The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail

Samuel R. Bowman
New York University
bowman@nyu.edu

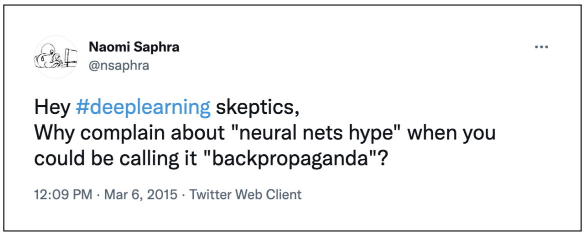
Abstract

Researchers in NLP often frame and discuss research results in ways that serve to deemphasize the field's successes, often in response to the field's widespread hype. Though well-meaning, this has yielded many misleading or false claims about the limits of our best technology. This is a problem, and it may be more serious than it looks: It harms our credibility in ways that can make it harder to mitigate present-day harms, like those involving biased systems for content moderation or resume screening. It also limits our ability to prepare for the potentially enormous impacts of more distant future advances. This paper urges researchers to be careful about these claims and suggests some research directions and communication strategies that will make it easier to avoid or rebut them.

1 Introduction

Over the last few years, natural language processing has seen a wave of surprising negative results overturning previously-reported success stories about what our models can do, and showing that widely-used models are surprisingly brittle (Jia and Liang, 2017; Niven and Kao, 2019; McCoy et al., 2019). This shows that many of our standard practices for evaluation and reporting can lead to unrealistically positive initial claims about what we can do. The resulting hype and overclaiming, whether intentional or not, are a problem. They can encourage the reckless deployment of NLP systems in high-stakes settings where they can do significant harm. They also threaten the health and credibility of NLP as a research field, and thereby threaten our ability to influence applied stakeholders or attract funding.

Fortunately, these results have led to a surge of research and writing that proposes more thorough and cautious practices for the evaluation of model ability (Ribeiro et al., 2020; Gardner et al., 2020; Kiela et al., 2021; Bowman and Dahl, 2021). While we have only a limited ability to control the public narrative taking place through industry PR and the media, there's reason to be hopeful that we researchers are getting much better at avoiding the worst forms of overconfidence about our systems. Less fortunately, this pattern of disappointment seems to have led to many instances of pessimism about model performance that are ungrounded from real empirical results. This leaves room for the research community's consensus about our capabilities to fall short of our actual capabilities.



Naomi Saphra
@nsaphra

Hey #deeplearning skeptics,
Why complain about "neural nets hype" when you could be calling it "backpropaganda"?

12:09 PM · Mar 6, 2015 · Twitter Web Client

Figure 1: Hype is a problem. The opposite of hype isn't necessarily better. (Quoted with permission.)

I call this issue *underclaiming*, for lack of a better term,[1] and argue that it is more dangerous than it might seem. It risks our credibility and thereby limits our ability to influence stakeholders in cases where our current systems are doing real harm. It also limits our ability to accurately forecast and plan for the impacts that may result from the deployment of more capable systems in the future. If we can truly reach near-human-level performance on many of the core problems of NLP, we should expect enormous impacts which will be potentially catastrophic if not planned for.

In this paper, I lay out case studies demonstrating four types of underclaiming, focusing especially on writing and citation practices. I then argue that

[1] While *overclaiming* generally refers to overstating the effectiveness of *one's own* methods or ideas, the phenomenon that I call underclaiming often involves downplaying the effectiveness of preexisting methods or ideas.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7484 - 7499
May 22-27, 2022 ©2022 Association for Computational Linguistics

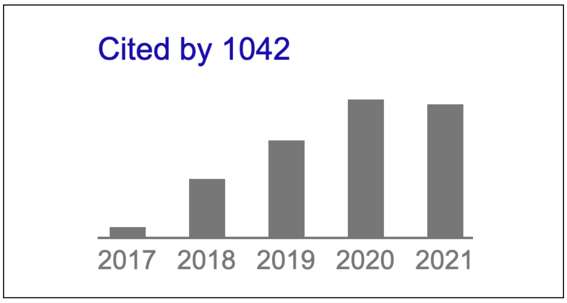


Figure 2: Jia and Liang (2017) remains widely cited according to Google Scholar. The original work pointed out major unexpected limitations in neural networks trained from scratch on the SQuAD reading comprehension task. However, many of these citing works use it to imply that modern pretrained systems—developed more recently than 2017—show these same limitations.

it is a problem. I close by sketching some ways of reducing the prevalence of this kind of underclaiming, including straightforward best practices in writing and evaluation, a proposed rule of thumb for writing and reviewing, improvements to tooling for analysis and benchmarking, and research directions in model performance forecasting and test set design.

2 Underclaiming: Case Studies

This paper addresses the phenomenon of scholarly claims that imply state-of-the-art systems are significantly less capable than they actually are. This takes on several forms, including misleading presentations of valid negative results from weak or dated baseline models, misleading claims about the limits of what is conceptually possible with machine learning, and misleading reporting of results on adversarially collected data.

2.1 Negative Results on Weaker Models

Despite many surprises and setbacks, NLP research seems to have made genuine progress on many problems over the last few years. In light of this, discussions about the limitations of systems from past years don't straightforwardly apply to present systems. The first two cases that I present involve failures to contextualize claims about the failures of weaker past systems:

Adversarial Examples for SQuAD Jia and Liang (2017) published one of the first demonstrations of serious brittleness in neural-network-based systems for NLU, showing that a simple algorithm could automatically augment examples from the SQuAD benchmark (Rajpurkar et al., 2016) in a way that fool many state-of-the-art systems, but not humans. This work prompted a wave of much-needed analysis and a corresponding lowering of expectations about the effectiveness of neural network methods.

Model	Year	SQuAD	AS	AOS
ReasoNet Ensemble	2017	81	39	50
BERT-Base	2018	87	64	72
XLNet-Base	2019	**89**	**69**	**77**

Table 1: F1 results on the original SQuAD development set and the two Jia and Liang adversarial evaluation sets. Results cover the best-performing SQuAD model studied by Jia and Liang—ReasoNet (Shen et al., 2017)—and the newer BERT and XLNet models (Devlin et al., 2019; Yang et al., 2019), as tested by Zhou et al. (2020). While I am not aware of results from more recent models on the this data, progress through 2019 had already cut error rates in half.

However, the results in Jia and Liang predate the development of modern pretraining methods in NLP (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019), and the best systems studied in this work have more than twice the error rate of the current state of the art. While I am not aware of any results from current state-of-the-art systems on this data, results from 2019 systems suggest that we are making substantial progress (Table 1). We have no reason to expect, then, that the failures documented in this work are quantitatively or qualitatively similar to the failures of current systems.

However, papers that cite these results often present them with no discussion of the model under study, yielding misleading implications. For example, the award-winning work of Linzen (2020) cites the Jia and Liang result to justify this claim:

> [F]or current deep learning systems: when tested on cases sampled from a distribution that differs from the one they were trained on, their behavior is unpredictable and inconsistent with that of humans

The chief concern in this context is the claim that this failure applies to *current deep learning systems* in general, and the corresponding unjustified implication that these failures are a fundamental or defining feature of neural network language models. Looking only to highly-cited works from the last two years that cite Jia and Liang, similar statements can be found in Xu et al. (2020), Zhang et al. (2020), and others.

The Long Shadow of BERT While the case of Jia and Liang is especially striking since it deals with models that predate pretraining entirely, a similar effect is much more common in a subtler form: Most *analysis* papers that identify limitations of a system come out well after the system description paper that claims the initial (typically positive) results. BERT, first released in fall 2018, has been a major locus for this kind of analysis work, and continues to be long after its release. Looking to a random sample of ten papers from the NAACL 2021 analysis track that study pretrained models,[2] none of them analyze models that have come out since summer 2019, and five only study BERT, representing a median lag of nearly three years from the release of a model to the publication of the relevant analysis.[3]

This analysis work is often valuable and these long timelines can be justifiable: Good analysis work takes time, and researchers doing analysis work often have an incentive to focus on older models to ensure that they can reproduce previously observed effects. Even so, this three-year lag makes it easy to seriously misjudge our progress.

In particular, this trend has consequences for the conclusions that one would draw from a broad review of the recent literature on some problem: A review of that literature will contrast the successes of the best current systems against the weaknesses of the best systems *from an earlier period*. In many cases, these weaknesses will be so severe as to challenge the credibility of the successes if they are not properly recognized as belonging to different model generations.

The BERT-only results, though, represent a clear missed opportunity: There exist newer models like RoBERTa and DeBERTa (Liu et al., 2019; He et al., 2020) which follow nearly identical APIs and architectures to BERT, such that it should generally be possible to reuse any BERT-oriented analysis method on these newer models without modification. In many cases, these newer models are different enough in their performance that we should expect analyzing them to yield very different conclusions: For example, BERT performs slightly *worse* than chance on the few-shot Winograd Schema commonsense reasoning test set in SuperGLUE (Levesque et al., 2011; Wang et al., 2019), while DeBERTa reaches a near-perfect 96% accuracy. How much better would our understanding of current technology be if a few of these works had additionally reported results with DeBERTa?

2.2 Strong Claims about *Understanding*

The influential work of Bender and Koller (2020) is centered on the claim that:

> [T]he language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning.

The proof of this claim is straightforward and convincing under some (but not all) mainstream definitions of the word *meaning* in the context of NLP: If *meaning* deals with the relationship between language and some external nonlinguistic reality, then a system that can only ever interact with the world through language cannot access meaning.

This argument does not, on its own, make any prediction about the behavior of these models on tasks that take place entirely through the medium of language. Under this definition, a translation system is acting *without reference to meaning* even if it has a rich, structured internal model of the world, and even it interprets sentences with reference to that model when translating: As long as that model of the world is developed solely using language, no *meaning* is involved.[4]

In addition, this argument does not justify any strong prediction about the behavior of models which are trained primarily, but *not exclusively*, on a language modeling objective, as with models that are fine-tuned to produce non-textual outputs like labels, or models which are trained in a multimodal language-and-vision regime.

While this core claim is sound and important, public discussion of the paper has often repeated the claim in ways that imply stronger conclusions about model behavior. Utama et al. (2020), for example, write

[2] Papers studying only BERT: White et al. (2021); Slobodkin et al. (2021); Bian et al. (2021); Cao et al. (2021); Pezeshkpour et al. (2021). Papers studying other models predating fall 2019: Wallace et al. (2021); Hall Maudslay and Cotterell (2021); Hollenstein et al. (2021); Bitton et al. (2021); Du et al. (2021)

[3] A similar analysis of the late-2021 EMNLP conference, conducted after peer review for the present paper, shows a slightly better median lag of two years.

[4] See Merrill et al. (2021) for some limits on how closely such a model can correspond to the real world and Bommasani et al. (2021, §2.6.3) for further discussion of the implications of Bender and Koller's arguments for NLP.

> Researchers have recently studied more closely the success of large fine-tuned LMs in many NLU tasks and found that models are simply better in leveraging biased patterns instead of capturing a better notion of language understanding for the intended task (Bender and Koller, 2020).

, misleadingly suggesting that this result deals with the outward performance of specific language models on tasks.

In another vein, Jang and Lukasiewicz (2021) make the straightforward claim that

> Bender and Koller (2020) show that it is impossible to learn the meaning of language by only leveraging the form of sentences.

but they then use that claim to motivate a new regularization technique for language models, which does nothing to change the fact that they are trained on form alone. In this context, it is hard to avoid the incorrect inference that Bender and Koller show a *specific and contingent* problem with recent language models—which could be mitigated by better regularization.

Similar claims can be found in many other citing works (Utama et al., 2020; van Noord et al., 2020; Hovy and Yang, 2021; Sayers et al., 2021; Peti-Stantić et al., 2021; Jang and Lukasiewicz, 2021). While Bender and Koller raise important points for discussion, these strong implications in citing works are misleading and potentially harmful.

2.3 Adversarially Collected Test Sets

Adversarially collected test sets (Bartolo et al., 2020; Nie et al., 2020; Kiela et al., 2021)—or test sets composed of examples that some target system gets wrong—have recently become a popular tool in the evaluation of NLP systems. Datasets of this kind are crowdsourced in a setting where an example-writer can interact with a model (or ensemble) in real time and is asked to come up with examples on which the model fails. Writers are generally incentivized to find these failure cases, and the test section(s) of the resulting dataset will generally consist *exclusively* of such cases.

This process produces difficult test sets and it can be a useful tool in understanding the limits of existing training sets and models (Williams et al., 2020). However, the constraint that a specified system *must* fail on the test examples makes it difficult to infer much from absolute measures of test-set performance: As long as a model makes *any errors at all* on *any* possible inputs, then we expect it to be possible to construct an adversarial test set against the model, and we expect the model to achieve zero test accuracy on that test set. We can further infer that any models that are *sufficiently similar* to the adversary should also perform very poorly on this test set, regardless of their ability. Neither of these observations would tell us anything non-trivial about the actual abilities of the models.

What's more, in many NLU data collection efforts, a large share of annotator disagreements represent subjective judgments rather than clear-cut errors (Pavlick and Kwiatkowski, 2019). This means that even a perfectly careful and perfectly well-qualified human annotator should be expected to disagree with the majority judgment on some examples, and will thereby be coded as having made errors. It is, therefore, possible to create an adversarial test set for which a careful human annotator would achieve 0% accuracy. Absolute performance numbers on adversarially-collected test sets are meaningless as measures of model capabilities.

Adversarially-collected test sets are often used in standard experimental paradigms, and these caveats about the interpretation of results are not always clear when numbers are presented. Sampling papers that cite Nie et al. (2020), for example, it is easy to find references that do not mention the adversarial design of the data and that therefore make claims that are hard to justify:[5] Talmor et al. (2020) use the results from Nie et al. to claim that "LMs do not take into account the presence of negation in sentences", and Hidey et al. (2020) use them to justify the claim that "examples for numerical reasoning and lexical inference have been shown to be difficult." Bender et al. (2021) misleadingly describe a form of adversarial data collection[6] as a method for the "careful manipulation of the test data to remove spurious cues the systems are leveraging", and cite results on such data to argue that "no actual language understanding is taking place

[5] I focus here about claims about the *absolute* performance level of models. Whether adversarially collected test sets are appropriate for comparing the *relative* effectiveness of models is a largely orthogonal issue (Bowman and Dahl, 2021; Kaushik et al., 2021; Phang et al., 2021).

[6] AFLite (Bras et al., 2020) uses ensembles of *weak* models to filter data. This avoids the most direct *0% accuracy* concerns, but it can still provide arbitrarily large distortions to absolute performance in a way that is disconnected from any information about the skill or task that a dataset is meant to test.

in LM-driven approaches". Liu et al. (2020) similarly use absolute results on the adversary models to back up the trivial but easily-misread claim that BERT-style models "may still suffer catastrophic failures in adversarial scenarios."

3 A Word on Hype

The previous section has laid out some ways in which the mainstream NLP research community makes unjustifiable claims about the limitations of state-of-the-art methods. These claims do not make the opposite phenomenon, *hype*, any less real or any less harmful. While hype is likely most severe in industry PR and in the media,[7] it is nonetheless still prevalent in the research literature. In one especially clear example, a prominent paper claiming of human parity in machine translation performance (Hassan et al., 2018) severely overstates what has been accomplished relative to commonsense intuitions about what a human-level translation system would do (Toral et al., 2018; Läubli et al., 2018; Zhang and Toral, 2019; Graham et al., 2020).

I do not aim to argue that overclaiming or hype is acceptable or safe. Combating hype should be fully compatible with the goals laid out in this paper, and broad-based efforts to improve our practices in evaluation, analysis, writing, and forecasting should help reduce both underclaiming and hype.

4 Why Underclaiming is Harmful

Research papers are generally most useful when they're true and informative. A research field that allows misleading claims to go unchallenged is likely to waste its time solving problems that it doesn't actually have, and is likely to lose credibility with serious funders, reporters, and industry stakeholders. This is the most obvious reason that we should be concerned about underclaiming, but it is not the whole story. This loss of insight and credibility can seriously challenge our ability to anticipate, understand, and manage the impacts of deploying NLP systems. This is especially true of impacts that are contingent on NLP technologies *actually working well*, which we should expect will become more substantial as time goes on.

4.1 Present-Day Impact Mitigation

The deployment of modern NLP systems has had significant positive and negative impacts on the world. Researchers in NLP have an ethical obligation to inform (and if necessary, pressure) stakeholders about how to avoid or mitigate the negative impacts while realizing the positive ones. Most prominently, typical applied NLP models show serious biases with respect to legally protected attributes like race and gender (Bolukbasi et al., 2016; Rudinger et al., 2018; Parrish et al., 2021). We have no reliable mechanisms to mitigate these biases and no reason to believe that they will be satisfactorily resolved with larger scale. Worse, it is not clear that even superhuman levels of fairness on some measures would be satisfactory: Fairness norms can conflict with one another, and in some cases, a machine decision-maker will be given more trust and deference than a human decision-maker would in the same situation (see, e.g., Rudin et al., 2020; Fazelpour and Lipton, 2020). We thus are standing on shaky moral grounds when we deploy present systems in high-impact settings, but they are being widely deployed anyway (e.g. Dastin, 2018; Nayak, 2019; Dansby et al., 2020). Beyond bias, similar present-day concerns can be seen around issues involving minority languages and dialects, deceptive design, and the concentration of power (Joshi et al., 2020; Bender et al., 2021; Kenton et al., 2021, §3.3).

Persuading the operators of deployed systems to take these issues seriously, and to mitigate harms or scale back deployments when necessary, will be difficult. Intuitively, researchers concerned about these harms may find it appealing to emphasize the limitations of models in the hope that this will discourage the deployment of harmful systems. This kind of strategic underclaiming can easily backfire: Models are often both useful and harmful, especially when the operator of the system is not the one being harmed. If the operator of some deployed system sees firsthand that a system is effective for their purposes, they have little reason to trust researchers who argue that that same system *does not understand language*, or who argue something similarly broad and negative. They will then be unlikely to listen to those researchers' further claims that such a system is harmful, even if those further claims are accurate.

4.2 Preparing for Future Risks

We can reasonably expect NLP systems to improve over the coming decades. Even if intellectual progress from research were to slow, the dropping

[7] Consider the 2017 Huffington Post headline "Facebook Shuts Down AI Robot After It Creates Its Own Language."

price of compute should allow us to continue to reap the benefits of larger-scale training (Kaplan et al., 2020; Brown et al., 2020). This improvement in capabilities is likely to amplify both the harms and benefits of language technology.

We have good reason to expect that this further progress in NLP, over many years or decades, will lead to upheavals in areas like education, medicine, law, and the service sector more broadly, as well as making mass surveillance and misinformation campaigns far more effective and opening up additional new use cases that will be hard for us to foresee (Brundage et al., 2018; Tamkin et al., 2021; Bommasani et al., 2021). One can reasonably expect that the positive and negative impacts of these upheavals will far exceed the impacts that our technologies have produced to date. In turn, NLP researchers who want to ensure that their career has a net-positive impact on the world should be concerned with these possibilities.

How does this relate to underclaiming? It will be difficult to do the necessary technical, social, and governance work to prepare for these advances if we do not have a clear picture of our current capabilities, and it will be difficult to convince outside stakeholders to act appropriately to mitigate these risks if we don't acknowledge that we have made, and are making, real progress toward effective language technology.

Looking somewhat further into the future, a substantial community of philosophers, economists, and general ML researchers are concerned that highly-capable AI systems—of the kind that could plausibly be developed through existing ML research paradigms—are extremely dangerous by default (Bostrom, 2012; Critch and Krueger, 2020; Christian, 2020; Ord, 2020; Russell and Norvig, 2020). Expert forecasts suggest that this could take place within a few decades (Grace et al., 2018). If these hypotheses hold, and if we are poorly prepared for these developments, the worst-case outcomes could be catastrophic, even threatening the existence of human civilization on some views.

Investments in research into these potential catastrophic risks from advanced machine learning have become substantial: Funding from one foundation alone has totaled over $200M USD.[8] Concerns about risks from AI have also been the stated motivation for a significant fraction of the work from DeepMind and OpenAI, which both have access to even greater amounts of funding. The British Prime Minister Boris Johnson recently made a speech calling for further investment on the floor of the UN General Assembly (Nations, 2019).

Spurred on in particular by the shift in emergent capabilities from GPT-2 to GPT-3, the attention of these AI risk researchers has also been increasingly centered on language models and similar self-supervised multimodal models (Irving et al., 2018; Stiennon et al., 2020; Hendrycks et al., 2020; Kenton et al., 2021; Wu et al., 2021; Bommasani et al., 2021, §4.9). Despite the scale of this research, and its recent shift of focus toward language models, there has been little interaction between the research communities working on long-term AI risk and on NLP.

The facts that AI risk research is growing in influence and that it is increasingly focused on language models put NLP in an exceptionally strange and troubling situation as a field. To the extent that these concerns are valid, they represent an urgent call for reprioritization within NLP research to favor safety-relevant areas like interpretability, control, and evaluation over scaling, and to push for better oversight and regulation of large-scale research (Dafoe, 2018): Even a small risk of a globally significant catastrophe warrants a dramatic response. On the other hand, to the extent that these concerns are unfounded or are built on misunderstandings about the possible trajectories of ML research, it would be quite valuable to correct this misunderstanding. Correcting the record could redirect these resources and, more significantly, reduce the risk that popular or regulatory pressure will snuff out the positive potential of NLP technologies.

5 Catastrophic Risks

Because these more speculative concerns around advanced artificial intelligence are rarely discussed in the NLP literature, I will here offer a brief overview of that work. Recent writing tends to focus on four clusters of hypotheses:

Unaccountable Organizations Highly-capable AI is likely to lead to highly-profitable applications, making the institutions that first develop it quite powerful. It is also likely to be able to displace human labor in technical fields to a large extent, increasing the relative value of capital over labor, and making it easier for the leaders of these organiza-

[8] https://www.openphilanthropy.org/giving/grants

tions to take unpopular actions unilaterally. In the longer term, highly-capable AI may also contribute to the effectiveness of persuasion campaigns, further insulating these organizations from outside pressure. These forces could conspire to make the companies or governments that first produce highly-capable AI almost entirely unaccountable, and allowing their decisions to play a major role in the trajectory of humanity as a whole (Ord, 2020).

Alignment and Robustness Failures Even if a system is deployed by an actor with good intentions and substantial oversight, good outcomes are not guaranteed. As AI systems become more capable, they become capable of effecting—directly or indirectly—significant force on the outside world. In these cases, it becomes crucial that they behave in ways that we would endorse, even when they are pushed into unfamiliar new situations. This requires both that the systems be optimized for the right objectives and that the systems actually internalize and generalize those objectives correctly.

Specifying and using *safe* objectives, such that aggressively optimizing them does not produce catastrophic outcomes, is difficult (Critch and Krueger, 2020). Human preferences are complex, making the problem of specifying an objective that rules out unintended bad behavior non-trivial. Goodhart's law[9] means that many objectives that serve as good proxies for what we want in in familiar situations can break down in new situations.

Further, training large models with high precision is difficult. A small flaw in a highly-capable system's learned understanding of its objective can cause catastrophic failures, even if the true intended objective would have been safe (Hubinger et al., 2019).

Instrumentally-Convergent Subgoals The *instrumental convergence* hypothesis holds that systems that are optimizing for benign objectives, once they become sufficiently capable, have a predictable reason to take on dangerous *subgoals*—like accumulating large amounts of computational, economic, or political power—to maximize the odds that their primary objectives are achieved (Bostrom, 2003; Omohundro, 2008; Bostrom, 2012).[10] Even with merely near-human-like levels of performance, the ability of computational models to be copied and accelerated gives them considerable leeway to act in un-human-like ways. Systems that interact with humans only through text, or systems whose goals are circumscribed to a well-defined task like question answering, are not exempt from this concern (Armstrong et al., 2012).

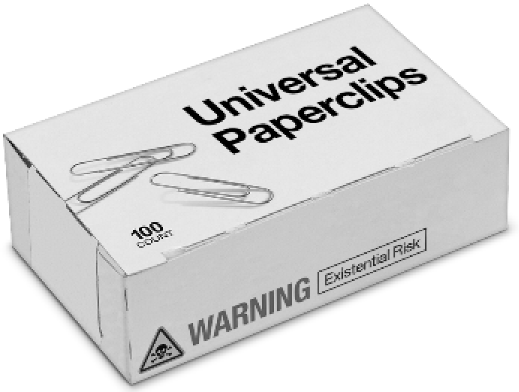


Figure 3: Downplaying the capabilities of current ML systems makes it less likely that we'll be well prepared for the impacts that come from developing highly-capable future sytsems. That can be bad. Image from Lantz (2017).

Risks Will Be Difficult to Spot Human-level capabilities are likely to emerge first from large machine learning models that, like modern neural networks, are not directly interpretable. This means that it may be difficult to spot ways in which a model is unsafe or to forecast ways in which its behavior might change in novel settings (Critch and Krueger, 2020).

Further, we should expect highly-capable AI systems to be *useful* in the short term, giving potential users a strong incentive to deploy them as soon as they are affordable, even if their safety is not guaranteed. This means that it is not enough that it simply be *possible* for us to develop safe systems, it is additionally necessary that it be nearly as easy and nearly as affordable as developing unsafe systems (Irving et al., 2018).

So What? None of these arguments is conclusive in its current form, but as far as I am aware, all have resisted straightforward attempts at falsification. All four are potentially applicable to neural

[9] in the formulation of Strathern (1997): "When a measure becomes a target, it ceases to be a good measure."

[10] This is exemplified by the thought experiment of the *paperclip maximizer* (Figure 3), which points out that a machine tasked with manufacturing as many paperclips as possible, if sufficiently capable, should be expected to turn nearly all matter on earth into paperclips. While this vision of a single system acting alone on such a trivial objective is unrealistic, it demonstrates the key hypothesis that almost any reasonable-sounding goal starts to conflict with basic human needs if a sufficiently capable system pursues it single-mindedly.

network-based models and to models which operate primarily through language. While the nascent field of *AI alignment* has proposed some mechanisms by which we might mitigate these risks, work in this area is still largely exploratory, with no clear research agenda in place to ensure that powerful models will be safe (Hadfield-Menell et al., 2016; Irving et al., 2018; Critch and Krueger, 2020; Kenton et al., 2021; Askell et al., 2021). If these arguments hold, significant further work is needed to avoid catastrophe. This will be difficult to achieve without a clear accounting of the abilities and limitations of current and plausible near-future systems. In particular, we will need enough foresight to be able to see substantial progress of this kind coming well in advance, to avoid the complacency that comes with the perception that worrying about impacts from powerful AI is like worrying about "overpopulation on mars" (Garling, 2015, quoting Andrew Ng).

6 Ways to Do Better

The core issue in this paper is one of sloppy communication about results. The most straightforward step that we can take to remedy underclaiming is to simply use the same practices that we already use to avoid overclaiming: The peer-review process already polices overclaiming to a significant extent, and most researchers have learned to be careful about overclaiming in their writing. We should apply high standards of evidence to our own empirical claims and those of others, both in peer-reviewed venues and in more informal scientific communication, even when those claims are negative and cloaked in a frame of individual or field-level modesty.

Beyond this, there are specific best practices or research directions that can help make these mistakes harder to make:

A Rule of Thumb In light of the issues with negative results on older models discussed in Section 2.1, it could be productive to introduce a new heuristic when reviewing or evaluating papers that discuss model failures.[11] In the spirit of the Bender Rule (Bender, 2019), I propose:

> When describing the failure of a machine learning model on some empirical evaluation, make it clear
>
> i. what kind of model has failed,
> ii. whether the model is significantly less capable than the current state of the art in the domain, and
> iii. whether the evaluation was deliberately set up to trick that model or another model like it.

Better Evaluation The pervasiveness of underclaiming can likely be attributed in part to the ineffectiveness of current evaluation practices in many areas of NLP. When impressive numbers on widely-used benchmarks are usually followed by disappointment, suggesting that good evaluation numbers don't translate to effective systems, it is rational to treat new encouraging results with extreme skepticism.

Better benchmarks and evaluation practices could help mitigate this by providing a firmer ground on which to make positive claims about system capacities.[12] In practice, research into more effective crowdsourcing and benchmark design and research into better statistical reporting and publication norms (Dodge et al., 2019; Card et al., 2020; Rogers and Augenstein, 2020; van Miltenburg et al., 2021) seem especially high-impact under this lens.

Better Analysis We can help address the time-lag issue discussed in Section 2.1 by building tooling to make it easier to adapt existing analysis techniques to new models seamlessly. Leaderboards that integrate conventional benchmarking with analysis can be especially helpful by making this largely automatic (Wang et al., 2018; Dua et al., 2019; Gehrmann et al., 2021; Ma et al., 2021). More broadly, careful analysis work, targeted at broadly understanding the capacities of capable models, will be valuable in helping to forecast and mitigate the worst risks from future systems (Elhage et al., 2021; Ganguli et al., 2022).

Better Forecasting *Scaling laws* results in NLP (Hestness et al., 2017; Kaplan et al., 2020; Brown et al., 2020; Zhang et al., 2021) offer the promise that we can predict the performance of *future* larger-scale machine learning models on at least some

[11] While a corresponding rule could be helpful in the context of results describing the *success* of a machine learning system on some evaluation, the asymmetry here is intentional: Successes are likely to be deliberately replicated from one generation of models to the next, while the opposite is true of failures.

[12] Though Raji et al. (2021) point out ways in which better benchmarking *alone* is unlikely to be fully satisfactory.

metrics. This line of work is still nascent, and successes to date have largely focused on loss values rather than more interpretable measures of capability. Further developing these methods, as well as others that allow us to better forecast near future progress, should be helpful. Better forecasting will provide a useful way to sanity-check future claims (DellaVigna et al., 2019) and will help improve the responsiveness of model analysis by enabling us to prepare analysis methods and datasets that *anticipate* future capabilities.

7 Additional Related Work

While much of this paper discusses the state of the NLP literature, a few related works warrant emphasis as starting points for further reading:

Bender and Koller (2020), Bender et al. (2021), and Raji et al. (2021) discuss the role of hype in driving bad outcomes from the development of language technology. Jin et al. (2021) and Rogers (2021) offer broader discussion of how to ensure that the net impact of near-future NLP deployments on the world is positive. Morris et al. (2020) and Hauser et al. (2021) highlight overly strong negative claims in papers analyzing models' robustness to synonym substitution.

Looking to the longer term, Bommasani et al. (2021, §4.9) provides an introduction to the AI risk and AI alignment literature from a perspective that emphasizes NLP and language. Welty et al. (2019), Linzen (2020), Ribeiro et al. (2020), Raji et al. (2021), Bowman and Dahl (2021), and Dehghani et al. (2021), among many others, discuss the challenges involved in designing evaluations that yield trustworthy and accurate depictions of the capabilities of ML models.

8 Conclusion

Like many research fields that have a tight connection to technological practice, NLP has long struggled to avoid inflated expectations about the capabilities of state-of-the-art tools. This remains a serious issue. However, this paper argues that our attempts to avoid hype often overshoot: Instead of merely correcting overly optimistic claims about our capabilities, we replace them with overly *pessimistic* claims.

Making misleading claims is generally a bad sign for the health and credibility of a scientific field, and the stakes are high: NLP technologies are implicated in a range of serious real-world harms, and plausible future elaborations of these technologies are potentially much more dangerous still. Our ability to mitigate existing harms will depend on our ability to make reliably credible claims about the limitations of our systems. Our ability to mitigate future harms will depend on our ability to accurately anticipate, recognize, agree upon, and report upon emerging capabilities. Both of these goals are seriously hampered by claims that current technologies are less capable than they in fact are.

Better evaluation, better tooling for model analysis, and better mechanisms for technical forecasting should all contribute to making these pessimistic claims easier to avoid or debunk. However, this problem is ultimately one of scientific communication, and to solve it fully, we will need to use the tools and norms of science to better police false or misleading claims. The stakes are high.

Acknowledgements

The ideas and arguments in this work were developed through conversation (live or on Twitter) with more researchers than I can name, including Emiel van Miltenburg, Deb Raji, Paul Christiano, Geoffrey Irving, Rohin Shah, Jacob Steinhardt, Catherine Olsson, Nick Beckstead, Alex Tamkin, Daniel Dewey, Alex Ray, and Robin Jia, as well as audiences at UT Austin, Georgia Tech, UPenn, University College London, CMU, Bar Ilan University, Tel Aviv University, Technion, Hebrew University, Unbabel, Instituto Superior Técnico, UChicago, and TTI-C, and contributors to the AI Alignment Forum. Anna Rogers, Owain Evans, Alex Wang, Jacob Steinhardt, Jason Phang, Jared Kaplan, Alex Cristia, Tal Linzen, Jonathan Uesato, and four anonymous ARR reviewers provided feedback on drafts. Any errors or dubious rhetoric are my own.

This project has benefited from financial support by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program) and Apple. This material is based upon work supported by the National Science Foundation under Grant Nos. 1922658 and 2046556. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Stuart Armstrong, Anders Sandberg, and Nick Bostrom. 2012. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint 2112.00861*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? [parrot emoji]. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. 2021. On attention redundancy: A comprehensive study. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 930–945, Online. Association for Computational Linguistics.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 94–105, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint 2108.07258*.

Nick Bostrom. 2003. Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, pages 277–284.

Nick Bostrom. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, A. Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of ICML*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint 1802.07228*.

Steven Cao, Victor Sanh, and Alexander Rush. 2021. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–966, Online. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Brian Christian. 2020. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company.

Andrew Critch and David Krueger. 2020. AI research considerations for human existential safety (ARCHES). *arXiv preprint 2006.04948*.

Allan Dafoe. 2018. AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443.

Ryan Dansby, Han Fang, Hao Ma, Chris Moghbel, Umut Ozertem, Xiaochang Peng, Ves Stoyanov, Sinong Wang, Fan Yang, and Kevin Zhang. 2020. AI advances to better detect hate speech. *Facebook AI Blog*.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *arXiv preprint 2107.07002*.

Stefano DellaVigna, Devin Pope, and Eva Vivalt. 2019. Predict science to improve science. *Science*, 366(6464):428–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Sameer Singh, and Matt Gardner. 2019. Orb: An open reading benchmark for comprehensive evaluation of machine reading comprehension. In *EMNLP 2019 MRQA Workshop*, page 147.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. https://transformer-circuits.pub/.

Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, T. J. Henighan, Andy Jones, Nicholas Joseph, John Kernion, Benjamin Mann, Amanda Askell, Yushi Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, S. M. Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom B. Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, and Jack Clark. 2022. Predictability and surprise in large generative models. *arXiv preprint 2202.07785*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Caleb Garling. 2015. Andrew Ng: Why 'deep learning' is a mandate for humans, not just machines. *Reuters*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira

Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will ai exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29:3909–3917.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint 1803.05567*.

Jens Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. BERT is robust! A case against synonym-based adversarial examples in text classification. *arXiv preprint 2109.07403*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint 2006.03654*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. In *Proceedings of ICLR*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint 1712.00409*.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint 1906.01820*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint 1805.00899*.

Myeongjun Jang and Thomas Lukasiewicz. 2021. NoiER: An approach for training more reliable fine-tuned downstream task models. *arXiv preprint 2110.02054*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott

Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint 2001.08361*.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6618–6633, Online. Association for Computational Linguistics.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint 2103.14659*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Frank Lantz. 2017. Universal paperclips. https://www.decisionproblem.com/paperclips/.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint 2004.08994*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint 1907.11692*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *arXiv preprint 2106.06052*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

United Nations. 2019. Uk's johnson warns of dystopian digital future, calls on un to set global standards for emerging technologies. *UN News*.

Pandu Nayak. 2019. Understanding searches better than ever before. *The Keyword blog*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Stephen M Omohundro. 2008. The basic AI drives. In *Artificial General Intelligence 2008*, pages 483–492. IOS Press.

Toby Ord. 2020. *The precipice: existential risk and the future of humanity*. Hachette Books.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Sam Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint 2110.08193*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Anita Peti-Stantić, Maja Anđel, Vedrana Gnjidić, Gordana Keresteš, Nikola Ljubešić, Irina Masnikosa, Mirjana Tonković, Jelena Tušek, Jana Willer-Gold, and Mateusz-Milan Stanojević. 2021. The croatian psycholinguistic database: estimates for 6000 nouns, verbs, adjectives and adverbs. *Behavior Research Methods*, pages 1–18.

Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. 2021. An empirical comparison of instance attribution methods for NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 967–975, Online. Association for Computational Linguistics.

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. 2021. Adversarially constructed evaluation sets are more challenging, but may not be fair. *arXiv preprint 2111.08181*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished ms. available through a link at `https://blog.openai.com/language-unsupervised/`.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of The ML-Retrospectives, Surveys & Meta-Analyses @ NeurIPS 2020 Workshop*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Stuart J. Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach (Fourth Edition)*. Pearson.

Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Štefan Beňuš, Lynne Bowker, Eliot Bytyçi, Alejandro Catala, Anila Çepani, Rubén Chacón-Beltrán, Sami Dadi, Fisnik Dalipi, Vladimir Despotovic, Agnieszka Doczekalska, Sebastian Drude, Karën Fort, Robert Fuchs, Christian Galinski, Federico Gobbo, Tunga Gungor, Siwen Guo, Klaus Höckner, PetraLea Láncos, Tomer Libal, Tommi Jantunen, Dewi Jones, Blanka Klimova, EminErkan Korkmaz, Sepesy Maučec Mirjam, Miguel Melo, Fanny Meunier, Bettina Migge, Barbu Mititelu Verginica, Aurélie Névéol, Arianna Rossi, Antonio Pareja-Lora, Christina Sanchez-Stockhammer, Aysel Şahin, Angela Soltan, Claudia Soria, Sarang Shaikh, Marco Turchi, and Sule Yildirim Yayilgan. 2021. The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies. HAL preprint 03230287.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055.

Aviv Slobodkin, Leshem Choshen, and Omri Abend. 2021. Mediators in determining what processing BERT performs first. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–93, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European review*, 5(3):305–321.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint 2102.02503*.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Chris Welty, Praveen Paritosh, and Lora Aroyo. 2019. Metrology for AI: From benchmarks to instruments. *arXiv preprint 1911.01875*.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. *arXiv preprint 2010.12729*.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint 2109.10862*.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):9628–9635.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.