On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations

Roy Schwartz **Gabriel Stanovsky**
School of Computer Science, The Hebrew University of Jerusalem
{roy.schwartz1,gabriel.stanovsky}@mail.huji.ac.il

Abstract

Recent work has shown that deep learning models in NLP are highly sensitive to low-level correlations between simple features and specific output labels, leading to overfitting and lack of generalization. To mitigate this problem, a common practice is to balance datasets by adding new instances or by filtering out "easy" instances (Sakaguchi et al., 2020), culminating in a recent proposal to eliminate single-word correlations altogether (Gardner et al., 2021). In this opinion paper, we identify that despite these efforts, increasingly-powerful models keep exploiting ever-smaller spurious correlations, and as a result even balancing all single-word features is insufficient for mitigating all of these correlations. In parallel, a truly balanced dataset may be bound to "throw the baby out with the bathwater" and miss important signal encoding common sense and world knowledge. We highlight several alternatives to dataset balancing, focusing on enhancing datasets with richer contexts, allowing models to abstain and interact with users, and turning from large-scale fine-tuning to zero- or few-shot setups.

1 Introduction

Effective human communication relies on our ability to understand extra-textual context based on common sense, world knowledge or shared cultural experiences, a property often cited as Grice's second maxim of quantity: "Do not make your contribution more informative than is required" (Grice, 1975, 1989). Studies have estimated that only 12% of the information conveyed by text is mentioned explicitly (Graesser, 2013; Tandon et al., 2020). To illustrate this, consider the question "*who is the president of the U.S.*?". To answer it, a human reader is likely to presume many unstated propositions, as exemplified in Tab. 1.

In contrast to humans, supervised models often fail to generalize and understand implicit context, instead resorting to low-level correlations in the data, leading to amplified bias (Zhao et al., 2017; Stanovsky et al., 2019) and brittle performance (Schwartz et al., 2017; Gururangan et al., 2018). To address this, recent approaches have suggested mitigating such correlations by *balancing* the dataset via either adding or removing certain instances (Goyal et al., 2017; Hudson and Manning, 2019; Zellers et al., 2018; Sakaguchi et al., 2020). In parallel, developers keep building larger and larger pretrained models (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020), which, when *fine-tuned* on these datasets, consistently manage to reach human performance. Taken together, these trends lead to an arms-race between data curation

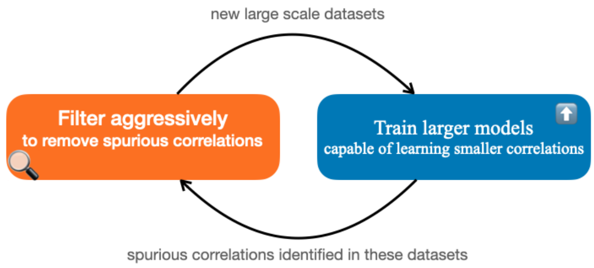

Figure 1: A high-level overview of the current state of supervised NLP research. Dataset developers create more aggressive filtering techniques (left), leading to larger models that are able to solve them by finding more elusive spurious correlations (right).

Who is the president of the U.S.?

Context	**Answer**
∅	Joe Biden
The year 2019	Donald Trump
The West Wing, season 1	Josiah "Jed" Bartlet

Table 1: Context, whether explicit or implicit, matters in textual understanding, as exemplified by the question "*who is the president of the U.S.*?". E.g., in the first line, given no other context, a QA system should provide the most sensible fallback answer (*Joe Biden*, at the time of writing).

and model development (Fig. 1).

In this position paper, we question the value of mitigating spurious correlations via dataset balancing, by showing that their existence in large training sets is both inevitable and to some extent even desired, as they are an inherent property of natural language understanding. We build on a recent result by Gardner et al. (2021), who assumed that every single-word feature correlation is *spurious*, i.e., can be used to mislead a model. We extend their argument, showing that balancing single-word features is insufficient for eliminating all spurious correlations, and that balancing feature combination is needed for that purpose. On the other hand, we show that balancing too much leads to datasets that contain no learnable signal either. We conclude by questioning whether mitigating all spurious correlations via dataset balancing is *practical*.

Following, we show that this practice is also *undesired*. We show that ignoring these correlations will hinder the learning of fallback options for both world knowledge facts (*Joe Biden is the president of the U.S.*) and common sense knowledge (*a person is happy when receiving a gift*), thus preventing models from using this knowledge in cases of uncertainty. We conclude that the existence of spurious correlations in training sets should not be solved by creating more balanced datasets.[1]

We then discuss alternatives to mitigating spurious correlations. We argue that models should be trained to understand constructions emanating from an apriori theory of language, such as negation, sarcasm, humor, and metaphors. We also suggest adopting modeling approaches that identify when the context is insufficient. We argue that in such cases, the model should *not* fallback to default assumptions, but rather abstain or interact with the user to clear ambiguities. Finally, we question the basic procedure of large-scale fine-tuning, and suggest focusing on zero- and few-shot learning instead (Liu et al., 2021b).

2 Dataset-Model Arms Race

This section provides a view of recent research in NLP as an *arms race* between models and datasets. Below we describe the conditions leading to this arms race, and present our main research question, challenging its value for making progress in NLP.

Figure 2: An example of dataset balancing (adapted from Goyal et al., 2017). For each (question, image) pair in the VQA dataset (left), VQA2.0 adds another image, for which the answer is different (right).

Models exploit spurious correlations While pretrained models consistently perform well across multiple tasks, various studies have pointed out that this is often achieved by exploiting spurious correlations in datasets, rather than improving on the underlying task (Glockner et al., 2018; Gururangan et al., 2018; Elazar et al., 2021), and that this phenomenon becomes more prominent as the models grow in size (Li et al., 2021).

Mitigating spurious correlations via balancing Various dataset curators have tried to prevent models from learning spurious correlations by modifying their training data via a careful control for the training label distribution, effectively striving for a *balanced* dataset. One approach is to *add* examples in order to balance the dataset (Goyal et al., 2017; Sharma et al., 2018; Hudson and Manning, 2019). For instance, the VQA2.0 dataset (Goyal et al., 2017) is built by taking every (question q, image i, answer a) triplet in the VQA dataset (Antol et al., 2015), and adding another triplet with the same question q, but a different image i', guaranteed to lead to a different answer a'. See Fig. 2 for an example.

Filtering as balancing A complementary balancing approach to augmentation is *filtering* examples out from datasets such that spurious correlations are minimized. This approach was taken in the creation of the SWAG dataset (Zellers et al., 2018), using "adversarial filtering" (AF). In AF, dataset instances that are easily solved by an adversarial model are filtered out. The AF approach and similar approaches were picked up by many datasets such as ReCoRD (Zhang et al., 2018), DROP (Dua et al., 2019), HellaSWAG (Zellers et al., 2019), αNLI (Bhagavatula et al., 2020), and

[1] We emphasize that balancing methods are still useful as they can lead to mitigation of *some* spurious correlations, and therefore better generalization (Le Bras et al., 2020; Swayamdipta et al., 2020), as well as potentially more efficient training. We argue that these methods are inherently limited in their ability to mitigate *all* spurious correlations.

WinoGrande (Sakaguchi et al., 2020).

Here we argue that approaches like AF converge to removing all low-level correlations,[2] and therefore a fully balanced dataset. As this approach relies on an external model, applying it with ever stronger models with higher capacity, will allow these models to pick up on subtler correlations (Li et al., 2021). At the extreme, the remaining instances that could not be solved by a fully capable model will have no statistical signal that can be exploited by that model, i.e., a balanced dataset. We henceforth refer to both augmentation and filtering as *balancing* methods.

Large models solve the new datasets In parallel to the efforts in dataset balancing, the leading *modeling* approach in recent years in NLP is pretraining large language models on raw text corpora, followed by fine-tuning them on supervised downstream applications. These models continue to grow in size (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Raffel et al., 2020), and their fine-tuning performance improves accordingly. This in turn leads to more aggressive balancing, setting in motion a kind of *arms race* between datasets and models (Fig. 1).

Evidently, a similar trend emerges for the previously mentioned datasets: (1) the first baselines, reflecting the state of the art at the time of dataset creation, perform relatively poorly, e.g., 59% on SWAG, 47% on ReCoRD, 47 F1 on DROP, 47% on HellaSWAG, 69% on αNLI, and 79% on WinoGrande; (2) model developers introduce increasingly larger and heavily-parameterized models, hill-climbing on these datasets; and eventually (3) models essentially solve the dataset within a year or two, often outperforming humans: 86% on SWAG (Devlin et al., 2019), 94% on ReCoRD (He et al., 2021b), 88 F1 on DROP (Chen et al., 2020), 93% on HellaSWAG (He et al., 2021b), 92% on αNLI (He et al., 2021a), and 90% on WinoGrande (Raffel et al., 2020). (4) new large-scale datasets are collected with more aggressive pruning techniques, thus repeating the cycle.

Based on these findings, our main research question is whether dataset balancing is the most promising method for mitigating spurious correlations. We note that an arms race between models and datasets might spur advances. Here we question a specific aspect of this arms race: the improvement of datasets by using more aggressive filtering techniques. Next we turn to present practical and conceptual limitations of this practice.

[2] Indeed, AFLite, an extension of AF, was designed to "systematically discover and filter *any* dataset artifact in crowdsourced commonsense problems" (Le Bras et al., 2020, emphasis in the original).

Name	**Description**
`ingenuine`	Correlations between features and output labels for no reason.
`ungeneralizable`	Correlations that do not generalize to new contexts.
`every-word`	Correlations between every single-word feature and output label.

Table 2: Different definitions of *spurious correlations*.

3 The Lost Battle Against Spurious Correlations

So far we have identified dataset balancing as a common way to mitigate spurious correlations. Next, we outline how different works define spurious correlations (Sec. 3.1), and then question whether dataset balancing is a viable way for mitigating them; we note that balancing too little is bound to leave spurious correlations in the data (Sec. 3.2), while balancing too much discards meaningful signal (Sec. 3.3). We finish by questioning whether this practice is even desired (Sec. 3.4).

3.1 What are Spurious Correlations?

Mitigating spurious correlations is frequently used as motivation for developing new balancing approaches. However, the term *spurious correlations* is often not clearly and consistently defined. The basic definition is a set of features that are correlated but not causally related.[3]

In NLP, several definitions of spurious correlations are typically used. One conceptual definition, denoted here `ingenuine` (e.g., Wang and Culotta, 2020; Rogers, 2021) is a feature correlated with some output label for no apparent reason. Such features often result from the annotation process (referred to as *annotation artifacts*; Gururangan et al., 2018). For instance, Gururangan et al. (2018) have shown that the words "cat" and "sleeping" are correlated with contradictions in the SNLI dataset (Bowman et al., 2015).

This definition is appealing: we want our models to learn real information about the world, and not properties of a given dataset. However, it is also

[3] https://en.wikipedia.org/wiki/Spurious_relationship

somewhat subjective, and could include features that might be referred to as genuine, such as the word "not" indicating NLI contradictions. Further, genuine features, i.e., those representing a real phenomenon in the world (e.g., "amazing" as a feature for positive sentiment), are also likely to lead models make to erroneous predictions in some contexts (e.g., negation or sarcasm; Gardner et al., 2021). Such features could thus harm generalization, so some might consider them spurious as well.[4]

In an alternative definition, denoted `ungeneralizable`, a spurious feature is one that works well for specific examples but does not hold in general (Chang et al., 2021; Yaghoobzadeh et al., 2021). This definition does not address the nature of the feature (genuine or not), but does make an implicit assumption that such features are of high importance (e.g., high pointwise mutual information values with the corresponding label; Gururangan et al., 2018). This definition is no longer subjective in terms of the genuineness of the feature, but is still subjective in the level of effect on generalizability (i.e., what is a *high* value of PMI?).

Gardner et al. (2021) relaxed the last constraint, and assumed that *every* simple correlation between single word features and output labels is spurious (henceforth `every-word`). They then defined a class of *competent* datasets, where the marginal probability for every feature is uniform over the class label, i.e., for any feature x_i and label $y \in Y$, $p(y|x_i) = \frac{1}{|Y|}$, thus limiting models from picking up any correlation between single features and output labels.

We next extend the `every-word` approach beyond single words, showing that models that can exploit single word features can also exploit some feature interactions, and therefore these should also be considered spurious. Tab. 2 summarizes the different definitions of spurious correlations.

3.2 Balancing too Little Leaves some Spurious Features

Gardner et al. (2021) assumed that as each word can appear in certain contexts that change its semantic meaning (e.g., negation, sarcasm), each word is potentially spurious. Here we note that the same argument can be applied to feature interactions, such as word n-grams. We start with a toy example to illustrate our argument for bigrams, and then extend it for larger values of n.

[4]See Eisenstein (2022) for discussion of different feature types.

Split	Text	Label
Train	very good	$+$
	very bad	$-$
	not good	$-$
	not bad	$+$
Test	not very good	$-$
	good	$+$

Table 3: A toy example of a training set (*Train*), which is balanced for unigrams, but not for bigrams. Relying on the bigram correlations (e.g., memorizing that "very good" leads to a positive sentiment) will lead to mispredictions on the test set (*Test*).

Consider the toy dataset for the task of sentiment analysis shown in Tab. 3, with vocabulary V={*good, bad, not, very*}, and label set $Y = \{ +, - \}$. The *Train* split is balanced with respect to single-word features, i.e., it is a *balanced* or *competent* dataset:

$$\forall w \in V, y \in Y \ : \ p(y|w) = \frac{1}{|Y|}$$

Assume the semantics of this dataset is that of English, while '$+$' means positive sentiment and '$-$' means negative.

A model trained on *Train* can achieve perfect training accuracy by learning the correct semantics. However, achieving perfect training accuracy can also be done by learning correlations between *two-word* features and the target label (i.e., memorizing all the training examples). In this case, the model would make the wrong prediction for the first test example in *Test* (as it has learned that *very good* is a feature that indicates positive sentiment), and similarly, will make a random prediction for the second test example, which does not contain any two-word feature seen during training.

This example highlights that balancing single-word features does not guarantee resiliency to spurious correlations, and therefore in order to mitigate all spurious correlations, balancing pairs of features is also required. One can construct similar examples for larger values of n, by similarly considering multi-word expressions and common co-occurrences (e.g., "jaw dropping", "worst day ever"). These could serve as spurious correlations in the same way single words do.

Another example is sarcasm. A model that fails to understand sarcastic contexts will misinterpret

Original Train Set		Augmented Samples	
Input	**Label**	**Input**	**Label**
0 0	0	*0 0	1
0 1	1	*0 1	0
1 0	1	*1 0	0
1 1	0	*1 1	1

Table 4: Left: a training set for the XOR function, balanced for unigrams. Right: requiring that bigrams are also balanced would prevent models from learning.

statements that appear in such contexts, even if it perfectly understands the base meaning of these statements. Thus, the entire reasoning process of such a model, whether relying on simple features, feature interactions, or other types of understanding, will result in mispredictions of certain inputs, and thus can be considered spurious.

As a result, to truly mitigate all spurious correlations in a dataset, balancing feature combinations is required as well. Accordingly, balancing too little will leave some spurious correlations in the dataset.

3.3 Too much Balancing Prevents Learning Valuable Semantic Knowledge

We observed that balancing too little does not allow models to fully eliminate spurious correlations. Here we show that too much balancing can prevent models from learning valuable knowledge.

Consider the training data for learning the XOR function presented in Tab. 4 (left). This dataset contains enough learnable signal when considering feature interactions despite being balanced for single words. Nonetheless, balancing this dataset for *pairs of features* would result in no information, and thus prevent any model from learning this function (Tab. 4, right).

Now consider a given natural language dataset D. Define n to be the length of the longest document in D. By definition, balancing every combination of up to n features leaves no learnable signal in D.[5] We conclude that balancing too much can prevent models from learning semantic knowledge.

Combining the two observations, we are left with the question of the potential intersection between balancing too much and balancing too little: does a sweet spot exist for which no spurious correlations are found in the dataset, but enough learnable signal is left? And even if so, would a balancing algorithm, whether by augmentation or filtering, be able to find it? We leave these questions for future work, but note that addressing them is a prerequisite for the theoretical and practical application of dataset balancing for mitigating spurious correlations.

[5]We assume the standard data collection process when using AF, in which the last step is balancing (Zellers et al., 2018; Dua et al., 2019), and longer instances cannot be added.

3.4 Dataset Balancing is *Undesired*

Even if a sweet spot exists between balancing too little and too much, do we really want to find it? Here we argue that perhaps not.

The practice of dataset balancing is designed to prevent models from learning that some words or expressions have a common fallback meaning that can stem from dataset artifacts (e.g., "cat" as an indicator of contradiction) but also from cultural and historical contexts (e.g., Biden is the U.S. president in 2022). Fallback meanings are crucial for understanding language, as contexts are often underspecified (Graesser, 2013). Indeed, relying on fallback meanings might make models fail to process some inputs correctly, and might not generalize to other domains where the fallback meaning is different. We argue that the ability to use them is a central ability of language understanding.

For example, substantial efforts are made to teach models *world knowledge*, such as that the president of the U.S. is Joe Biden, the capital of Brazil is Brasília, and France is the soccer world champion. These efforts include building world knowledge datasets (Wang et al., 2021), developing methods for enhancing models with this information (Zhang et al., 2019; Peters et al., 2019), and evaluating how well models capture it (Rubinstein et al., 2015; Roberts et al., 2020). But many of these world-knowledge facts are context dependent: the capital of the Brazil has changed in 1960, the president of the U.S., as well as soccer world champions potentially change every 4 years, etc.

Another example is *common sense knowledge*, such as "people are happy when they receive a gift", "an elephant is taller than a zebra", and "a statue that doesn't fit into a suitcase is too large". A large body of work has been carried out to create benchmarks that measure the common sense abilities of models (Liu and Singh, 2004; Levesque et al., 2012; Zellers et al., 2018; Sakaguchi et al., 2020; Bisk et al., 2020), as well as augmenting models with such abilities (Qin et al., 2020; Bosselut et al., 2021).

Common sense reasoning is, by definition, stochastic and reliant on understanding presupposed, underspecified context. One could imagine a person unhappy to receive a gift (e.g., because it is not what they wanted), a fantastically large zebra compared to a tiny elephant, and a suitcase with multiple compartments which prevent a small statue from fitting in it.

These examples illustrate that a model that learns these correlations and relies exclusively on them to make predictions is limited and is bound to make mistakes in some contexts. One way to avoid these mistakes is to balance these correlations out, and prevent models from knowing these assertions to begin with. We argue that this solution is not a *desired* solution. In essence, an interpreter's task (be it human or machine) is to infer the most probable context in which a statement is made, and as a result, it *should* have a fallback option for such world knowledge and common sense assertions.

Discussion We recognize that a balanced dataset may not be balanced with respect to the appearance of common-sense or world-knowledge assertions *in a given context*. E.g., a model might balance-out the general fact that Joe Biden is the U.S. president, but *not* that he is the president in 2022. As in many cases much of the context is unobserved (Graesser, 2013), the question is whether we want models to make a prediction in cases of uncertainty based on the fallback option. We argue that doing so is a desired strategy in many cases (though a preferred strategy might be to interact of abstain from making a decisive prediction, see Sec. 4.2).

We also acknowledge that correlations in the real world can be misleading. For instance, people often mistake the biggest commercial city in some countries for their capital (e.g., Istanbul in Turkey), potentially due to the high correlation between the two. In such cases, relying on the fallback option might lead to prediction error. However, we argue that following the human strategy of relying on a fallback option in cases of uncertainty will promote models' communication abilities.[6]

We want to stress that balancing methods can result in mitigating *some* of the spurious correlations, and therefore lead to increased generalization (Le Bras et al., 2020; Swayamdipta et al., 2020). Moreover, the process of filtering the data naturally results in smaller datasets, which leads to lower training costs (Swayamdipta et al., 2020). While such contribution is meaningful in terms of, e.g., environmental concerns (Strubell et al., 2019; Schwartz et al., 2020), it is orthogonal to our research question. Overall, despite the important contributions of balancing techniques, this paper shows that even the perfect balancing method might not mitigate all spurious correlations in a satisfying way.

[6] A counter example is social biases, where we want to explicitly discourage models from having a fallback option (see Sec. 4.4 for discussion).

Current Practice	**Proposal**
Dataset balancing	Richer contexts (§4.1)
A closed label set	Abstain/interact (§4.2)
Large-scale fine-tuning	Few-shot learning (§4.3)

Table 5: Our suggestions for mitigating the effects of spurious correlations, listing three current practices, each with an alternative proposal.

So how can we make models more resilient to spurious correlations without balancing the data? Below we discuss several ideas for doing this.

4 Ways to Move Forward

So far, we presented limitations of dataset balancing as a means to mitigate spurious correlations. In this section we discuss several alternatives to this practice, summarized in Tab. 5. We note that none of these proposals is particularly novel. Rather, we intend to survey alternatives proposed in literature and argue that these may be promising for addressing the drawbacks of spurious correlations, and that more efforts should be put into studying them.

4.1 Augmenting Datasets with Rich Contexts

The implicit assumption of dataset balancing is that in order to mitigate spurious correlations the model has to *unlearn* them, that is, they should be removed altogether from the training set. We argue that instead we should be focusing on learning and modeling richer contexts.

As an example, consider negation. A model that generalizes well, should learn the meaning of words such as *not*, and should be able to negate new words, even those that were seen only in positive contexts at training time. For example, if a model only sees during training words like "amazing" or "happy" with positive sentiment, and thus learns that these words bear positive meaning, we would still expect it to interpret their negated appearance (e.g., *not amazing*) as an indicator of *negative* sentiment. Such generalization is crucial for language learning, and should ideally allow models to not rely exclusively on spurious correlations. Despite the immense progress in the field in the past decade, negation still poses a challenge to modern NLP tools (Hossain et al., 2020, 2022).[7]

We suggest taking into account different types of contexts during dataset design. In particular, collecting training examples with contexts such as negation (Morante and Blanco, 2012), humor (Weller and Seppi, 2019; Annamoradnejad and Zoghi, 2020), sarcasm (Davidov et al., 2010; Oprea and Magdy, 2020), or metaphors (Tsvetkov et al., 2014; Mohammad et al., 2016). This recommendation applies to both supervised tasks, and perhaps more so to pretrained data. We suggest adding documents with such contexts throughout the pretraining corpus, or as a continued pretraining step to existing large-scale models.[8]

To incorporate contexts from a wide range of phenomena, we can leverage the vast literature on broad-coverage semantics (Baker et al., 1998; Steedman and Baldridge, 2006; Banarescu et al., 2013; Abend and Rappoport, 2013).[9] This line of work proposes theories of language, composing inventories of linguistic constructions with an algebraic formulation of their inter-relations in terms of truth value, factuality, and more. These inventories often include the phenomena discussed above, such as negation, sarcasm, and presupposition.

4.2 Interaction and Abstention to Cope with Underspecified Contexts

Most NLP tasks are designed with a closed label set that forces models to make a concrete prediction for each test instance, without an option to abstain or interact with the user to get more information. Even for tasks with a large label set (e.g., language modeling), models still have to output a valid vocabulary item. Here we argue that this practice creates an *inductive bias* towards using spurious correlations in cases of uncertainty, as the model has "nothing to lose" in case of low certainty, and is encouraged to always make some prediction, potentially relying on spurious correlations.[10]

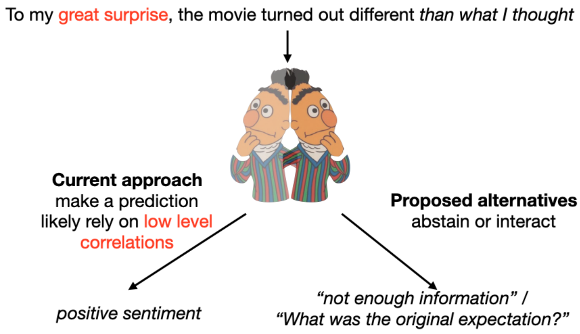

Figure 3: An example of abstention/interaction in cases of uncertainty. For the task of sentiment analysis, models currently assign a label to each input, even for ambiguous or underspecified ones (top). This may lead the model to over-rely on spurious correlations (marked in red, bottom left). Models that abstain or interact (bottom right) might learn to rely less on such correlations.

To further illustrate this point, consider the ambiguous sentence "*To my great surprise, the movie turned out different than what I thought.*", in the context of sentiment analysis. The reader cannot infer whether the writer is pleasantly surprised (a positive review) or disappointed (a negative review). We argue that in such cases models might lean towards a positive sentiment based on the words "great" and "surprise", which are typically correlated with a positive sentiment.

To test this, we ran a RoBERTa-large model (Liu et al., 2019) fine-tuned on SST-2 (Socher et al., 2013) on that example.[11] As expected, the model returns a *positive* label, with 99.99% confidence. Interestingly, three different interpretation methods (simple gradient visualization, Simonyan et al., 2014; integrated gradient visualization, Sundararajan et al., 2017; and SmoothGrad, Smilkov et al., 2017) all find the word "great" or "surprise" to be one of the three most influential words on the model's prediction. While this example does not prove the prevalence of this problem, it does demonstrate its existence.

To address this problem, we suggest adopting approaches that allow models to abstain and interact when they cannot make a decision with high confidence (Chow, 1957; Hellman, 1970; Laidlaw and Feizi, 2019; Balcan et al., 2020). See Fig. 3. This can be achieved by building datasets with unanswerable questions (Ray et al., 2016; Rajpurkar et al., 2018; Sulem et al., 2021), but also by designing models that abstain in cases of low certainty for

[7] Though we should continually assess the challenge negation poses on the most recent models (Bowman, 2022).

[8] We recognize that editing pretrained corpora poses significant challenges due to their immense size, as demonstrated by recent efforts such as corpus analysis (Dodge et al., 2021) and deduplication (Lee et al., 2022).

[9] See Abend and Rappoport (2017) for a survey.

[10] We recognize that in some cases we do want the model to make a prediction under cases of uncertainty (see Sec. 3.4). The ability to detect when is it reasonable to make an educated guess is an important property of an intelligent agent, and an exciting research question.

[11] We used the AllenNLP demo (https://demo.allennlp.org/sentiment-analysis/).

all inputs, even those with an unambiguous gold label.[12] We hypothesize that encouraging the model to provide this output when it is unsure, rather than making a semi-educated guess, potentially based on spurious correlations, could reduce its dependency on such correlations.

4.3 The End of Large-Scale Fine-Tuning?

This paper has demonstrated the limitations of mitigating spurious correlations via dataset balancing. A naive way to mitigate spurious correlations is to stop using large-scale datasets altogether. We echo recent calls (Liu et al., 2021b) and argue that for supervised learning (i.e., large-scale fine-tuning), recent advances in zero- and few-shot learning might make this option possible.

Large pretrained models such as T5 (Raffel et al., 2020) or GPT-3 (Brown et al., 2020), trained on vast amounts of data, arguably learn enough about the world to acquire many of the skills currently learned through supervised learning. Indeed, the large increase in the size and capacity of pretrained language models has led to a new wave of few-shot and zero-shot methods (Schick and Schütze, 2021; Shin et al., 2020; Gu et al., 2022), which are able to reach human-level performance on certain tasks using only a few dozens of training examples (Schick and Schütze, 2021). Given these impressive results, it is not clear whether there is still value in fine-tuning models on large-scale datasets for all tasks. In the context of this work, focusing on few-shot learning might allow models to not learn some of the correlations that result from manual annotation (Schwartz et al., 2017; Gururangan et al., 2018; Poliak et al., 2018), as they will not be exposed to many of them to begin with.

We note that this proposal is not a perfect solution. First, some spurious correlations may be picked up by the small number of examples. This is less of a problem in the zero-shot setting, or in cases where the model parameters are not updated in few-shot settings (Brown et al., 2020), but studying the extent to which spurious correlations are picked up in other few-shot settings is an important avenue for future research. Second, some spurious correlations might be picked up during the pretraining stage (Gehman et al., 2020; Birhane et al., 2021; Dodge et al., 2021). Continuing to quantify this phenomenon and finding ways to mitigate it is another important line of research.

An important question in this context is the tasks for which supervised learning is still needed. It seems plausible that excelling in language modeling tasks requires mastering the skills that stand in the base of many NLP tasks, such as sentiment analysis, syntactic parsing, and NER. However, it is similarly plausible that this is not the case for other tasks, e.g., summarization, simplification and dialogue. We are cautious in making concrete recommendations for which tasks to apply this principle, but suggest the following intuitive rule of thumb: for datasets or tasks for which the state of the art is close to or surpasses the human baseline, we should consider moving to few-shot setups.

Finally, dataset creation is still a valuable and important line of research. Our recommendation to stop building large scale training sets does not make this task redundant, to both spur the design of better models, and to better test their capabilities. We suggest that instead of building large training sets and small validation and test sets, authors should consider building large test sets, as a means for achieving improved statistical power (Card et al., 2020).

4.4 A Note on Social-Bias Correlations

So far, we discussed the problems with unlearning spurious correlations, and advocated instead for more elaborate context modeling. One exception might be the case of social biases. Textual data often reflects human stereotypes such as spurious correlations between labels and protected group attributes, e.g., alignments between professions and gender or race. Unlike other types of knowledge discussed in Sec. 3.4, in this case there is an incentive to prevent models from learning this type of correlation as means for actively reducing the harms of such biases, especially in commercial and public-facing applications, such as machine translation (Stanovsky et al., 2019) or automated financial decision-making (Bartlett et al., 2021). As a result, methods for dataset balancing are no longer *undesired* for mitigating such spurious correlations.

Nonetheless, as demonstrated in Sec. 3, methods for dataset balancing are a limited solution for mitigating spurious correlations, including social-bias ones. In contrast, the methods proposed in this section for mitigating spurious correlations might also

[12] Model calibration techniques (DeGroot and Fienberg, 1983; Guo et al., 2017; Card and Smith, 2018) are often used both for allowing models to abstain (Cortes et al., 2016; Shrikumar et al., 2019) and identifying unanswerable questions (Kamath et al., 2020; Zhang et al., 2021).

assist in mitigating social biases, or at least slow down their amplification (Zhao et al., 2017).

5 Related Work

This paper discusses the arms-race between models and datasets. Previous works criticized one side of this arms race—the increasing size of pretrained models—due to ethical and environmental concerns (Schwartz et al., 2020; Bender et al., 2021), or questioning its ability to learn meaningful abstractions from raw text (Bender and Koller, 2020; Merrill et al., 2021). This work studies the second part of this arms race, regarding the efforts to mitigate spurious correlations through dataset balancing. The release of such datasets is often motivated by their potential to spur progress in modeling, and to help tease apart qualitative differences between models. Liu et al. (2021a) showed that this is not necessarily the case, by observing that the ranking of reading comprehension models on small and synthetic benchmarks is similar to that of the (large) SQuAD dataset (Rajpurkar et al., 2016).

Raji et al. (2021) recently criticized the concept of benchmarks as a whole, arguing that they are only capturing specific skills and not "general" capabilities. Our paper raises related concerns about training sets implicitly containing spurious correlations, and suggests reconsidering the practice of building large-scale training sets.

Finally, concurrent to this work, Eisenstein (2022) discussed several types of spurious correlations in the context of causality theory (Pearl, 2009), and used a toy example to demonstrate their different effects on models. They concluded that domain knowledge is required to identify the correlations that are indeed spurious, i.e., those that might challenge the generalization ability of models.

6 Conclusion

Spurious correlations in large textual corpora can result in model brittleness, lack of generalization, and an inflated sense of the state of the art. Mitigating their negative side-effects is an important research goal of the NLP community. In this paper we presented practical and conceptual limitations of dataset balancing as a means for doing so. We proposed alternative ways for mitigating spurious correlations, including adding richer contexts to textual corpora, and allowing models to abstain or interact in cases of uncertainty. We concluded by suggesting to reconsider the practice of fine-tuning pretrained models on large-scale training sets.

7 Broader Impact and Ethical Consideration

Our work did not involve any new data or annotation collection, and as such did not require crowdsourced or in-house workers, or introduces any new models and related risks. Instead, we examine existing resources and common data balancing approaches. In Section 4.4 we specifically discuss the relation between these practices and implications on social bias in models.

Acknowledgements

We would like to thank Matt Gardner and Will Merrill for the in-depth discussion. We would also like to thank Omri Abend, Yoav Goldberg, Inbal Magar, and the anonymous reviewers for their feedback. This work was supported in part by research gifts from the Allen Institute for AI.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proc. of ACL*.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proc. of ACL*.

Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using bert sentence embedding for humor detection. arXiv:2004.12765.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *Proc. of ICCV*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. of ACL*.

Maria-Florina Balcan, Avrim Blum, Dravyansh Sharma, and Hongyang Zhang. 2020. On the power of abstention and data-driven decision making for adversarial robustness. arXiv:2010.06154.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of LAW VII & ID*.

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2021. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*.

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 *Proc. of FAccT*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of ACL*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proc. of ICLR*.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *Proc. of AAAI*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proc. of AAAI*.

Samuel R. Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how nlp systems fail. In *Proc. of ACL*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 *Proc. of NeurIPS*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proc. of EMNLP*.

Dallas Card and Noah A. Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proc. of NAACL*.

Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proc. of EMNLP: Tutorial Abstracts*.

Kunlong Chen, Weidi Xu, Xingyi Cheng, Zou Xiaochuan, Yuyu Zhang, Le Song, Taifeng Wang, Yuan Qi, and Wei Chu. 2020. Question directed graph attention network for numerical reasoning over text. In *Proc. of EMNLP*.

Chi-Keung Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, 6:247–254.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Boosting with abstention. In *Proc. of NeurIPS*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proc. of CoNLL*.

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proc. of EMNLP*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL-HLT*.

Jacob Eisenstein. 2022. Uninformative input features and counterfactual invariance: Two perspectives on spurious correlations in natural language. In *Proc. of NAACL*.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proc. of EMNLP*.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proc. of EMNLP*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proc. of ACL*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proc. of CVPR*.

Arthur C Graesser. 2013. *Prose comprehension beyond the word*.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*.

Paul Grice. 1989. *Studies in the Way of Words*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proc. of ACL*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proc. of ICML*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL-HLT*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proc. of ICLR*.

Martin E. Hellman. 1970. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6:179–185.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of EMNLP*.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proc. of ACL*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proc. of CVPR*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proc. of ACL*.

Cassidy Laidlaw and Soheil Feizi. 2019. Playing it safe: Adversarial robustness with an abstain option. arXiv:1911.11253.

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

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proc. of ACL*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proc. of KR*.

Xiang Lorraine Li, Adhiguna Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2021. Do language models learn commonsense knowledge? arXiv:2111.00607.

Hugo Liu and Push Singh. 2004. Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4).

Nelson F. Liu, Tony Lee, Robin Jia, and Percy Liang. 2021a. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. arXiv:2102.01065.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.13586.

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:1907.11692.

Will Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form:what will future language models understand? *TACL*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proc. of *SEM*.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proc. of *SEM*.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proc. of ACL*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proc. of EMNLP*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proc. of *SEM*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proc. of EMNLP*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *Proc. Of NeurIPS Benchmarks and Datasets track*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. of ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in VQA: Identifying non-visual and false-premise questions. In *Proc. of EMNLP*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proc. of EMNLP*.

Anna Rogers. 2021. Changing the world by changing the data. In *Proc. of ACL*.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proc. of ACL*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Proc. of AAAI*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proc. of NAACL*.

Timo Schick and Hinrich Schütze. 2021. True few-shot learning with prompts – a real-world perspective. arXiv:2111.13440.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *CACM*, 63(12).

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proc. of CoNLL*.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proc. of ACL*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proc. of EMNLP*.

Avanti Shrikumar, Amr Alexandari, and Anshul Kundaje. 2019. A flexible and adaptive framework for abstention under class imbalance. arXiv:1802.07024.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proc. of ACL*.

M. Steedman and J. Baldridge. 2006. Combinatory categorial grammar. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 610–621. Elsevier, Oxford.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proc. of ACL*.

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do we know what we don't know? studying unanswerable questions beyond SQuAD 2.0. In *Findings of EMNLP*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proc. of ICML*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proc. of EMNLP*.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proc. of EMNLP*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proc. of ACL*.

Luyu Wang, Yujia Li, Ozlem Aslan, and Oriol Vinyals. 2021. WikiGraphs: A Wikipedia text - knowledge graph paired dataset. In *Proc. of TextGraphs*.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of EMNLP*.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proc. of EMNLP*.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proc. of EACL*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proc. of EMNLP*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proc. of ACL*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. arXiv:1810.12885.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of ACL*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proc. of ACL*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*.