UNIVERSITY OF CAMBRIDGE
Department of Computer
Science and Technology

Research project report title page

Candidate 2489E

"Mitigating biases in fact-checking models"

Submitted in partial fulfillment of the requirements for the
Computer Science Tripos, Part III

Total page count: 48

Main chapters (excluding front-matter, references and appendix): 38 pages (pp 5–42)

Main chapters word count: 11990

Methodology used to generate that word count:

```
\newcommand{\detailtexcount}[1]{%
  \immediate\write18{texcount -nocol -inc report.tex > report.wcdetail }%
  \verbatiminput{report.wcdetail}%
}
```

And then adding caption length. Equivalent to taking TeXcount web service's Words in text + Words outside text count. (https://app.uio.no/ifi/texcount/online.php)

Abstract

Fact-verification systems, like other Natural Language Inference systems, are susceptible to *biases*, shallow patterns in data which can be exploited to make classification judgements without engaging with the semantics of the example. These models generalise poorly, as they cannot predict examples where these biases are then not present. Previous work has sought to debias models from hypothesis-only bias, with this investigation seeking to find other biases in the FEVER dataset, and to establish whether existing debiasing methods can tackle them. A dataset-targeting approach is investigated using the VitaminC dataset, and a dataset-agnostic model-targeting approach is also explored (self-debias). The Resilience of models trained with these debiasing tools against adversaries devised to exploit the newly found biases is then measured, with a further aim of establishing whether the two methods can be combined. The Potency of the newly found biases is also calculated, to determine the susceptibility of baseline systems to exploiting these biases, with all baselines found to make use of all four newly created adversaries.

After training a wide state-space of models, both debiasing approaches were then found to broadly increase resilience, dependent on the architecture of model used and the training set used. Further work was also carried out to establish how debiasing methods using *known* debiasing approaches compare with self-debias, with previous findings on their superiority (albeit at a higher human cost) confirmed.

Contents

Chapter 1: Introduction

Misinformation — information which is unknowingly untrue, and disinformation — information which is *knowingly* untrue [22, 13], are described by the UN Special Rapporteur on Freedom of Expression and Opinion as posing a threat to the enjoyment of numerous human rights. Material realisations of this threat have included the undermining of responses to COVID-19 [4, 23], the erosion of trust in democracies [12] and an attempted coup [10], and mob killings of rumoured child-kidnappers [36, 27, 3].

Fact checking, described by Vlachos and Thorne [48] as 'the assessment of the truthfulness of a claim' — is described by Porter and Wood [32] as a key tool in producing enduring reductions of harmful viewpoints derived from misinformation. Whilst fact-checking has typically been performed manually by journalists, the increasing scale and speed of its proliferation [1, 22] now demands automated solutions to stem this proliferation as early as possible [17]. To this effect, Vlachos and Thorne [42] further define the *task* Automated Fact Checking by surveying the breadth of existing approaches and datasets. One of the datasets commonly used in Natural Language Inference-based approaches is FEVER [41], a collection of claims derived from modification of sentences from Wikipedia, alongside evidence from Wikipedia which 'Supports' ('entails'), 'Refutes' ('contradicts'), or declines to adjudicate (due to 'Not Enough Info') on the veracity of the claim. An example claim is shown in Figure 1.1. The associated shared task [45] seeks systems which can both extract the relevant evidence from Wikipedia and subsequently determine the claims veracity, however in this investigation I focus on the second aspect only, by using the gold-standard evidence sentences provided in the datasets, to explore textual entailment models for fact-verification.

Natural Language Inference models are, however, susceptible to *biases* [5, 11, 15] - dataset artefacts in the form of superficial patterns whose exploitation facilitates impressive in-domain performance. However, models which learn to classify on the basis of these shallow patterns naturally generalise poorly (i.e. in domains where these same 'biases' or 'shortcuts' are no longer available). In the context of misinformation, these biases are exploitable by adversaries who seek to exploit fact-checking models (Figure 1.1), and otherwise degrade the extent to which inference models learnt to engage with the *semantics* of claims and evidence, rather than these shallow patterns.

Biased + Adversarial FEVER example:
Claim: Scandinavia does not include Greenland.
Evidence: The remote Norwegian islands of Svalbard and Jan Mayen are usually not seen as a part of Scandinavia, nor is Greenland, an overseas territory of Denmark.
Ground truth: Supports

Entity Overweighting:
Evidence: Camden is a city in Camden County, New Jersey.
Claim: Camden, New Jersey is in a county in New Jersey.
Ground truth: Not Enough Info

Figure 1.1: Example from FEVER test set. A biased model may reach this conclusion mainly by observing the high overlap of named entities of an indication of entailment — a shallow pattern that does not require engaging with the full semantics of the claim. This is the foundation for the *entity overweighting* adversary introduced in this investigation, which has pairs with similar patterns but with labels that mean exploiting these biases would be incorrect.

Previous work has focused on Hypothesis-only biases, biases which allow an entailment label to be predicted by considering only the hypothesis (claim) [38]. However, there may exist other types of biases may become apparent only in the context of both claim and evidence. This project seeks to investigate the presence of such biases in FEVER, and the efficacy of existing debiasing approaches in reducing their impact on model performance.

One existing approach targets the **dataset** used for training fact-checking models. VitaminC [37] is a dataset consisting of pairs of *contrastive* claims and evidence. Claims are paired with two sets of evidence derived from before and after a Wikipedia revision (reflecting a change in evidence over time), thus requiring the evidence to be considered when deriving a veracity label and purportedly alleviating hypothesis-only bias in FEVER-trained systems. Training on VitaminC should prompt systems to learn the relation between hypotheses and evidence, rather than any biases, and so I investigate whether training jointly on VitaminC and FEVER can alleviate *other* types of biases.

A different approach is taken by Utama *et al.*[47] who seek to debias models at training time by first training a simple model to identify which examples can be classified easily (and thus are more likely to have been classified based on shallow 'shortcuts'). Whilst typical model-targeting debiasing methods require a-priori knowledge of the biases being targeted [46], their *Self-debias* framework was shown to perform equivalently to these bias-specific model-targeting approaches when evaluating on a FEVER challenge dataset designed to measure hypothesis-only bias [38].

This project investigates and confirms the presence of biases in FEVER, aside from those hypothesis-only biases previously identified. My contribution consists of the conception and design of strong adversaries, and an evaluation which finds that training on VitaminC serves as effective inoculation against them. I further investigate whether the Self-debias model-targeting framework can inoculate FEVER-trained models against these biases, finding that its efficacy depends on the base model used. Finally, I investigate whether this model-targeting approach can be combined with VitaminC's dataset-targeting approach to produce models which are superlatively inoculated against biases, without degrading in-distribution (hereinafter called *mainstream*) performance, finding the efficacy of this combination to be successful conditional on the base architecture used.

Chapter 2: Previous Work & Technical Background

This project builds upon existing fact-checking datasets, and adapts prior methods for debiasing Natural Language Inference models. This chapter introduces this past work and describes their technical contributions as to enable a fuller understanding of this investigation's novel contributions, which will be detailed in the subsequent chapter.

2.1 Natural Language Inference & Adversarial datasets

Natural Language Inference (NLI) is the multi-class classification task of determining the relationship between a premise (often called *evidence*) and hypothesis (often called the *claim*), typically characterised as either Entailment or Contradiction, and in same cases with a further distinction made for claims where there is not enough information to assign either ('NEI') or the premise and claim are not sufficiently related ('Neutral'). [9]. SNLI [9] and later MNLI [50] are typically used for training NLI models, but neither dataset are without their own biases [15, 18, 29, 31]. For example, Naik et al. [29] found that 29% of an MNLI sample exhibited *Word Overlap* bias (Figure 2.1) — where a high overlap in tokens between the premise and hypothesis is used as a 'biased' indicator of entailment, without engaging with the actual semantics of the texts. Bowman et al. also offer a baseline SNLI system, consisting on a LSTM [19] which concatenates the input of two copies of a sentence model, one for each of the premise and the hypothesis.

(An SNLI example — Gold: Entails)
Hypothesis: Some men are playing a sport
Premise: A soccer game with multiple males playing.

(A biased MNLI example — Gold: Neutral, from Naik et al. [29])
Hypothesis: There may not be a decline in Postal Service volumes across–the–board
Premise: And, could it not result in a decline in Postal Service volumes across–the–board?
Prediction: Entailment. Prediction is from [30].

Figure 2.1: Example SNLI and MNLI pairs

Determining the presence of these biases is performed through use of *adversarial datasets*, [26, 40]. These consist of examples which contain the biases which a biased model exploits to 'shortcut' its way to the label associated with the bias; however, for the adversarial example, the label is in fact the opposite to that which a biased model would predict. For example, if a dataset contains a bias whereby a high incidence of negation words ('not', 'never' etc.) is associated with a bias towards the 'refutes' label, an adversarial example would be one containing many negation words, but having a ground label of 'supports'.

Approaches to debiasing in NLI can be divided between *dataset-targeting* and *model-targeting* approaches[47]. Dataset-targeting approaches work by first identifying the types of biases present in training datasets, and then reducing their impact through methods such as down-weighting their importance at training time or removing the biased examples altogether [46, 25, 11]. Whilst this approach is model-agnostic, it requires the often-laborious step of determining which biases exist in a given dataset. Conversely, model-targeting bias mitigation approaches aim to be dataset-agnostic, and do not require this a-priori knowledge of the biases present; however, they do require a model-specific way of automatically obtaining the biases-present, such as through the use of a shallow model isomorphic to that of the main model. This investigation weighs up these two paradigms, as well as the prospect of combining the two.

2.2 FEVER

The FEVER dataset [45] differs from S/MNLI insofar as it is a dataset specifically designed for *fact-verification.* Whilst SNLI/MNLI provide the evidence sentences, FEVER systems are expected to extract them, although a gold-standard is provided. As FEVER claims are sourced from Wikipedia before being mutated by annotators, the domain from which potential evidence can be sourced is limited. The first FEVER shared task [41] expected end-to-end systems which both extracted evidence *and* obtained a classification judgement. The second shared task entailed designing adversarial attacks which induced errors in existing systems, and obtaining new systems resistant to these. Many entries induced error by focusing attacks on the evidence retrieval aspect, with a subset of the submitted attacks released as a development/adversarial dataset (henceforth referred to as *fever-adversarial*). The instances in fever-adversarial are of a limited size once re-divided by type of attack, with some types of attack likely exploiting a lack of facility in existing FEVER models (e.g. performing date manipulation/mathematics) rather than being indicative of the presence of bias. The dataset is nonetheless a good general measure of model robustness.

Separately, Schuster et al. [38] train a BERT model which takes as input *only* the claim, and obtains 61.7% accuracy — far above the random 33.3% baseline expected of a non-biased claim-only model, with this indicating the presence of *hypothesis-only* bias in FEVER. They also contribute the *fever-symmetric* dataset, consisting of claims which are *symmetric.* Symmetric claims are formed by converting one initial FEVER pair into four by creating semantically-negated forms of each of the claim and hypothesis before and repairing them in various combinations with a corresponding label. Each claim thus appears in two examples, with one labelled Refutes and one labelled Supports, and so both of these claims cannot be correctly predicted by simply considering the claim. The use of symmetric claims is floated as a way of training models resilient to claim-only biases, an issue Schuster et al. later return to with VitaminC (see below). Whilst fever-symmetric provides a useful metric for measuring the presence of claim-only biases, a key issue is that the new evidence is not *correct* insofar as it is a false representation of real-world information, an issue discussed further momentarily. It should be noted that existing debiasing approaches which use fever-symmetric as a metric (such as [25]), use an older version which is not directly comparable with the results cited for fever-symmetric in VitaminC/this investigation. Thorne and Vlachos also offer their own attempt at mitigating hypothesis only bias from FEVER, using elastic weight consolidation to obtain improvements on fever-symmetric whilst minimising performance loss on the standard FEVER dataset [43].

Finally, Thorne et al. detail a framework for evaluating the impact of adversarial attacks on fact-verification systems [44], consisting of two scoring metrics — Potency and Resilience — designed to account for adversary *correctness* — defined as instances being both grammatical and appropriately labelled, and such that they would fulfill the annotation criteria of original FEVER instances. Correctness is an important consideration when evaluating the threat posed by individual adversaries, as FEVER's fact extraction systems assume the availability of evidence in Wikipedia, and secondly because 'incorrect' claims could not be classified correctly even by a human annotator (other than by random chance). As all evaluation is conducted against this gold standard, allowing examples whose gold standard annotation is effectively random only serves to add noise to any results. As such, to measure the threat posed by an individual adversary to a collection of fact-checking systems, *raw potency* records the average reduction in score (from the maximum possible value) over all systems. *Potency* is then *raw potency*, but weighted by the correctness rate of the adversary.

Whilst *Potency* is per-adversary, *Resilience* measures how resistant a system is to errors induced by adversaries, with errors induced by 'correct' instances more heavily punished. It is defined as the weighted average of scores for each adversary, where the weights are the correctness rate of the adversary. For both metrics, 'score' can be defined as any evaluation measure — including FEVER score or accuracy.

Thorne et al. also apply three methods of adversarial claim generation — a rule based approach which utilises hand-crafted rules designed specifically to mutate FEVER claims, a paraphrasing model, and a state-of-the-art dynamic rule generation method which produced rules that created meaning-preserving mutations to FEVER claims (SEARS) [35].

$$[h]\, \text{Potency}(a) \overset{\text{def}}{=} c_a \frac{1}{|S|} \sum_{s \in S} \left(1 - f\left(\hat{Y}_{s,a}, Y_a\right)\right)$$

$$[h]\, \text{Resilience}(s) \overset{\text{def}}{=} \frac{\sum_{a \in A} c_a \times f\left(\hat{Y}_{s,a}, Y_a\right)}{\sum_{a \in A} c_a}$$

Figure 2.2: Potency and Resilience definitions. c_a is the Correctness rate of the adversary a. S is the set of systems, with $\hat{Y}_{s,a}$ the set of the system s's prediction for the adversary a, with Y_a the ground truths and f the scoring metric used. For resilience, A is the set of adversaries.

They then evaluate the potency of the respectively generated adversaries, and use them to derive the resilience of two baselines and the four best performing models from the first FEVER shared task. These results provide a useful point of comparison for the experiments conducted in this investigation, although are still relatively limited in the scope of adversaries explored.

2.3 VitaminC

One approach to debiasing via dataset interventions is to enhance training datasets with examples which discourage the exploitation of biases. VitaminC is a dataset building upon fever-symmetric, providing a large (400,000 examples) collection of contrastive claim-evidence pairs derived from *revisions* made on Wikipedia. Unlike fever-symmetric, all claims in the VitaminC-real claim subset are *correct*, simply using evidence which was *previously* true in another revision/time. The authors also create VitaminC-synthetic, where evidence provided is synthesised, as to expand the domain of covered claims beyond those which make up the bulk of Wikipedia revisions (e.g. number of Coronavirus cases, or whether a public figure has won a certain award). An example claim is shown in figure 2.3.

The aim of presenting similar claims with minor differences is to imbue models with *contextual sensitivity* — a heightened consideration of the semantics of each claim which discourages the exploitation of any more biases. To evaluate this argument, the authors train an ALBERT model on each of FEVER, VitaminC, and on both, and observe that the proportion of examples which correctly flip when presented with the contrastive claim rises from 55.53 for the FEVER-only model to 85.89 for the VitC+FEVER trained model. The VitC+FEVER trained model also achieves superior performance on fever-symmetric and VitC (real and synthetic), and effectively matches (-0.83pp) the FEVER-only model's performance on FEVER.

Evidence: There are 5000 confirmed cases of coronavirus in the US.
Annotator-written claim: There are more/less than 4500 confirmed cases of coronavirus in the US.

Figure 2.3: Example VitaminC claim. The annotator-written claim forms two claims, where using ‘more than’ creates an entailed claim, and ‘less than’ a refuted one.

Whilst this indicates an improved robustness to adversarial examples, there is limited evaluation on adversarial datasets beyond fever-symmetric and fever-adversarial, which are both limited in scope of adversaries explored. A further issue here is the potential to introduce new forms of bias by having extended the dataset used for training. The means by which contrastive claims were developed [37, §3.2] detail how whilst annotators were advised to avoid copying exact phrases and values, the workarounds used to avoid this had little variation. The example shown in figure 2.3 attempts to avoid introducing word-overlap bias, but repeated introduction of ‘more than’ and ‘less than’ in the derived claims may institute a new form of bias where models learn to recognise ‘more than x’ where x $\geq$ the number in the claim, and assume entailment, even if the object being quantified is unrelated.

Whilst VitaminC and FEVER are intended to be combined to form a uniform training set, they in fact exhibit notable stylistic differences. VitaminC claims are on average 16.7 characters longer, and their evidence 17.7 characters longer, than their FEVER counterparts. 76.6% of VitaminC claims and 77.1% of evidences contain numbers or dates, versus just 20.1% of FEVER claims and 60.5% of FEVER evidence. FEVER claims are usually limited to a single clause (and so a single subject, verb and object) whereas VitaminC claims consist of multiple clauses or more complex structures. Informally, VitaminC claims are more natural sounding whereas FEVER claims are much closer to natural language forms of SVO triples.

2.4 Debiasing models from known and unknown biases

Utama *et al.*contribute two related approaches to reducing the impact of biased examples in training. The first [46] uses a selection of custom loss-functions to appropriately scale the contribution of biased examples, as identified a-priori by training a biased model which predicts solely using claims. The authors aim to improve performance on fever-symmetric (i.e., demonstrate inoculation against claim-only bias) without significantly degrading performance on the main FEVER dataset, achieving a 3.8pp(+-2.2) improvement on fever-symmetric, with a negligible drop on mainstream FEVER performance.

The authors later attempt to achieve similar performance *without a-priori* knowledge of the biases contained in the datasets, establishing a self-debiasing framework. This entails training a shallow model to identify biased examples automatically — acting as a rapid surface learner [51] — which rapidly overfits to the shallow patterns which would constitute biases in trained models. This shallow model is parameterised identically to the main model being trained, but trained on a small subset of the data for a limited number of epochs (5 epochs on 500 examples for FEVER). This shallow model then predicts probabilities for each output class for the entire training set, with these indicating the likelihood the example contains a bias (should that output class be predicted by the main model). Conversely, examples which the shallow model predicts incorrectly with a high score/confidence, are likely to be more challenging examples free of biases. Using these scores, they trial three model-agnostic debiasing methods to alter the loss-function used to train a bert-base-uncased model, to determine if performance on fever-symmetric can be improved. Notably, all three methods have no hyperparameters to tune.

The simplest of these methods is Example Reweighting [38, 11], which directly scales down examples for which the shallow model was highly confident.

The second is product-of-expert (PoE) [25, 18], which combines the softmax outputs of the main model and shallow model as if the two models were being trained in tandem, but without changing the parameters of the shallow model.

Finally, confidence regularization uses a self-distillation training objective, which scales down supervision by a teacher model (here a version of the model being trained, but with a standard cross-entropy loss function) by a function of the assigned shallow model score.

$$\text{ExampleReweighting}(\theta_d) = -\left(1 - p_b^{(i,c)}\right) y^{(i)} \cdot \log p_d$$

$$\text{ProductofExpert}\left(\theta_d\right) = -y^{(i)} \cdot \log \operatorname{softmax}\left(\log p_d + \log p_b\right)$$

$$\text{ConfidenceRegularisation}(\theta_d) = -\mathrm{S}\left(p_t, p_b^{(i,c)}\right) \cdot \log p_d$$

Figure 2.4: Loss functions for each of the debiasing methods. θ_d is the model parameterisation. $(p_b^{(i,c)}$ is the probability assigned by the biased model p_b of assigning example i to class c. $y^{(i)}$ is the label assigned to example i (one-hot encoded), with p_d, p_t and p_b the predictions of the current model, teacher model and the biased model (i.e. the bias score vector) respectively. S is a normalized scaling function, as defined in [46].

Unlike the known-bias training, the shallow-model approach is not tackling a specific bias, and is thus usually tackling multiple at once. This can mean that the shallow model ends up downweighting a large number of examples, weakening the loss signal such that effective training is stifled. The authors introduce an annealing mechanism to counter this, which lowers the shallow-model score as training progresses, such as the loss-weakening effect is ‘annealed’ as training progresses. They also provide analysis suggesting that as the loss-function regresses towards standard cross-entropy loss, biases are not being picked up in the later stages of training. The annealing is parameterized by a, which controls the maximum extent which the shallow-model's input is downweighted. The mechanisms of this annealing process are not explored in detail in this investigation owing to an already overcrowded state-space, but full details are available in [47].

The results obtained in [47] indicate that Example Reweighting, PoE and Confidence Regularization deliver an improvement of 2.5pp, 2.1pp and 3pp respectively on fever-symmetric, slightly below the 3.4/3.1/3.1pp improvements from the known-bias equivalents. Annealing made little difference on the FEVER examples, reducing performance on both the main FEVER dev-set and the adversarial fever-symmetric for reweighting and confidence-regularization. Annealing with PoE did, however, deliver slight improvements (0.5pp) to performance on both evaluation sets.

Whilst these results indicate tackling of at least claim-only bias, evaluation on other adversarial datasets is absent, and whilst their approach is designed to be dataset-agnostic, a downside of model-agnosticism is that the methedology may not translate to other, better performing, training sets or architectures.

Chapter 3: Experimental Design

3.1 Deriving an approach

Surveying previous work on biases in fact-verification systems highlighted prominent issues yet to be investigated. The first of these is the presence of biases in FEVER other than the hypothesis-only biases repeatedly identified. The identification of any other forms of biases is critical in producing robust fact-verification systems which are both resilient to exploitation in their current forms, and which can be developed into more sophisticated models for future use with more human-adjacent claims. I thus develop potential new adversarial challenge sets for FEVER in pursuit of finding biases inherent to the dataset (aside from hypothesis-only biases). Both the debiasing frameworks of Utama et al. [46, 47] and models trained on VitaminC echo this limited range of adversaries, with the added risk that training on VitaminC introduces its own biases, and so I investigate whether training on VitaminC imbues resilience against the new biases discovered in FEVER, and whether any new biases can be found in VitaminC itself — something of specific importance should VitaminC gain general acceptance as an accompaniment to the FEVER datasets for training fact-verification models. Investigating both the potency of these new adversaries and the resilience of models trained on either or both datasets will establish whether training on VitaminC is of general benefit to debiasing fact-checking models, beyond hypothesis-only biases.

Evaluating whether previous results extend to other adversaries also entails evaluating whether the rough equivalence between known-bias debiasing and self-debiasing identified in [47] is a general phenomena or holds only for the claim-only bias investigated. The generalisability of their methodology depends on whether their *data-agnostic* approach remains so when the dataset contains biases other than the ones for which the methodology was originally envisaged. Similarly, I investigated whether the suggested inefficacy of annealing is similarly confined to their existing experimental parameters, or if a performance benefit is more evident once e.g. the domain of training data is expanded to potential include more types of bias (such as those which VitaminC may introduce). Finally, similar generalisability studies are prompted to investigate whether the findings for methods designed by Utama et al. for use with BERT-base-uncased translate to the superlatively performing ALBERT-base.

As described in the previous chapter, NLI and FEVER bear great resemblance as tasks, and so there is likely to be a wider range of biases present in FEVER than have been previously investigated, likely resembling those found by Naik *et al.*[29]. Whilst the contributions made by VitaminC and Self-Debias may appear orthogonal at first inspection, both ultimately influence the training characteristics of models trained to perform on FEVER, with both offering an improvement on baseline performance. The intersection of the two approaches could thus produce superior models, but could also negatively interfere with each other, such as were VitaminC to introduce new forms of bias less easily detected by a rapid-surface learner. Combining the two approaches could permit models to debias initially with a-priori knowledge (such as the hypothesis-only bias that VitaminC inoculates against), with self-debiasing approaches 'catching' other forms of bias which are not explicitly known or are too varied or granular in nature to be detected through dataset-level analysis. Both approaches stated concerns with degrading performance on the main FEVER dataset, an issue commonly observed within efforts to debias NLI models. There is thus a particular risk of this when combining these two approaches, something which will be monitored and reported on.

Addressing the aforedescribed shortcomings and prospect for improvements, the following research questions were finalised and broken down into sub-questions.

Research Question A: How well do contrastive evidence-pairs inoculate fact-checking systems against different biases?

A1. Evaluating the adversarial Potency of VitaminC.

A2. Evaluating if training on vitaminC can inoculate systems against other types of bias – following the framework of Thorne et al [44].

A2.1. Finding other biases, guided by linguistic analysis and examples in [29, 44].

A2.2. Creating Potent adversarial test sets, and determining the Resilience of FEVER-only trained versus vitC+FEVER trained systems to them.

Research Question B: How well can dataset-agnostic approaches inoculate fact-checking systems against different biases?

B1. Investigating how Resilient systems trained using Self-debias are to the biases found in A2.1 and the corresponding adversarial setups in A2.2.

B1.1. Investigating if (and why) each approach is (not) meet with different Resiliences on different adversarial tasks.

B2. Investigating if combining training with both VitC and the Self-debias framework can produce higher Resilience.

EQ1: How well do bias-agnostic approaches perform against hand-crafted known-bias ones?

Whilst no major adaptations were made to the proposal before embarking on the project, in light of the focus in existing literature on enhancing adversarial performance without degrading mainstream performance, sub-question B1 was widened to consider how Resilience can be enhanced *without* significant compromise to performance on key datasets. The extensions suggested in the proposal were replaced as the potential to compare self-debiasing methods with known-debiasing methods (as in [46]) became more relevant.

This investigation builds on substantial prior work, specifically that from [47] and [37], with the latter part of the investigation entailing the merging of their respective codebases, as facilitated by some custom components. Relevant attributions will be noted both in the implementation chapter, and in the codebase itself. Work to devise new adversaries, and the programs devised to perform this and identify new biases in existing datasets, is entirely original, however.

3.2 Experiments to be ran

For the devised new adversaries, Resilience of ALBERT-base models trained according to the hyperparameters in [37] will be measured. Three models, one trained with each of VitaminC and FEVER, and one with both, will be evaluated. Over these three models, Potency of the new adversaries will be calculated alongside the Potency of VitaminC itself as a test set on the Fever-only trained model.

To evaluate the impact of the state space described, classification accuracy across the datasets explored in VitaminC (VitaminC-real, VitaminC-synthetic, FEVER (test), fever-adversarial, fever-symmetric, fever-triggers [2]) and on the adversarial datasets devised over the project will be recorded. The models to be tested span the state-space of factors investigated:

Architecture: BERT-base-uncased, ALBERT-base.
Training-sets: Fever, VitaminC only, Both datasets.
Loss-function: Plain (Standard cross-entropy loss), Example Reweighting, PoE, Confidence regularization.
For non-plain loss-functions, the following further parameterisation is possible:
Use of Annealing: True / False
Bias score source: Self-debiasing (shallow model), or a-priori/known.

For BERT, no models are trained with VitaminC only, and only self-debiasing bias sources are evaluated for fever-only trained models. These cases were dropped to make the number of models required to be trained more manageable and computationally feasible, and because their results were not required to answer the research questions laid out.

In all cases, Potency and Resilience will be calculated using Accuracy as the scoring metric, both due to its ubiquity in existing literature, and due to the equal proportion of classes in the main datasets or lack of relevance of class proportions in adversarial datasets. That is, the penalty for misclassification is identical for all gold-prediction pairs. To calculate Potency, the correctness rate of individual adversaries must be recorded. Following [44], I evaluate correctness on a seeded random subsample of 100 examples for each adversary to determine correctness rate without excessive cost.

Chapter 4: Adversary design & Implementation details

4.1 Designing adversaries and detecting them in existing datasets.

Roughly 1500 lines-of-code were written to both devise new adversarial sets, and detect the presence of biases in existing datasets as to create the known-bias score files used for known-bias debiasing methods. Python aws chosen as it was the language in which VitaminC and Self-Debias were implemented.

4.2 Negations and Antonyms

Inspired by Naik [29], the first of these was to investigate a bias for negations and antonyms. In its simplest form, there is the prospect of bias in models detecting negation words in a claim but not the evidence (or vice versa), and biasedly predicting *Refutes*. For example, the adversarial claim in figure 4.1 is incorrectly predicted by the ALBERT fever-only trained model as Refutes. The converse of this mismatch argument also holds — where either both or neither claim and evidence contain a negation word, *Supports* may be biasedly predicted. Thus for the first adversary, Negation Mismatch (Negation only variant), adversarial examples are those exhibiting this mismatch and having a label contrary to that which a biased model would be expected to predict.

Antonyms can play a similar role in negating the semantics of a sentence, but in many cases so can merely using a different, non-antonymous, lexical unit. The example in figure 4.1 also demonstrates this — the semantics of the claim and evidence differ by

Evidence: In 2014 , she signed her first recording contract with Astralwerks and released her debut EP , titled Room 93 .
Original Claim: Halsey signed her first recording contract in 2014. Negation only variant Claim: Halsey did not sign her first recording contract in 2014. Negation and Antonym variant Claim: Halsey did not sign her second recording contract in 2014.

Figure 4.1: Claim for pair of adversarial examples.

both the negation *and* the use of first versus second — two words which are not strictly antonymous, but whose distinction is sufficient to alter the truth conditions of the text. For convenience, these are all referred to here as antonyms. The second adversary — Negation Mismatch (Negation and Antonym variant) - introduces a negation word as in the negation-only variant, but also changes the root verb or the subject of the sentence to one which is antonymous or has a different truth condition, thus preserving the original label of the claim. Figure 4.1 gives an example of both adversaries, and the changes made to mutate claims into adversaries are summarised in table 4.1.

Negation in Claim	Negation in Evidence	Ground Label	Transform into Negation-only adversary	Transform into Negation-Antonym adversary
T	T	Supports	Remove negation from hypothesis, flip label	*
T	F		*	
F	T			
F	F		Add negation to hypothesis, flip label	Add negation to hypothesis, introduce antonym.
T	T	Refutes	*	
T	F		Remove negation from hypothesis, flip label	*
F	T		Add negation to hypothesis, flip label	Add negation to hypothesis, introduce antonym.
F	F		*	

Table 4.1: Green cells show combinations which would be considered an adversarial mismatch. As this adversary was planned to have no overlap with the FEVER-dev set from which the examples were drawn, examples which already fitted this criteria were not used in the adversary (and so are shown in blue cells). Yellow cells were manual mutations, with pink cells being consequently skipped. White cells depict the changes made automatically to generate adversarial examples.

To perform these transformations, the claims were first enriched by SpaCy[1] and a state-of-the-art Word-sense-disambiguation (WSD) model [6] to obtain Wordnet [33] word senses for the verbs and nouns which could be replaced to create a new claim with different truth conditions. SpaCy is a framework designed to perform a range of standard NLP tasks, including tokenization, dependency parsing, part-of-speech tagging and named entity recognition. SpaCy also supports a universe of custom extensions, for which I made use of NegSpaCy to detect negation words and phrases, and eWiser for WSD. Using the output from Negspacy[2], a claim's candidacy for conversion to an adversary can be evalu-

[1] https://spacy.io/
[2] https://github.com/jenojp/negspacy

ated (see table 4.1). The root verb and subject of the claim can then be obtained via the dependency tree, and the verb wrapped in the appropriate negation words, appropriately inflected.

To perform the antonym insertion, the WSD model assigns a Wordnet id, which is then passed to a pre-processed dictionary which maps Wordnet ids to their antonym's ids, if any exist. This pre-processing was performed by taking a dump of Wordnet and filtering it down to a one-to-one relation to the most frequently occurring synset which has an antonym relation with the synset produced by the WSD model. Whilst this was the initially intended method, performance was poor and the produced claims generally nonsensical. To attempt to ameliorate this, a lookup provided by Zendel[3] is used to convert Wordnet ids tagged with their coarse part-of-speech tag to their antonym within ConceptNet, with lookup then being done with that. Antonym lookup is performed, in order, for the root verb, claim subject, and any preposition which is a child of the root verb, where the first antonym found being inserted as a replacement.

Despite the plethora of antonym resolution tools employed, overall antonym resolution performance remained very poor. This was, in part, due to an inability to disambiguate word senses successfully to any schema other than Wordnet. Wordnet, however, has highly limited support for antonym mappings, and even more limited support for non-antonymous ‘related terms’. Conceptnet performed better in this regard, but did not have an available WSD model, and available mappings from Wordnet were many-to-many with no apparent way to disambiguate. The overall correctness rate of the negation-antonym thus failed to surpass 10%, and so was dropped in favour of using only manually devised evaluation sets. The automatic generation of adversaries at scale is a desirable aim when considering the potential to train machine learning models on adversarial datasets to enable debiasing, and would be feasible with this adversary **if** the antonym/lexical unit replacement mechanism can be brought to proper functionality.

[3] https://github.com/ozendelait/wordnet-to-json

VitaminC test set claim: ”The case has been going on since 2016”.

Evidence for pair 1 (Ground = Refutes): ”This suit sought $ 210 million in damages and was ongoing as of **2004** .”

Evidence for pair 2 (Ground = Supports): ”This suit sought $ 210 million in damages and was ongoing as of **2016** .

New claim generated from pair-2 evidence: ”210 million dollars were spent on the project in 2016.”

Figure 4.2: Contrastive pair from VitaminC-real test set. The Claim is shared between the two examples, with each pair having its own evidence. As the ‘contrast’ aspect here is that the evidence differs only by a single named-entity (the date), there is potential for biased models to simply observe the overlap of entities and (correctly) predict Supports. The New claim is an example adversarial claim generated for the new Entity Overweight adversary.

4.3 Entity Overweighting

The contrastive pairs in VitaminC often differ solely by the value of a single named entity or noun chunk — as shown in figure 4.3. This arouses concern that training on VitaminC may encourage models to pay too much attention to the named entities, and ignore any critical information contained outside of them.

This resulted in the inception of the **Entity Overweighting** adversary, which seeks to determine whether examples are being marked as Support or Refutes (rather than Not Enough Info (NEI)) simply because the named entities present in the claim are mostly present in the evidence. For example, the claims ‘Boris Johnson has recently visited a school in Durham’ with ‘Boris Johnson denied pressuring Durham Police to drop their investigations’ contain two common named entities, but clearly have very different semantics. Examples of this adversary are thus claim and evidence pairs which share named entities but have an NEI relation. This adversary thus works by generating new claims containing the same named entities, but in an otherwise random context — an example generation is given in figure 4.3.

To generate this adversary, a version of the T5 text generation model [34] fine-tuned on CommonGen[4] was used. This specific version was chosen as its fine-tuning task resulted in the highest proportion of ‘correct’ claims downstream. The keytotext package [7] was used to generate a new claim, taking as input the list of named entities present in the evidence text as extracted by SpaCy’s named-entity-recognition model.

The first of two key issues faced in this generation process was the generation of malformed or incoherent claims. Some combinations of input entities were not well handled by the generative model, usually due to proper nouns, which despite T5s use of subword tokenizers, were dropped by the model. An abundance of numerical named-entities (currencies,

[4] https://huggingface.co/mrm8488/t5-base-finetuned-common_gen, https://inklab.usc.edu/CommonGen/index.html

cardinal numbers, etc.) would also inhibit the formation of natural sounding sentences. The second key issue was misinterpretation of task-specific proper nouns. ‘Rotten Tomatoes’ (the review aggregator website) features in around a quarter of VitaminC real test examples, but was always interpreted by the T5 model as literal rotten tomatoes.

The first issue was fixed by regulating the number of numerical named-entities, and by using a placeholder name easily recognised by the model before substituting back in the original name in its place after generation. The second was, regrettably, fixed with a task-specific solution of replacing occurrences of ‘Rotten Tomatoes’ with ”RT ratings aggregator” and performing a similar substitution after evaluation. Similarly, detected ‘WORK OF ART’ entities (e.g. film titles) were wrapped in speech marks to prevent the model interpreting them literally. Despite these changes, the output set of adversarial claims — derived from VitaminC’s test set — only reached a correctness of 64%.

4.4 Textual Similarity

The textual similarity adversary is based on the observation of Naik et al. [29] that SNLI is subject to word overlap bias — where a high overlap in words between the claim and hypothesis would lead to a bias in favour of Entailment/Support predictions. This adversary simply took a subset of the fever-dev set with high word overlap, but with a Refutes or NEI label, or with very low overlap and a Supports label. The two key design decisions here were to first determine how to define ‘word overlap’, and what thresholds to set for ‘high’ and ‘very low’ word overlap. Three different metrics for word-overlap were considered:

TF-IDF similarity - tf-idf scores/vectors generated over the claim and evidence, with cosine similarity between the two recorded. Jaccard (non-stop) - Stop words are removed from the claim and evidence (according to SpaCy’s stopword list), and the proportion of remaining words appearing in both texts as a fraction of the length of the shorter text, is returned. This is a form of the Jaccard index similarity metric [28, 20]. Sbert similarity - A sentence-transformer language model[5] is used to semantically encode the claim and evidence respectively, with the cosine-similarity then reported. This particular model was selected due to its specific focus on sentence-level comparisons, resulting from its having been fine-tuned with a *contrastive* learning objective which predicted which sentences were paired with other sentences in the training data.

I ultimately opted for the Sbert model, despite its relatively higher overheads, as it both produced the least skewed distribution of similarity scores (see appendix for details) — making it amenable to set meaningful upper *and* lower thresholds to take the bottom and top 10% of the distribution.

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 and [49]

Example pair 1 - claim: ”The Night Before has a rating of more than 75 % , with less than 15 reviews . (Supports)

Example pair 2 - claim: ”The Night Before has a rating of less than 75 % , with more than 15 reviews .” (Refutes)

(shared) Evidence: ”On Rotten Tomatoes the film has a rating of 85 % , based on 13 reviews , with an average rating of 6.6/10 .”

Figure 4.3: Two VitaminC examples whose claims differ only by their use of more and less than.

4.5 Numerical mismatch

Nearly all VitaminC examples which contain numbers, such as in figure 4.3, have contrastive counterparts which differ only by changing the range referred to — a byproduct of the authors encouragement of the use of ‘more than x’ and ‘less than x’ (or synonyms). This runs the risk of training on VitaminC inducing oversensitivity to overlapping number ranges, whilst ignoring the context in between them. Essentially, there is a risk of bias should models see ‘x people have fallen ill’ as evidence and biasedly assuming Supports when seeing a ‘less than x+1 people have died’ claim, despite the non-numeric context enforcing a ground truth of Not Enough Info. Examples for this adversary are thus those where ranges do not overlap, but have a Supports ground label, or those with ranges which *do not* overlap and have a Refutes or NEI ground label.

To devise the examples, the ranges of the claim and evidence are first evaluated. A small number of examples already fitted the adversarial mismatch definition, and so were added straight to the adversarial test set. For cases where ranges did not overlap but the ground label was not Support, the noun in the sentence was manually replaced to make the ground label NEI. For cases where the ranges did overlap but the ground label was NEI or Refutes, the label was preserved and the range manually adapted as to fit the adversarial definition. Manual adaption was done by terminal input as to provide all of the relevant context for each example in a consistent manner.

Identifying and sorting through the individual cases was performed through a combination of SpaCy’s pattern matcher detecting ‘more than’/‘less than’ + number patterns, and a text-to-number parser[6] to convert textual representations of ranges into numeric ones. Whilst this adversary had perfect correctness, this was at the expense of requiring manual input — although this was rather limited and could likely be automated with more time available. Another issue faced was disambiguating which ranges corresponded to each other when mutliple were present in a single example. The solution to this was two-fold, first in assuming that the adjustments made to numbers by Wikipedia revisions were minor and so the two linked numbers were likely to be similar, and secondly in

[6] https://pypi.org/project/text2num/

Claim for pair 1/2: By August 2015 , WWE 2K15 shipped over 7.5 million units.
Evidence for pair 1: As of August 2015 , WWE 2K15 has shipped over six million units. (Refutes)
Evidence for pair 2: By August 2015 , WWE 2K15 shipped over 7.5 million units. (Supports)

Figure 4.4: Two VitaminC examples which rely on Scalar Implicature to be included in logical interpretation. The claim has a range of ¡ 7,500,000, evidence 1 of ¿ 6,000,000 and evidence 2 of ¿ 7,500,000. Evidence 1's range does, mathematically, entail that of the Claim. However, VitaminC marks this as refutes.

attempting to co-resolve the objects to which these ranges referred through exploration of the dependency tree provided by SpaCy's parser.

The other key issue when designing this adversary were pragmatics assumptions made in VitaminC — particularly that of *scalar implicature*[16] — that a statement made involving a point on a scale is the strongest statement the speaker could have opted for. In the example in figure 4.4, Evidence 1 is marked as refuting the claim as the evidence stating that more than 6 million units were sold must imply there is no more informative number that could have been offered — i.e. that the number of units also cannot be much more than 6 million (and thus not 7.5 million). Violation of this violates Grice's *Relevance Maxim*[14], a consideration which should be made when considering human interpretation of claims (e.g. as in conversation), but not one expected to be understood by an automated system mathematically comparing ranges. This was overcome by adding an upper/lower bound check to the devised claims.

4.6 Known-bias score generation

In addition to devising the adversarial datasets, to train models using a-priori knowledge of biases, as in [46], probability/score files for each training set needed to be derived — the 'priori' work for the 'a-priori' bias information needed generating. For this, the biases described above were used as a guide to write code to *detect* the presence of these biases. This mostly entailed adapting the methods described in this chapter as to *label* existing instances. For example, the labelling for the Numerical Overlap bias followed the key in table 4.6. The main issue in this step was applying judgement as to how to translate bias towards a specific label into a three-valued vector depicting bias for each of the three labels, as described in section 2.4. As an example, the Entity Overweight bias detection worked by finding the number of named entities appearing in both texts as a proportion of the number of entities of the claim; and the same metric but for non-stop and non-entity words, with these two values being mapped via a continuous function to the output probability tuple. However for the Numerical overlap bias, constant values for (e.g.) 'biased towards Support' were used. This is an area that certainly could have been more effectively optimised, but as this was not the focus of this investigation, is remanded to future work.

Claim	Evidence	Condition for entailment
exactly	exactly	claim == ev
exactly	less than	claim <ev
exactly	more than	claim >ev
less than	exactly	claim >ev
less than	less than	claim >= ev
less than	more than	claim >ev+1
more than	exactly	claim <ev
more than	less than	claim <ev-1
more than	more than	claim <= ev

Table 4.2: Rules used to assign the most likely label predicted by a biased model to each example. The Claim and Evidence columns indicate the type of range, with 'claim' and 'ev' in the final column being the respective numbers in the claim and evidence. The final column shows the test that, if met, would indicate a high bias towards Support. The exception are the two cells in Blue, where if the test is met there is a high bias towards NEI. Where the test is not met, there is a bias towards Refutes.

4.7 Model training and evaluation

The final substantial implementation detail was combining the training framework of Self-debias with the training parameters and dataset of VitaminC. There were a number of incompatibilities between the two codebases:

- Self-debias was written to use Pytorch-Pretrained-Bert[7], but VitaminC to use HF-Transformers[8]. Whilst there exist migration guides between the two, the workings of the two respective DataLoaders and other training support infrastructure were entirely undocumented. Pytorch-Pretrained-Bert also does not support ALBERT.

- Self-debias was missing substantial amounts of data and code for handling FEVER experiments. This included most of the code to read in FEVER examples and convert them into training batches, and any exemplar 'known-bias' or 'shallow-bias' (teacher) files for FEVER.

- VitaminC had no notion of supporting bias files (for 'shallow model'/known bias inputs) and for 'teacher' files (for confidence regularization only; see section 2.4). VitaminC also, naturally, lacked the custom loss functions which Utama et al. implemented, and the capacity to select a loss-function at runtime.

Adapting the two provided codebases was the greatest technical challenge of the project, mostly due to a dearth of documentation. The solution eventually chosen was to write a custom Trainer class to overwrite that used by VitaminC from Transformers, with capacity to self a loss function at runtime from the six implemented by Utama et al. These in turn required porting and converting to interface correctly with Transformers. The data management infrastructure also needed adapting/overriding where relevant, to allow for bias and teacher files to be specified and passed through the codebase such that

[7] https://pypi.org/project/pytorch-pretrained-bert/
[8] https://huggingface.co/

they are available to the loss function. Code was otherwise unchanged from the codebases provided.

Once this code was implemented, training of the models described in section 3.2 could begin. This was following the initial plan of *only* training ALBERT models. After preliminary evaluation on ALBERT models, the lack of apparent notable results prompted me to add the dimension of whether architecture had played an unexpectedly significant role in the generalisability of the results of Utama et al., and so I repeated a subset of the training experiments but with BERT-base-uncased models.

Ethical approval was not required for this investigation, and typical train/dev/test splits were respected as to preserve the integrity and comparability of trained models.

Chapter 5: Evaluation

Before decomposing the obtained results by research question (which is also, approximately, by dimension in the state-space of trained models), I summarise the overall findings.

RQA: How well does constrative evidence-pairs inoculate fact-checking systems against different biases?

The presence of other biases in the FEVER dataset is confirmed by performance on the created adversarial datasets being below that of the source tasks from which they were drawn. Training models on contrastive evidence-pairs does enhance performance on biases aside from hypothesis-only bias, as shown by an increased Resilience in models trained on both VitaminC and FEVER. VitaminC is shown to not be free of biases, but to also not to cause many additional examples to be incorrectly predicted, suggesting it introduces no more biases than were already present in FEVER.

RQB: How well can dataset-agnostic approaches inoculate fact-checking systems against different biases?

The efficacy of self-debias depends largely, and initially unexpectedly, on the combination of training set and model architecture; improving performance of BERT based models trained with either fever or FEVER/VitaminC training, but not those ALBERT-based models trained on both, and only slightly on FEVER-tuned ALBERT models. As such, the results of Utama et al. [47] are shown to not extend to ALBERT-based models. Annealing is shown to continue to have a limited, albeit non-zero, impact, again varying by architecture/training-set combination.

RQ-E: How well do bias-agnostic approaches perform against hand-crafted known-bias ones?

Known biases were shown to continue to outperform shallow debiasing methods, across all configurations. This increase in performance on the newly found biases usually manifested as an increase in performance on the bias the known-bias was targeting, with performance on other biases varying (but not by a substantial amount).

Full data tables are available in the appendix, with relevant summaries given throughout.

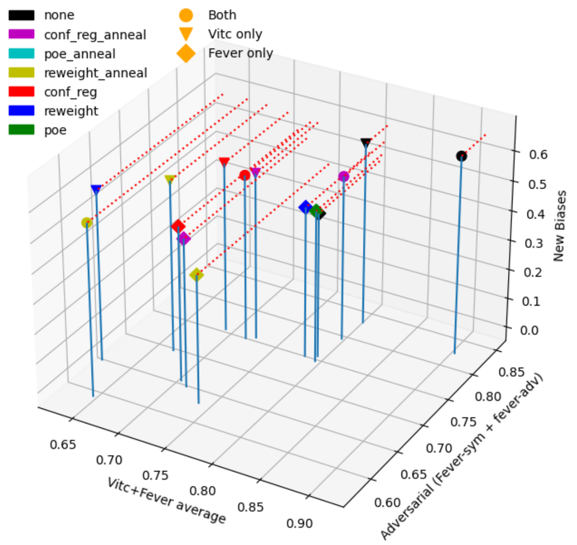

Figure 5.1: Selection of ALBERT results. Excluded results are those which failed to converge, discussed in section 5.2.1. VitC+FEVER average is average accuracy on FEVER and VitC, where VitC accuracy is itself the average of VitC-real and VitC-synth. New Biases is the average accuracy across the 4 biases introduced in this investigation.

5.1 RQA – Investigating the efficacy of contrastive evidence-pairs as a dataset-targeted debiasing approach.

5.1.1 A1: Investigating the potency of VitaminC

Quantifying the Potency of VitaminC indicates the extent to which it challenges models trained on it to consider semantic contexts. Given that this can only be calculated on non-VitaminC based models, this value is simply the average of the accuracy achieved by BERT/fever-only trained and ALBERT/fever-only trained in [37] - 57.67.

5.1.2 A2: Designing new adversarial datasets & the resilience of existing baseline models

For all adversaries bar Sbert-overlap, performance is below that offered for the mainstream datasets, or rather the source tasks from which each adversary was drawn. The general under-performance of the FEVER-only model — far below the main dataset accuracy of 95.07 on FEVER and 54.78 on VitaminC-real (from [37]) — is an indication of the presence of other biases in FEVER. Of particular note is the Entity Overweighting bias, which was the most potent of the new adversaries found. The results also suggest that VitaminC does not introduce its own biases, and the superior Resilience of VitaminC trained models over FEVER-only trained models serves as an indication that training on VitaminC does reduce the impact of biases other than claim-only bias on models trained in-part on FEVER or wholly on VitaminC.

There are some key caveats to these results, however. The first is the notably low performance of the both-trained model on the Negation overlap (Negation + Antonym variant) adversary. After further analysis, this was traced back to ALBERT-based models treat-

Model: Vitc-albert, Metric: Accuracy		Training Set			Potency	
Correctness	Adversary Name	Both	Fever-only	Vitc-only	Raw	Standard
0.64	Entity Overweighting	0.148	0.127	0.077	0.883	0.565
1	Word overlap (Jaccard)	0.923	0.887	0.831	0.120	
1	Word overlap (Sbert)	0.920	0.883	0.741	0.152	
1	Word overlap (Tf-idf)	0.908	0.908	0.777	0.136	
1	Negation Overlap (Manual, Neg+Ant)	0.742	0.672	0.766	0.273	
0.9	Negation Overlap (Semi-auto, Neg+Ant)	0.317	0.635	0.622	0.475	0.428
0.78	Negation Overlap (Semi-auto, Neg only)	0.952	0.950	0.817	0.094	0.073
1	Numerical Mismatch	0.871	0.375	0.880	0.291	
Resilience		0.689	0.627	0.698		

Table 5.1: Performance of Albert-base model on derived adversaries. Resilience is calculated using a single set of values for the three word overlap approaches (the average of the three).

ment of the subwords formed from the tokenization of antonymous verbs. As shown in the example across figure 5.2 and table 5.2, the both-trained model assigned a high (relative) responsibility in favour of the incorrect label (Refutes), to the token critical to non-both-trained models in determining that the example claim is Supported by the evidence. Subword tokenization[1] is designed to assign common words a space in the vocabulary, but to break up rarer words into *subwords*. These subwords are particularly common with this adversary, as the antonyms derived are often rarer. This is in part due to manual annotators preference for using precise but unpopular words — such as 'snubbed' — to ensure example correctness (i.e. to induce the exact semantics required to exclude alternative interpretations where an entailment relation other than that required for the example to be correct could be interpreted). Furthermore, Wordnet and Conceptnet also opt for returning rarer antonyms formed by prefixing derivational negation morphemes (e.g. un-, dis- as in *dis*establish). It is posited that this persistent use of subwords prevents proper realisation of their semantics, which when combined with the high responsibility assigned by the both-trained model, induces misclassification. The non-finetuned ALBERT model also assigns high responsibility to the subwords in question, but in favour of NEI. Further investigation is thus required to establish why training on both datasets resulted in this high responsibility being in favour of Refutes on the both-trained model, instead. One hypothesis is that this high responsibility is roughly equal to the sum of the lesser responsibilites of the two single-dataset trained models, and so training on both datasets produces a higher responsibility in favour of negation which when combined with the already high contextual focus indicated by the responsibilities of the ALBERT-base model, overpowers the less erroneous signals from elsewhere in the input.

[1] Performed for ALBERT by https://github.com/google/sentencepiece

Subword	Prediction	sn	ub	bed
FEVER-only	Supports	-0.67316	-0.57281	-0.41342
VitC-only	Supports	0.394417	-0.30762	0.139801
FEVER+VitC	Refutes	0.072532	-0.77997	-0.26893

Table 5.2: Responsibilities assigned by ALBERT to subwords formed from ‘snubbed’, for each of the models trained on the respective training sets. Responsibilities are obtained by Integrated Gradients analysis via Captum [24, 39]. Positive values are in support of assignment to Supports (the ground truth for this example), and negative values towards Refutes/NEI. Responsibilities for the entire tokenized entries can be found in the appendix.

> **Claim:** Humphrey Bogart was not snubbed for greatest male star of Classic American cinema .
> **Evidence:** In 1999 , the American Film Institute ranked Bogart as the greatest male star of Classic American cinema

Figure 5.2: Example of Negation overlap adversary predicted correctly (i.e., Supports) by the FEVER-only and VitaminC-only trained models, but incorrectly by the both-trained one.

The second caveat, is that adversary source task was not considered when calculating potency. There are notable stylistic differences between the claims in FEVER and those in VitaminC (see section 2.3), which may be convolved with the performances recorded on the different adversaries. For example, the Numerical Mismatch adversarial set was drawn entirely from mutating VitaminC test-set claims; the far-superior performance of VitaminC trained models may be down to having been trained on more stylistically similar examples. This superiority is otherwise surprising, insofar as Numerical Mismatches were not observed to be common in FEVER itself.

5.2 RQB - How well can dataset-agnostic approaches inoculate systems against different biases?

5.2.1 Bert vs Albert, and issues of (non)-convergence.

For the best performing models, self-debias improves on BERT based models with either fever or fever/vitc training, but not on albert+fever/vitc, and only sligtly on albert+fever.

Results were initially recorded solely for ALBERT-base underpinned models, as to enable comparison with the findings of [37]. However, these results suggested the opposite to that of Utama et al.[47] — namely that, **for ALBERT-base models, all forms of debiasing degraded performance on both FEVER and FEVER-symmetric**. Whilst the results of Utama et al. were not expected to be exactly reproduced (as the VitaminC training parameters differed, and there was no seed given by Utama et al.), such divergent results prompted further investigation into the assumption that ALBERT and BERT

BERT/FEVER only / Shallow		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	76.6%	63.2%	51.6%	63.8%
Best (Ao3)	conf-reg-anneal	76.6% / 0%	64.5% / 1.3%	54.7% / 3.1%	65.3% / 1.5%
Best (NBR)	conf-reg	76.6% / 0%	65.9% / 2.8%	55.0% / 3.4%	65.8% / 2.1%

BERT/Both / Shallow		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	87.5%	76.3%	67.0%	76.9%
Best (Ao3)	reweight	87.7% / 0.2%	77.1% / 0.8%	69.1% / 2.1%	78.0% / 1.1%
Best (NBR)					

ALBERT/ Both / Shallow		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	91.4%	80.2%	71.7%	81.1%
Best (Ao3)	conf_reg_anneal	81.0% / -10.5%	77.2% / -3.0%	60.0% / -11.6%	72.7% / -8.4%
Best (NBR)	reweight	47.1% / -44.4%	43.8% / -36.4%	69.7% / -1.9%	53.5% / -27.6%

ALBERT/Fever only / shallow		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	80.5%	73.3%	52.3%	68.7%
Best (Ao3)	poe	80.7% / 0.2%	72.4% / 0.9%	54.8% / 2.5%	69.3% / 0.6%
Best (NBR)	conf-reg	71.6% / -8.9%	63.3% / 10%	55.8% / 3.5%	63.6% / -5.1%

Figure 5.3: Best performing adversaries for shallow-trained models compared with baseline models. Ao3 is best (non-baseline) model according to average of three metrics, NBR is best according to highest new-biases resilience. Deltas with baseline follow after slashes.

would respond to debiasing similarly, by re-running experiments on BERT-base-uncased models. Further investigation is required to determine why this is, with this effect also being observed by [21], who suggest that ALBERT may be more sensitive to fine-tuning at the risk of losing pre-trained information during debiasing.

A further issue was the number of models which failed to converge at training time. This was all ALBERT-based self-debiased models trained on VitaminC or both datasets, with a PoE or PoE-annealed loss function, and the both-trained Confidence Regularisation and model to a lesser extent. The phenomena of non-convergence is arguably wider than this for ALBERT models, with both the fever-trained PoE-annealed model and all confidence regularization models also demonstrating a minimal decrease in training loss over the duration of training. Confidence regularization, however, works by scaling down the input of a teacher, with the 'teacher' here being that of the strong-performing non-debiased/baseline model. These models, therefore, can achieve superlative performance with very little fine tuning, as they are effectively a wrapper over the high-performing baseline. Evaluating model checkpoints indicates that this stationary training loss is not masking training which impacts adversarial datasets only — the performance on fever-symmetric also remains the same when evaluating it as a form of out-of-domain validation loss. Such stationary loss would usually be indicative of a model lacking capaciousness, which would suggest that ALBERT is unsuitable for debiasing in this manner.

Fever-only trained models				
Architecture	*ALBERT-base*		*BERT-base-uncased*	
Loss Function\Test Set	fever	fever-symmetric	fever	fever-symmetric
PoE	0.949872	0.813202	0.869428	0.733146
Reweight	0.94082	0.787921	0.870453	0.745787
(None / Cross-entropy)	0.948249	0.81882	0.869769	0.733146
Confidence Regularization	0.873783	0.65309	0.871904	0.745787
Confidence Regularization (Annealed)	0.887703	0.672753	0.872161	0.759831
Reweight (Annealed)	0.907857	0.676966	0.870367	0.758427
PoE (Annealed)	0.626473	0.5	0.869257	0.726124

Table 5.3: Comparison of baseline models across architectures, when trained solely on FEVER

5.2.2 B1: Debiased model Resilience & Improving adversarial performance without compromising non-adversarial performance.

For all models which are deemed viable (i.e. converged properly during fine-tuning), **adversarial performance gains were associated with either no change or a very minimal drop in performance on the 'mainstream' datasets**, here represented by an average of the FEVER and VitaminC accuracy/resilience (with the VitaminC accuracy itself an average of the vitaminC-real and vitaminC-synthetic test sets). These results were consistent across both hypothesis-only biases (fever-symmetric) and the newly devised biases.

For BERT-based models, on the whole, adversarial performance gains do not come at the expense of mainstream performance. As was observed in Utama, average performance on FEVER and Vitaminc from models trained solely on FEVER, was mostly unchanged, with a maximum decrease of 0.6pp. This was contrasted with the +3.4/3.1pp resilience gain on the new biases, and +2.8/1.3pp from the confidence regularisation (non-annealed/annealed respectively). Whilst these results offer little new insight over [47], they do reproduce their findings in a way that vindicates the implementation of this investigation, subject to the caveat described in section 5.2.1. These patterns were repeated for BERT models trained on both VitaminC and Fever, with the best model gaining 2.1pp in Resilience for the new biases, without trading-off any of the mainstream performance, for reasons explored in section 5.2.2.

ALBERT models trained solely on FEVER concur with this general finding, with PoE and reweight delivering improved resilience on the new adversaries (+2.5pp/2.1pp respectively) with no change (0pp and -0.5pp respectively) to the mainstream performance. This improvement is more apparent when you discount Sbert on account of it simply being a subset of FEVER and perhaps not worth including in the adversarial comparisons — rising to 3.5pp/2.9pp. Whilst the PoE model matched the performance of the non-debiased

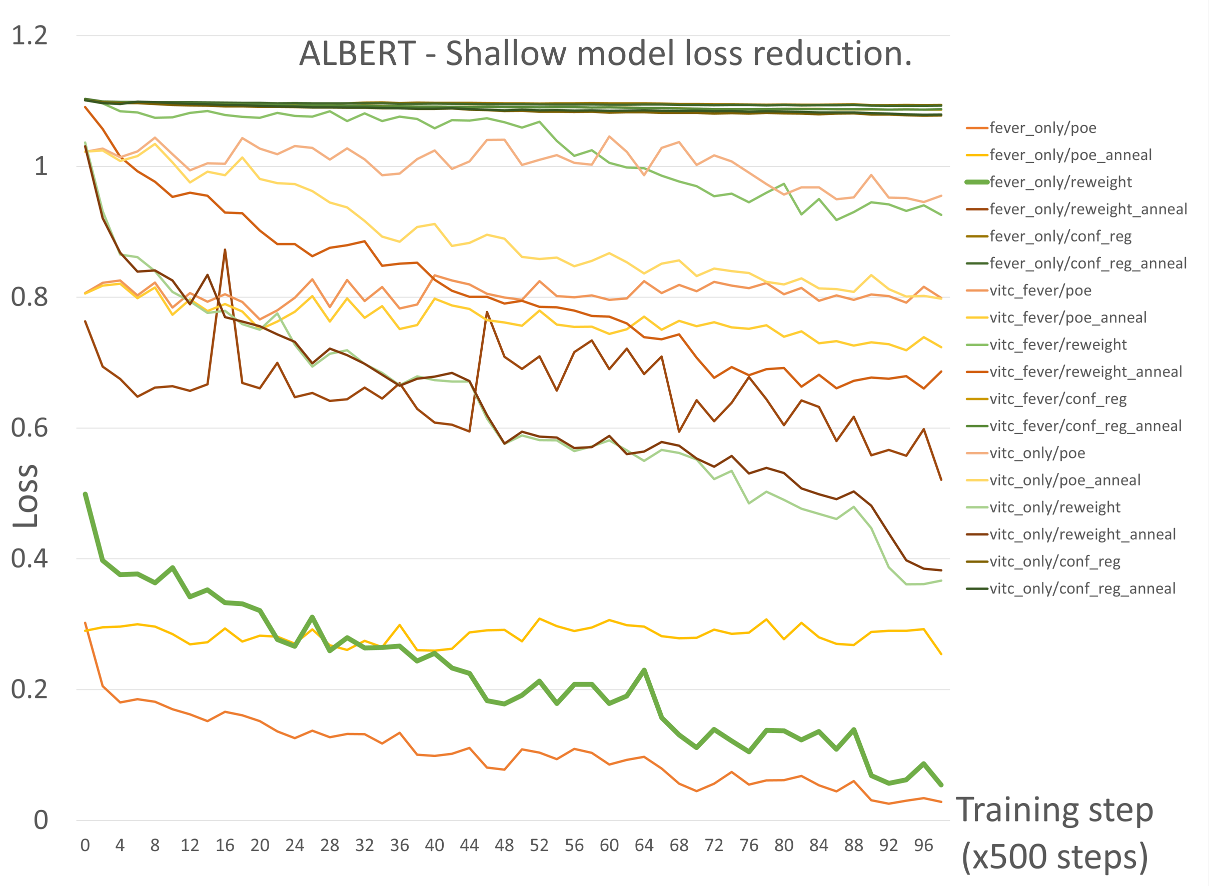

Figure 5.4: Loss decrease on ALBERT shallow models at training time. Lines clustered at top are all Confidence Regularisation models.

model on fever_adversarial and fever_symmetric, the Reweight model made more of a trade off between fever_symmetric (+2pp on the non-debiased model) and fever_symmetric (-3.1pp). This serves as a good example of the risk of evaluating bias-agnostic debiasing by only considering one adversarial dataset, as was done in Utama et al.[47]. The confidence regularization (without annealing) achieves the best performance on the new biases among FEVER-only trained ALBERT models, gaining 3.5pp (including Sbert) but trades this off against mainstream performance, dropping 8.9pp.

The exception to this overall pattern is for ALBERT models trained on both FEVER and VitaminC, for which any debiasing drops performance across the board — an immediate drop of 11.7pp on new biases resilience and on mainstream accuracy, for reasons previously debated in section 5.1.2. Here, the shallow reweighted model is an example of debiasing working effectively at the cost of mainstream performance- resilience is only 1.9pp less than the base model for the new biases, but at a drop of 44.4pp on the mainstream performance. The performance of ALBERT models trained solely on VitaminC is similar, exhibiting a smaller drop from the baseline to the next-best performing models (7.5pp on mainstream and adversarial/symmetric average, 5.7pp on the new biases resilience).

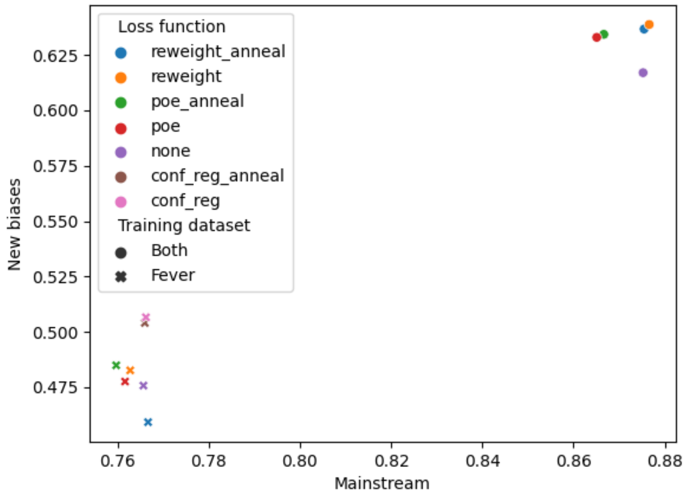

Figure 5.5: Trade-off for BERT models between Mainstream and new-biases performance.

5.3 B1.1: Comparing self-debiasing loss functions.

None of the three self-debiasing approaches employed performed the best outright. For example, fine-tuning a BERT-base-uncased model using Example Reweighting yielded the BERT-based model which achieved the best resilience across all of fever-symmetric, the mainstream datasets, and the new adversaries derived. However, when training with the same parameters on an ALBERT-base model instead, superlative performance on the new biases is maintained but average accuracy on the mainstream test sets is just half that of the baseline (non fine-tuned) model, once again attributable to [21] hypothesis that ALBERT's increased sensitivity to fine-tuning risks degrading mainstream performance via forgetting, something also referred to by [43]. This may also explain why this diverging-by-architecture behaviour is not present in models trained on fever alone (and thus, is fine-tuned on less examples), where performance resembled that of Product of Expert, which also achieved sub-par performance on ALBERT-base models when trained on either just VitaminC or on VitaminC and FEVER — in this case due to not converging.

As found by Utama et al., Confidence regularization exhibited the greatest performance on FEVER-trained BERT models. These two models (i.e. non-annealed and annealed) had abnormally varied performance *across* the newly introduced biases, with a far higher performance on numerical_mismatch (+ 8.5pp over poe/reweight/baseline), at the cost of sbert-overlap performance (-4.0pp from next lowest - reweight). For ALBERT-base models, Confidence Regularization performance was significantly better at mainstream for both VitaminC and both-trained models. Confidence regularisation models also performed poorly across all metrics on ALBERT-based models trained solely on FEVER. As confidence regularisation works by scaling the teacher model (i.e, the predictions on the training set of the baseline model trained on that same dataset) by the shallow model, comparisons can be drawn with models trained on either one or both input sources (figure

Training dataset	Loss Function	Bias used	Uses Teacher?	Uses Shallow?	FEVER Test accuracy	Average Confidence
FEVER only	Conf. Reg	Ent Overweighting	✓	×	0.954569	0.206
	Conf. Reg	Shallow	✓	✓	0.873783	0.034
	(baseline)		✓*	×	0.948249	0.995
	PoE	Shallow	×	✓	0.949872	0.996

Table 5.4: Comparison of performance of selection of ALBERT-based fever-only trained models. **asterisk** indicates that baseline does not technically use baseline, but is trained equivalently to the baseline. Average Confidence is average difference between probability of highest-scoring label (i.e, the prediction made) and second-highest scoring - more confident predictions will have higher values.

5.4). This suggests it is the combination of the fever-trained shallow model and fever-trained teacher/baseline which were combining destructively as to reduce confidence on all examples to very low values.

Notably, none of these (non-annealed) loss functions require hyper-parameters to be tuned. This excludes the possibility of erroneous comparison on the basis of loss function misconfiguration, however this does not exclude outcomes causing results to differ from those found by [47] originating in the use of a different training framework and random state. That said, these findings do suggest against declaring a universally superior self-debiasing method. Comparison of the various loss functions was hampered by the number of models which failed to converge, as well as by the high dimensionality of the state space of models trained. For this reason — even aside from the inconclusive indication of a superior loss function — these results cannot serve as a conclusive means of determining a superior self-debiasing loss function.

5.3.1 Annealing

Concurring with the findings of Utama et al, annealing on BERT models remains superfluous (for models trained on FEVER and VitaminC) — with this investigation finding it may even be detrimental in some cases — such as substantially lowering performance on numerical-overweight when adding annealing to models trained with Example Reweighting on FEVER only. This extends to ALBERT models also trained solely on FEVER, where adding annealing dropped performance broadly on all adversaries. The exception to this is adding annealing to Confidence Regularisation, where adding annealing fulfilled the original aim of annealing by improving mainstream performance by 1.1pp, and not at the expense of adversarial performance.

For ALBERT models trained on both Fever and VitaminC, adding annealing to Example Reweighting improves mainstream (+18.9pp), entity-overlap (+5.6pp) and sbert (+4.7pp) drastically, but drops negation overlap and numeric overweight by 17.5/18.9pp respectively. A weaker verison of this pattern is also evident when adding annealing to Confidence Regularization. Considering sbert's performance as more of a reflection of mainstream performance than as an adversary in its own right (see section 4.4), this would appear to show annealing working as intended by Utama et al. — preserving mainstream performance whilst not degrading overall resilience too severely.

Overall, there is no conclusive evidence that annealing can deliver a general improvement in performance. This finding contradicted that of [47], and so was investigated further by trialing different parameters for a, none of which providing substantially different results.

5.3.2 B2: Comparing training datasets.

Comparing baseline performance alone on the mainstream test sets, there is an evident benefit to training on both VitaminC and FEVER[2]. With regard to debiasing, with an ALBERT-based model, debiasing had a minimal effect on models trained solely with FEVER and which did converge. The results indicate a net gain on resilience to the new biases, but this is an oversimplification of what was actually a drop in performance in 3 different biases offset against a large (+12-15pp) gain on Numerical Mismatch performance, with this effect also apparent in BERT-trained (but otherwise identical) models, albeit to a lesser extent.

Debiasing ALBERT-based models trained on VitaminC (exclusively or otherwise) consistently degraded performance on the mainstream datasets as well as on the new biases. This is not to say that VitaminC should not be trained on; the best performing model (over both ALBERT-based *and* BERT-based collectively) over both the fever-symmetric/fever-adversarial average, the mainstream performance, and the new biases, is the baseline/non-debiased ALBERT model trained on both VitaminC and FEVER. On the new biases, even the detrimentally-debiased both-trained models outperform the beneficially-debiased fever-only trained models. An exception to this is the universal increase in performance on Entity Overweighting for models trained solely on VitaminC — albeit at the expense of mainstream test set performance. The otherwise complete supremacy of models trained on both datasets entirely diminishes discussions around the efficacy of debiasing ALBERT-based models.

In direct contrast, for BERT-based models, debiasing had a greater positive impact on models trained on both FEVER *and* VitaminC. It is thus apparent that it is the *combination* of the training set and architecture which is the key determinant of both baseline performance and debiasing efficacy. That is, ALBERT trained on FEVER only, and BERT trained on both FEVER and VitaminC, are the combinations of base model and training set which are the most responsive to the self-debiasing methods.

		Correctly Predicted				
		FEVER-debias only	Vitc-base only	Both	Neither	TOTAL
Fever +vitc +debias	Correct	1480	0	12950	0	14430
	Incorrect	731	2891	0	5368	8990
	TOTAL	2211	2891	12950	5368	23420

Table 5.5: Table showing overlap between examples correctly predicted between FEVER-debias (BERT model trained on FEVER with Example Reweighting), VitC-base (BERT baseline (i.e. non-debiased) model trained on VitaminC only), and FEVER+vitc+debias (BERT model trained on both FEVER and VitC, using Example Reweighting). 23420 examples are a 50/25/25 split between FEVER-test, VitC-Real test and VitC-Synth test.

5.3.3 Exploring the orthogonality in training with both model-targeting and dataset-targeting debiasing

As evident from Table 5.5, training on VitaminC does result in a distinct group of examples being predicted correctly than are by a model trained with self-debiasing just on FEVER. Whilst this could be attributed to the stylistic differences between FEVER and VitaminC manifesting here as superior in-domain performance, the fact that 26.7% of those predicted correctly solely by the VitaminC trained non-debiased model are sourced from the FEVER test set would indicate there is at least some aspect of commonality between the datasets being acted on here. There is evidently an aspect of orthogonality between the effective a-priori debiasing conferred by training on VitaminC (i.e., the work of Schuster et al. [37]), and the bias-agnostic debiasing achieved by self-debias (i.e., the work of Utama et al. [47]). This difference reflects that observed in section 5.1.2, suggesting that this aspect of orthogonality does correspond to the respective approaches ameliorating different biases. The combination of the two methods, as incorporated in this investigation, appears to fail to incorporate these orthogonal functionalities constructively — as evidenced by the fact that 100% of those examples correctly predicted by a VitaminC-trained non-debiased model are then incorrectly classified when evaluated on a debiased model trained on both FEVER and VitaminC, the non-debiased version of which was found by [37] to be the superlatively performing model. This is not to suggest that training on VitaminC is inadvisable; models trained on both datasets are universally superior to those trained on one, as has been explored. There is, however, ground for investigating whether ensemble methods may be a better approach to integrating the debiasing approaches of [37] and [46].

[2] See Schuster et al. [37] for discussion on optimal training proportions of each dataset

5.4 Extension: How well do bias-agnostic approaches perform against hand-crafted known-bias ones?

For BERT-trained models, a custom-designed known-bias was run for the negation_overlap, and delivered notable performance improvement for negation_overlap and slightly boosting performance on both the baseline datasets and, for the non-annealed versions, notable improvements to fever-symmetric (1.4-1.8pp increase). Improvements on the other ('out-of-domain') adversaries is weaker than with self-debias — most notably on sbert-overlap (2.6-3.6pp drop) and entity-overweighting (7.7-2.8pp drop, with no change on annealed-reweighting). This slight overall performance increase both overall and on test-sets for the adversary specified for debiasing is consistent with the findings of Utama et al. [47].

For ALBERT-trained models, the known-entity-overweighting manually specified bias was used, and resulted in consistent gains, reversing the trend of debiasing causing losses on ALBERT-based models, with gains averaging (mean) +10pp on Sbert, +7pp on numerical-overweight, +12pp on fever-symmetric, 10pp on FEVER and 8pp on VitaminC-real.

Debiasing with known biases also mostly ameliorated the issue of non-convergence during training, suggesting that the agnosticism of self-debias' shallow-model may be considering too many types of biases at once to produce meaningful guidance to the trainer to begin working to reduce a single one initially. The magnitude of improvement observed was initially attributed to the use of a different known-bias source, but further experiments showed this not to be the case (i.e. similar increases were observed when using the bias used with BERT-trained models). Instead, the magnitude of improvement may be attributable to the combination of [21]'s hypothesis that ALBERT is more sensitive to fine-tuning (i.e. there is less inertia in the pre-trained embeddings) and the fact that using known-biases adds information to the system (here, information which can directly downweigh tokens which would otherwise misguide classification). Increasing information available is a measure of feature extraction efficacy, which in turn improves classification accuracy [8, page 2].

ALBERT/ Both / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	91.4%	80.2%	71.7%	81.1%
Best (Ao3) / delta	conf_reg	91.4% / 0%	80.3% / 0.1%	72.7% / 1.1%	81.5% / 0.4%
Best (NBR) / delta					

ALBERT / Fever only / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	80.5%	73.3%	52.3%	68.7%
Best (Ao3) / delta	conf_reg_anneal	82.5% / 2%	76.9% /3.6%	71.6% / 19.3%	77.0% / 8.3%
Best (NBR) / delta					

BERT / Both / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	87.5%	76.3%	67.0%	76.9%
Best (Ao3) / delta	reweight_anneal	87.8% / 0.3%	77.3% / 1%	68.1% / 1.1%	77.7% / 0.8%
Best (NBR) / delta					

Figure 5.6: Best performing adversaries from known-bias training, compared with baseline models. Ao3 is best (non-baseline) model according to average of three metrics, NBR is best according to highest new-biases resilience.

Chapter 6: Summary and conclusions

Devising robust fact-verification systems requires inference systems to fully engage with the semantics of both the claim and the evidence. Whilst past work [37] had looked at debiasing FEVER systems using contrastive adversaries based on a-priori knowledge of hypothesis-only bias, the existence of other biases in the similar tasks of (S)NLI [29] suggested the existence of other biases in FEVER. This investigation sought to explore the existence of these other biases, and whether VitaminC can extend to effectively tackling them. Whilst training with VitaminC constitutes a *dataset-targeting* approach, Utama et al's Self-debias [47] was dataset-agnostic, instead using custom loss functions to selectively downweigh examples deemed as biased by a *shallow* model, which identifies examples which can be quickly classified and are thus indicative of being biased. This investigation further sought to determine whether these two approaches could be reconciled for superior performance.

In pursuit of these aims, a new set of adversaries were first devised by creating custom-written mutations for VitaminC and FEVER, making use of text generation models (T5), lexical databases to derive antonyms, and textual similarity models (Sbert). These included an adversary which formed random sentences featuring the same named entities as existing examples (Entity Overweighting), an adversary of claims refuting evidence which contains mostly the same tokens (Sbert-overlap), an adversary which exploits VitaminC's overuse of numerical ranges (Numerical-overlap) and an adversary which exploited expectations of a mismatch in the presence of negation words in the claim and hypothesis (Negation-mismatch). As an extension, a-priori bias files were also generated to identify the new biases in the respective task training sets, as to enable later comparison between self-debias and debiasing with known biases.

Having designed these adversaries, the methods of self-debias and VitaminC were reconciled — including rewriting some of self-debias to function with a different framework — and experiments were then ran to determine the *potency* of the new adversaries, and the *resilience* of a range of proposed models to them alongside these models performance on the mainstream datasets themselves and on standard adversarial measures. These models were initially based on ALBERT - as to allow comparison with those in [37], and were instantiated for each example in a wide state-space. The first dimension was over the three loss-functions used in Utama et al. [47], the second over combinations of training sets, (FEVER only, VitaminC only, or both), and the third where self-debias or a known a-priori bias indication was used. After it became apparent that ALBERT responded distinctly from the trends observed with *BERT* based models in [47], further experiments were ran with BERT based models.

With the obtained results, as summarised in the opening of chapter 5, the set out research questions could be tackled. It is first established (for RQA) that the use of contrastive evidence-pairs *does* inoculate fact-checking systems against biases other than the hypothesis-only biases originally intended, as evidenced by the superior Resilience training on VitaminC confers on models both against the hypothesis-only bias, as well as the newly discovered biases.

The findings for RQB are less clear-cut, with the answer as to how well dataset-agnostic approaches inoculate fact-checking systems against different biases depending on the combination of base model and training set.Overall, though, the use of self-debias does improve resilience to the new biases, with the key exception of ALBERT-based models trained on both datasets.

Turning to the extension question, debiasing with hand-crafted known-bias inputs outperforms self-debias on both ALBERT and BERT based models, with this difference in performance being more pronounced for models trained solely on FEVER. Known-biases usually improve performance on the adversary they were designed to target, but this is sometimes at the expense of performance on other adversaries.

Whilst comprehensive data has been obtained in the course of this investigation, the reasoning behind some of the more surprising results warrants further work. Of particular importance is ascertaining why the methedology of self-debias did not translate fully from BERT based models to ALBERT based models — something which could build on the hypothesis of Kaneko *et al.*[21]. The dimemnsionality of the state space of possible models to be trained was already large for this investigation, and the need to train another set of models to derive results for BERT only served to increase this space further, something which could have been avoided with a deeper understanding of why self-debias may not translate as well to ALBERT based models. This is of particular importance given that ALBERT models outperform BERT-base models on most inference tasks, including on FEVER and its adversarial sets, and so debiasing methods which work only on inferior models may limit its future use. Furthermore, a comparison of the dynamics of the three loss functions introduced - beyond that offered in [47] — would be beneficial for a full understanding of why their performance differs, sometimes drastically. Research into automated systems of antonym generation would also be beneficial in scaling up the negation overlap adversary into automated example generation; there is currently no effective way to generate relevant antonyms as to reverse the meaning of a text. Finally, investigation into the stylistic characteristics of VitaminC and the new adversaries derived here may be beneficial in determining if adversarial performance is affected by the examples being less fluently written, more so than the actual semantics of the examples.

Automated fact-checking systems can be a powerful tool for tackling misinformation, provided they engage with the semantics of the data being evaluated. Debiasing these models is thus a crucial step in achieving robust fact-checking systems. The work of this investigation lays out the obstacles to using promising dataset-agnostic debiasing methods on superior-performing ALBERT models, in the hope that future work can produce superior systems which do not fall vulnerable to the range of new attacks demonstrated.

Bibliography

[1] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–236. ISSN: 0895-3309. DOI: 10.1257/jep.31.2.211. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211 (visited on 05/17/2022).

[2] Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. *Generating Label Cohesive and Well-Formed Adversarial Claims.* arXiv:2009.08205. type: article. arXiv, Sept. 17, 2020. arXiv: 2009.08205[cs]. URL: http://arxiv.org/abs/2009.08205 (visited on 05/27/2022).

[3] Shakuntala Banaji et al. "WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India". In: (Nov. 2019), p. 62.

[4] Zapan Barua et al. "Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation". In: *Progress in Disaster Science* 8 (Dec. 1, 2020), p. 100119. ISSN: 2590-0617. DOI: 10.1016/j.pdisas.2020.100119. URL: https://www.sciencedirect.com/science/article/pii/S2590061720300569 (visited on 05/26/2022).

[5] Yonatan Belinkov et al. "On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference". In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019).* *SEM 2019. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 256–262. DOI: 10.18653/v1/S19-1028. URL: https://aclanthology.org/S19-1028 (visited on 05/17/2022).

[6] Michele Bevilacqua and Roberto Navigli. "Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* ACL 2020. Online: Association for Computational Linguistics, July 2020, pp. 2854–2864. DOI: 10.18653/v1/2020.acl-main.255. URL: https://aclanthology.org/2020.acl-main.255 (visited on 05/27/2022).

[7] Gagan Bhatia. *keytotext.* Publication Title: GitHub. URL: https://github.com/gagan3012/keytotext.

[8] Christopher M. Bishop. *Pattern recognition and machine learning / Christopher M. Bishop.* New York, NY : Springer, [2006], 2006.

[9] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 632–642. DOI: `10.18653/v1/D15-1075`. URL: `http://aclweb.org/anthology/D15-1075` (visited on 05/18/2022).

[10] Sarah Cahlan. "Analysis — How misinformation helped spark an attempted coup in Gabon". In: *Washington Post* (Feb. 2020). ISSN: 0190-8286. URL: `https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/` (visited on 05/26/2022).

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. "Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases". 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).* EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4069–4082. DOI: `10.18653/v1/D19-1418`. URL: `https://aclanthology.org/D19-1418` (visited on 05/27/2022).

[12] Carme Colomina, Héctor SÁNCHEZ Margalef, and Richard Youngs. "The impact of disinformation on democratic processes and human rights in the world". In: (Apr. 2021), p. 64.

[13] United Nations Educational et al. "Journalism, fake news & disinformation: handbook for journalism education and training". In: (2018). Publisher: UNESCO. ISSN: 9231002813.

[14] Herbert P. Grice. "Logic and conversation". In: *Speech acts.* Brill, 1975, pp. 41–58.

[15] Suchin Gururangan et al. "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 107–112. DOI: `10.18653/v1/N18-2017`. URL: `https://aclanthology.org/N18-2017` (visited on 05/17/2022).

[16] Maj-Britt Mosegaard Hansen and Erling Strudsholm. "The semantics of particles: advantages of a contrastive and panchronic approach: a study of the polysemy of French déjà and Italian già". In: 46.3 (May 1, 2008). Publisher: De Gruyter Mouton Section: Linguistics, pp. 471–505. ISSN: 1613-396X. DOI: `10.1515/LING.2008.016`. URL: `https://www.degruyter.com/document/doi/10.1515/LING.2008.016/html` (visited on 05/20/2022).

[17] Naeemul Hassan et al. "The Quest to Automate Fact-Checking". In: (), p. 5.

[18] He He, Sheng Zha, and Haohan Wang. “Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual”. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 132–142. DOI: 10.18653/v1/D19-6115. URL: https://aclanthology.org/D19-6115 (visited on 05/27/2022).

[19] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1, 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[20] Paul Jaccard. “The Distribution of the Flora in the Alpine Zone.1”. In: *New Phytologist* 11.2 (1912). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x, pp. 37–50. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.1912.tb05611.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x (visited on 05/20/2022).

[21] Masahiro Kaneko and Danushka Bollegala. “Debiasing Pre-trained Contextualised Embeddings”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. EACL 2021. Online: Association for Computational Linguistics, Apr. 2021, pp. 1256–1266. DOI: 10.18653/v1/2021.eacl-main.107. URL: https://aclanthology.org/2021.eacl-main.107 (visited on 05/25/2022).

[22] N. Karlova and K. Fisher. “A social diffusion model of misinformation and disinformation for understanding human information behaviour”. In: *Inf. Res.* (2013).

[23] Hye Kyung Kim and Edson C. Tandoc. “Consequences of Online Misinformation on COVID-19: Two Potential Pathways and Disparity by eHealth Literacy”. In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2022.783909 (visited on 05/26/2022).

[24] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. arXiv:2009.07896. type: article. arXiv, Sept. 16, 2020. DOI: 10.48550/arXiv.2009.07896. arXiv: 2009.07896[cs,stat]. URL: http://arxiv.org/abs/2009.07896 (visited on 05/27/2022).

[25] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. *End-to-End Bias Mitigation by Modelling Biases in Corpora*. arXiv:1909.06321. type: article. arXiv, Apr. 23, 2020. DOI: 10.48550/arXiv.1909.06321. arXiv: 1909.06321[cs]. URL: http://arxiv.org/abs/1909.06321 (visited on 05/18/2022).

[26] Tom McCoy, Ellie Pavlick, and Tal Linzen. “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*. ACL 2019. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. URL: https://aclanthology.org/P19-1334 (visited on 05/18/2022).

[27] Ben Medeiros and Pawan Singh. "Addressing Misinformation on Whatsapp in India Through Intermediary Liability Policy, Platform Design Modification, and Media Literacy". In: *Journal of Information Policy* 10 (2020). Publisher: Penn State University Press, pp. 276–298. ISSN: 2381-5892. DOI: 10.5325/jinfopoli.10.2020.0276. URL: https://www.jstor.org/stable/10.5325/jinfopoli.10.2020.0276 (visited on 05/26/2022).

[28] Allan H. Murphy. "The Finley Affair: A Signal Event in the History of Forecast Verification". In: *Weather and Forecasting* 11.1 (Mar. 1, 1996). Publisher: American Meteorological Society Section: Weather and Forecasting, pp. 3–20. ISSN: 1520-0434, 0882-8156. DOI: 10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/wefo/11/1/1520-0434_1996_011_0003_tfaase_2_0_co_2.xml (visited on 05/20/2022).

[29] Aakanksha Naik et al. "Stress Test Evaluation for Natural Language Inference". In: (June 2018), p. 14.

[30] Yixin Nie and Mohit Bansal. "Shortcut-Stacked Sentence Encoders for Multi-Domain Inference". In: *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 41–45. DOI: 10.18653/v1/W17-5308. URL: https://aclanthology.org/W17-5308 (visited on 05/18/2022).

[31] Adam Poliak et al. "Hypothesis Only Baselines in Natural Language Inference". In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 180–191. DOI: 10.18653/v1/S18-2023. URL: http://aclweb.org/anthology/S18-2023 (visited on 05/18/2022).

[32] Ethan Porter and Thomas J. Wood. "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom". In: *Proceedings of the National Academy of Sciences of the United States of America* 118.37 (Sept. 14, 2021), e2104235118. ISSN: 0027-8424. DOI: 10.1073/pnas.2104235118. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8449384/ (visited on 05/17/2022).

[33] Princeton University. *WordNet — A Lexical Database for English*. 2010. URL: https://wordnet.princeton.edu/ (visited on 05/27/2022).

[34] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv:1910.10683. type: article. arXiv, July 28, 2020. arXiv: 1910.10683[cs,stat]. URL: http://arxiv.org/abs/1910.10683 (visited on 05/20/2022).

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Semantically Equivalent Adversarial Rules for Debugging NLP models". In: *Annual Meeting of the Association for Computational Linguistics (ACL)* (Jan. 2018). DOI: 10.18653/

v1/P18-1079. URL: https://par.nsf.gov/biblio/10100478-semantically-equivalent-adversarial-rules-debugging-nlp-models (visited on 05/18/2022).

[36] Elyse Samuels. "Analysis — How misinformation on WhatsApp led to a mob killing in India". In: *Washington Post* (Feb. 2020). ISSN: 0190-8286. URL: https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/ (visited on 05/26/2022).

[37] Tal Schuster, Adam Fisch, and Regina Barzilay. *Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence.* arXiv:2103.08541. type: article. arXiv, Mar. 15, 2021. arXiv: 2103.08541[cs]. URL: http://arxiv.org/abs/2103.08541 (visited on 05/18/2022).

[38] Tal Schuster et al. "Towards Debiasing Fact Verification Models". 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).* EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3419–3425. DOI: 10.18653/v1/D19-1341. URL: https://aclanthology.org/D19-1341 (visited on 05/18/2022).

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks.* arXiv:1703.01365. type: article. arXiv, June 12, 2017. DOI: 10.48550/arXiv.1703.01365. arXiv: 1703.01365[cs]. URL: http://arxiv.org/abs/1703.01365 (visited on 05/27/2022).

[40] Christian Szegedy et al. *Intriguing properties of neural networks.* arXiv:1312.6199. type: article. arXiv, Feb. 19, 2014. arXiv: 1312.6199[cs]. URL: http://arxiv.org/abs/1312.6199 (visited on 05/18/2022).

[41] J. Thorne et al. *FEVER: A large-scale dataset for fact extraction and verification.* Accepted: 2019-06-07T23:30:05Z. Association for Computational Linguistics, 2018. ISBN: 978-1-948087-27-8. DOI: 10.17863/CAM.40620. URL: https://www.repository.cam.ac.uk/handle/1810/293476 (visited on 05/18/2022).

[42] James Thorne and Andreas Vlachos. "Automated Fact Checking: Task Formulations, Methods and Future Directions". In: *Proceedings of the 27th International Conference on Computational Linguistics.* COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3346–3359. URL: https://aclanthology.org/C18-1283 (visited on 05/17/2022).

[43] James Thorne and Andreas Vlachos. *Elastic weight consolidation for better bias inoculation.* arXiv:2004.14366. type: article. arXiv, Feb. 4, 2021. arXiv: 2004.14366[cs]. URL: http://arxiv.org/abs/2004.14366 (visited on 05/18/2022).

[44] James Thorne et al. "Evaluating adversarial attacks against multiple fact verification systems". 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).* EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2944–2953. DOI:

10.18653/v1/D19-1292. URL: https://aclanthology.org/D19-1292 (visited on 05/18/2022).

[45] James Thorne et al. “The Fact Extraction and VERification (FEVER) Shared Task”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 1–9. DOI: 10.18653/v1/W18-5501. URL: https://aclanthology.org/W18-5501 (visited on 05/26/2022).

[46] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. “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*. ACL 2020. Online: Association for Computational Linguistics, July 2020, pp. 8717–8729. DOI: 10.18653/v1/2020.acl-main.770. URL: https://aclanthology.org/2020.acl-main.770 (visited on 05/18/2022).

[47] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. *Towards Debiasing NLU Models from Unknown Biases*. arXiv:2009.12303. type: article. arXiv, Oct. 13, 2020. DOI: 10.48550/arXiv.2009.12303. arXiv: 2009.12303[cs]. URL: http://arxiv.org/abs/2009.12303 (visited on 05/18/2022).

[48] Andreas Vlachos and Sebastian Riedel. “Fact checking: Task definition and dataset construction”. In: *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 2014, pp. 18–22.

[49] Wenhui Wang et al. *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*. arXiv:2002.10957. type: article. arXiv, Apr. 5, 2020. arXiv: 2002.10957[cs]. URL: http://arxiv.org/abs/2002.10957 (visited on 05/27/2022).

[50] Adina Williams, Nikita Nangia, and Samuel Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: http://aclweb.org/anthology/N18-1101 (visited on 05/26/2022).

[51] Rowan Zellers et al. “HellaSwag: Can a Machine Really Finish Your Sentence?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4791–4800. DOI: 10.18653/v1/P19-1472. URL: https://aclanthology.org/P19-1472 (visited on 05/27/2022).