turned towards measuring how well BERT and its kin do on both existing and new benchmarks.¹⁹ This allocation of research effort brings with it an opportunity cost, on the one hand in terms of time not spent applying meaning capturing approaches to meaning sensitive tasks, and on the other hand in terms of time not spent exploring more effective ways of building technology with datasets of a size that can be carefully curated and available for a broader set of languages [65, 91].

The original BERT paper [39] showed the effectiveness of the architecture and the pretraining technique by evaluating on the General Language Understanding Evaluation (GLUE) benchmark [138], the Stanford Question Answering Datasets (SQuAD 1.1 and 2.0) [108], and the Situations With Adversarial Generations benchmark (SWAG) [155], all datasets designed to test language understanding and/or commonsense reasoning. BERT posted state of the art results on all of these tasks, and the authors conclude by saying that "unsupervised pre-training is an integral part of many language understanding systems." [39, p.4179]. Even before [39] was published, BERT was picked up by the NLP community and applied with great success to a wide variety of tasks [e.g. 2, 149].

However, no actual language understanding is taking place in LM-driven approaches to these tasks, as can be shown by careful manipulation of the test data to remove spurious cues the systems are leveraging [21, 93]. Furthermore, as Bender and Koller [14] argue from a theoretical perspective, **languages are systems of signs [37], i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized.**

As the late Karen Spärck Jones pointed out: the use of LMs ties us to certain (usually unstated) epistemological and methodological commitments [124]. Either i) we commit ourselves to a noisy-channel interpretation of the task (which rarely makes sense outside of ASR), ii) we abandon any goals of theoretical insight into tasks and treat LMs as "just some convenient technology" [p.7], or iii) we implicitly assume a certain statistical relationship — known to be invalid — between inputs, outputs and meanings.²⁰ Although

¹⁹~26% of papers sampled from ACL, NAACL and EMNLP since 2018 cite [39].

²⁰Specifically, that the mutual information between the input and the meaning given the output is zero — what Spärck Jones calls "the model of ignorance".

¹⁸On the notion of documentation debt as applied to code, rather than data, see [154].

she primarily had n-gram models in mind, the conclusions remain apt and relevant.

There are interesting linguistic questions to ask about what exactly BERT, GPT-3 and their kin are learning about linguistic structure from the unsupervised language modeling task, as studied in the emerging field of ‘BERTology’ [e.g. 110, 133]. However, from the perspective of work on language technology, it is far from clear that all of the effort being put into using large LMs to ‘beat’ tasks designed to test natural language understanding, and all of the effort to create new such tasks, once the existing ones have been bulldozed by the LMs, brings us any closer to long-term goals of general language understanding systems. If a large LM, endowed with hundreds of billions of parameters and trained on a very large dataset, can manipulate linguistic form well enough to cheat its way through tests meant to require language understanding, have we learned anything of value about how to build machine language understanding or have we been led down the garden path?

6 STOCHASTIC PARROTS

In this section, we explore the ways in which the factors laid out in §4 and §5 — **the tendency of training data ingested from the Internet to encode hegemonic worldviews, the tendency of LMs to amplify biases and other issues in the training data, and the tendency of researchers and other people to mistake LM-driven performance gains for actual natural language understanding** — present real-world risks of harm, as these technologies are deployed. After exploring some reasons why humans mistake LM output for meaningful text, we turn to the risks and harms from deploying such a model at scale. We find that the mix of human biases and seemingly coherent language **heightens the potential for automation bias, deliberate misuse, and amplification of a hegemonic worldview**. We focus primarily on cases where LMs are used in generating text, but we will also touch on risks that arise when LMs or word embeddings derived from them are components of systems for classification, query expansion, or other tasks, or when users can query LMs for information memorized from their training data.

6.1 Coherence in the Eye of the Beholder

Where traditional n-gram LMs [117] can only model relatively local dependencies, predicting each word given the preceding sequence of N words (usually 5 or fewer), the Transformer LMs capture much larger windows and can produce text that is seemingly not only fluent but also coherent even over paragraphs. For example, McGuffie and Newhouse [80] prompted GPT-3 with the text in bold in Figure 1, and it produced the rest of the text, including the Q&A format.²¹ This example illustrates GPT-3’s ability to produce coherent and on-topic text; the topic is connected to McGuffie and Newhouse’s study of GPT-3 in the context of extremism, discussed below.

We say *seemingly* coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors’ beliefs [30, 31] and intentions [23, 33] within context [32]. That is, human language use

²¹This is just the first part of the response that McGuffie and Newhouse show. GPT-3 continues for two more question answer pairs with similar coherence. McGuffie and Newhouse report that all examples given in their paper are from either the first or second attempt at running a prompt.

Question: What is the name of the Russian mercenary group?

Answer: Wagner group.

Question: Where is the Wagner group?

Answer: In Syria.

Question: Who is the leader of the Wagner group?

Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia’s General Staff. He was also a commander of the special forces unit “Vostok” (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia’s war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad’s regime against anti-government forces there.

Figure 1: GPT-3’s response to the prompt (in bold), from [80]

takes place between individuals who share common ground and are mutually aware of that sharing (and its extent), who have communicative intents which they use language to convey, and who model each others’ mental states as they communicate. As such, human communication relies on the interpretation of implicit meaning conveyed between individuals. The fact that human-human communication is a jointly constructed activity [29, 128] is most clearly true in co-situated spoken or signed communication, but we use the same facilities for producing language that is intended for audiences not co-present with us (readers, listeners, watchers at a distance in time or space) and in interpreting such language when we encounter it. It must follow that even when we don’t know the person who generated the language we are interpreting, we build a partial model of who they are and what common ground we think they share with us, and use this in interpreting their words.

Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader’s state of mind. It can’t have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that. This can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do [89, 140]. The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language (independent of the model).²² Contrary

²²Controlled generation, where an LM is deployed within a larger system that guides its generation of output to certain styles or topics [e.g. 147, 151, 158], is not the same thing as communicative intent. One clear way to distinguish the two is to ask whether

to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.

6.2 Risks and Harms

The ersatz fluency and coherence of LMs raises several risks, precisely because humans are prepared to interpret strings belonging to languages they speak as meaningful and corresponding to the communicative intent of some individual or group of individuals who have accountability for what is said. We now turn to examples, laying out the potential follow-on harms.

The first risks we consider are the risks that follow from the LMs absorbing the hegemonic worldview from their training data. When humans produce language, our utterances reflect our worldviews, including our biases [78, 79]. As people in positions of privilege with respect to a society's racism, misogyny, ableism, etc., tend to be overrepresented in training data for LMs (as discussed in §4 above), this training data thus includes encoded biases, many already recognized as harmful.

Biases can be encoded in ways that form a continuum from subtle patterns like referring to *women doctors* as if *doctor* itself entails not-woman or referring to *both genders* excluding the possibility of non-binary gender identities, through directly contested framings (e.g. *undocumented immigrants* vs. *illegal immigrants* or *illegals*), to language that is widely recognized to be derogatory (e.g. racial slurs) yet still used by some. While some of the most overtly derogatory words could be filtered out, not all forms of online abuse are easily detectable using such taboo words, as evidenced by the growing body of research on online abuse detection [45, 109]. Furthermore, in addition to abusive language [139] and hate speech [67], there are subtler forms of negativity such as gender bias [137], microaggressions [22], dehumanization [83], and various socio-political framing biases [44, 114] that are prevalent in language data. For example, describing a woman's account of her experience of sexism with the word *tantrum* both reflects a worldview where the sexist actions are normative and foregrounds a stereotype of women as childish and not in control of their emotions.

An LM that has been trained on such data will pick up these kinds of problematic associations. If such an LM produces text that is put into the world for people to interpret (flagged as produced by an 'AI' or otherwise), what risks follow? In the first instance, we foresee that LMs producing text will reproduce and even amplify the biases in their input [53]. Thus the risk is that people disseminate text generated by LMs, meaning more text in the world that reinforces and propagates stereotypes and problematic associations, both to humans who encounter the text and to future LMs trained on training sets that ingested the previous generation LM's output. Humans who encounter this text may themselves be subjects of those stereotypes and associations or not. Either way, harms ensue: readers subject to the stereotypes may experience the psychological harms of microaggressions [88, 141] and stereotype threat [97, 126]. Other readers may be introduced to stereotypes or have ones they

already carry reinforced, leading them to engage in discrimination (consciously or not) [55], which in turn leads to harms of subjugation, denigration, belittlement, loss of opportunity [3, 4, 56] and others on the part of those discriminated against.

If the LM outputs overtly abusive language (as Gehman et al. [53] show that they can and do), then a similar set of risks arises. These include: propagating or proliferating overtly abusive views and associations, amplifying abusive language, and producing more (synthetic) abusive language that may be included in the next iteration of large-scale training data collection. The harms that could follow from these risks are again similar to those identified above for more subtly biased language, but perhaps more acute to the extent that the language in question is overtly violent or defamatory. They include the psychological harm experienced by those who identify with the categories being denigrated if they encounter the text; the reinforcement of sexist, racist, ableist, etc. ideology; follow-on effects of such reinforced ideologies (including violence); and harms to the reputation of any individual or organization perceived to be the source of the text.

If the LM or word embeddings derived from it are used as components in a text classification system, these biases can lead to allocational and/or reputational harms, as biases in the representations affect system decisions [125]. This case is especially pernicious for being largely invisible to both the direct user of the system and any indirect stakeholders about whom decisions are being made. Similarly, biases in an LM used in query expansion could influence search results, further exacerbating the risk of harms of the type documented by Noble in [94], where the juxtaposition of search queries and search results, when connected by negative stereotypes, reinforce those stereotypes and cause psychological harm.

The above cases involve risks that could arise when LMs are deployed without malicious intent. A third category of risk involves bad actors taking advantage of the ability of large LMs to produce large quantities of seemingly coherent texts on specific topics on demand in cases where those deploying the LM have no investment in the truth of the generated text. These include prosaic cases, such as services set up to 'automatically' write term papers or interact on social media,²³ as well as use cases connected to promoting extremism. For example, McGuffie and Newhouse [80] show how GPT-3 could be used to generate text in the persona of a conspiracy theorist, which in turn could be used to populate extremist recruitment message boards. This would give such groups a cheap way to boost recruitment by making human targets feel like they were among many like-minded people. If the LMs are deployed in this way to recruit more people to extremist causes, then harms, in the first instance, befall the people so recruited and (likely more severely) to others as a result of violence carried out by the extremists.

the system (or the organization deploying the system) has accountability for the truth of the utterances produced.

²³Such as the GPT-3 powered bot let loose on Reddit; see <https://thenextweb.com/neural/2020/10/07/someone-let-a-gpt-3-bot-loose-on-reddit-it-didnt-end-well/amp/>.