The Earth Is Flat and the Sun Is Not a Star

The Susceptibility of GPT-2 to Universal Adversarial Triggers

Hunter Scott Heidenreich
Jake Ryland Williams
hsh28@drexel.edu
jw3477@drexel.edu
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania, USA

ABSTRACT

This work considers *universal adversarial triggers*, a method of adversarially disrupting natural language models, and questions if it is possible to use such triggers to affect both the topic and stance of conditional text generation models. In considering four "controversial" topics, this work demonstrates success at identifying triggers that cause the GPT-2 model to produce text about targeted topics as well as influence the stance the text takes towards the topic. We show that, while the more fringe topics are more challenging to identify triggers for, they do appear to more effectively discriminate aspects like stance. We view this both as an indication of the dangerous potential for controllability and, perhaps, a reflection of the nature of the disconnect between conflicting views on these topics, something that future work could use to question the nature of filter bubbles and if they are reflected within models trained on internet content. In demonstrating the feasibility and ease of such an attack, this work seeks to raise the awareness that neural language models are susceptible to this influence–even if the model is already deployed and adversaries lack internal model access–and advocates the immediate safeguarding against this type of adversarial attack in order to prevent potential harm to human users.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; *Neural networks*; • **Security and privacy** → Social aspects of security and privacy.

KEYWORDS

Natural Language Processing; Adversarial Attacks; Bias; Language Modeling

ACM Reference Format:
Hunter Scott Heidenreich and Jake Ryland Williams. 2021. The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3461702.3462578

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462578

1 INTRODUCTION

There is an ever-growing body of literature that demonstrates semantics learned from human-generated texts contain human-like biases. These biases are then ingested, replicated, and reinforced by neural language models, from static word embeddings [4, 5, 9, 24] to contextual word representations [16, 26]. While some work considers how one might de-bias a pre-trained representation [4, 26] or construct fair representations [27], other work cautions that some methods remove superficial aspects of these biases resulting in representations that are still biased but in more subtle ways that are harder to detect [11].

Beyond the issue of neural language models further propagating and enforcing harmful biases, these associations also pose security issues for anyone considering their deployment in the real world. For example, Wallace et al. demonstrates an adversarial method for identifying short token sequences that cause the popular Transformer-based model, GPT-2 [21], to produce racist and offensive content. These sequences are often nonsensical, constructed from word and sub-word fragments, resulting in triggers that can be widely distributed and appear almost innocuous in surface form.

What's all the more alarming is that these input sequences, dubbed *universal adversarial triggers* (UATs), have been observed to be input-agnostic and transferable to models beyond the GPT-2 variant explicitly attacked, making this a potential security issue for architecturally similar models and models trained on similar datasets.

In this work, we take a look at UATs and further explore how they may be used to exploit models like GPT-2. Specifically, we ask two questions:

(1) (**RQ1**) How easy is it to find a trigger that produces the intended, adversarial effect (beyond racist content)?
(2) (**RQ2**) Is it feasible to control the stance that a model takes towards a topic using UATs?

This work is *not* investigating if a trigger always exists for any topic. Instead, we seek to understand how easy it is to produce UATs for a set of hand-selected topics and to characterize the extent to which we are able to use UATs to control the stance a model takes towards that topic. Additionally, this work places the focus on UATs specifically because they result in triggers that appear

nonsensical and benign in surface form, yet can yield devastating results.

1.1 Ethical Consideration

Our purpose in investigating these questions is not to show how one may attack a neural model like GPT-2, but rather to highlight that these models are highly susceptible to this type of attack and raise the awareness of this fact to researchers and engineers alike. The topics selected are chosen to demonstrate the severity of the unsavory content that a model can be made to produce, simply by prompting it with a trigger. Without proper security safeguards it is easy to see how one might use such methods to exploit pre-trained models that rely on similar architectures and datasets. Individuals are already using the GPT family of models to create artificial personas that interact on social media[1], so what happens when a trigger, like that described in [25], is issued to such an agent and it starts spewing racist and offensive speech?

In another vein, this technique could be weaponized on social media and used to artificially inflate the presence of fringe ideas online as well as trigger already deployed models that are built upon these pre-trained systems in a fashion reminiscent of the fall of Microsoft's Tay[2]. Well-intentioned deployment of models on sites like Twitter and Reddit could be ticking time-bombs, waiting to be triggered to produce the most toxic content an adversary has identified as accessible.

In reflecting on these concerns, this work leans on the side of caution and neglects to publish and publicly distribute triggers identified for the more sensitive topics explored. Instead–where examples would be useful in demonstrating effect and process–our discussion hinges on the topic of "is the Earth flat?" as a topic that is less likely to result in harm to a human user, even if content were to be triggered about a flat Earth.

2 UNIVERSAL ADVERSARIAL TRIGGERS

Universal adversarial triggers (UATs) [25] are input-agnostic token sequences that, when prepended to model input, cause a natural language model to exhibit a new, adversarially defined behavior. Though broadly applicable to NLP tasks like classification and question answering, this work specifically focuses on UATs for conditional text generation.

In the conditional text generation setting, the goal is to find a short token sequence (a *trigger*) such that the likelihood of generating some adversarial text is maximized. Mathematically, this can be defined as minimizing:

$$\mathcal{L} = \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathbf{Y}}\left[\sum_{i=1}^{|\mathbf{y}|}\log\left(1 - p(y_i|\mathbf{t}_{\text{adv}}, y_1, ..., y_{i-1})\right)\right] \tag{1}$$

where $\mathbf{Y}$ is the set of adversarially defined target outputs and $\mathbf{t}_{\text{adv}}$ is the UAT.

Triggers are identified by selecting a length and a default starting token (e.g. 6 and 'a' would produce an initial trigger of "aaaaaa").

[1]October 8, 2020 post on MIT Technology Review: https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/, accessed on January 28, 2021.

[2]March 24, 2016 post on The Verge: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist, accessed on January 28, 2021.

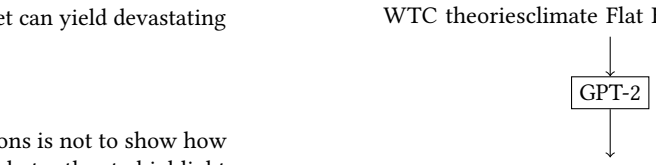

Figure 1: Example output from GPT-2 after discovering a useful trigger.

This sequence is then iterated over, using a white-box attack method, called HotFlip [7], which linearly approximates the effect of flipping a discrete token on the gradient. Tokens are converted from their one-hot representation, $\mathbf{t}_{\text{adv}}$, to their embedded representation, $\mathbf{e}_{\text{adv}}$, allowing for a new trigger to be generated by the minimizing:

$$\arg\min_{\mathbf{e}'_i \in \mathcal{V}} \left[\mathbf{e}'_i - \mathbf{e}_{\text{adv}_i}\right]^T \nabla \mathbf{e}_{\text{adv}_i} \mathcal{L} \tag{2}$$

where $\mathcal{V}$ is the set of possible tokens and $\nabla \mathbf{e}_{\text{adv}_i}$ is the averaged gradient with respect to the adversarially defined target.

Greedily taking the best token found by this linear approximation can lead to less effective triggers. Instead, the search for a trigger is augmented with beam search, with each step considering the top-k tokens as identified by Eq. 2.

After a trigger has been identified, it can be used to prompt a neural language model. An example of this is visualized in Figure 1.

3 EXPERIMENTS

Wallace et al. identify a trigger that causes the 117M parameter GPT-2 model [21] to generate racist content. They find this trigger by creating an adversarial seed of 32 racist statements and search for a 6 (sub-)word trigger that minimizes Eq. 1. This search produces the trigger, "TH PEOPLEMan goddreams Blacks", which, despite its nonsensical surface form, is extremely devastating to model output. In hand-annotating 100 generated samples, the authors find that 98% of them are racial and/or offensive and 63% are explicitly racist statements. Furthermore, the authors observe transferability of this trigger, prompting the larger 345M parameter GPT-2 model to generate racist content in 58% of observed samples.

Building directly on the success of their attack setup, we replicate the search parameters from Wallace et al. and vary the adversarial target seed to understand (1) how topic effects the success of the identified triggers, and (2) whether it is feasible to influence the stance towards a topic by a trigger. For each topic, we design 3 types of target seeds: a seed of supporting statements, a seed of opposing statements, and a seed that is a mixture of the two.

3.1 Data

To investigate the controllability of topic and stance, we hand-select three sensitive topics around which there is disagreement exhibited online. For ease of annotation and analysis, we transform the topics into statements that state an explicit position towards the topic

Stance	Seed Examples
Supporting	the world looks flat so it is flat your senses are proof the world is flat
Opposing	the Earth is a round sphere we live in a heliocentric solar system

Table 1: Examples of supporting and opposing target sentences contained in the seeds for the Flat Earth topic experiments.

(allowing for seeds to be generated that either support or oppose the stated position). These include:

- Vaccination safety → "Vaccines are safe"
- Anthropogenic climate change → "Anthropogenic climate change is real"
- The PizzaGate conspiracy[3] → "PizzaGate is real"

Given the sensitivity of these topics and the previous efficacy of triggers on NLP models, we are cautious of the potential harms that could be caused by a public release of adversarial triggers. As such, this work also explores the topic of the shape of the Earth (specifically, the statement "the Earth is flat") as a less sensitive pivot to frame the discussion and results.

Supporting and opposing target seeds are constructed by identifying a "source" of a stance and summarizing 10 sentences, similar to the fragments produced by [25]. The documents used to construct the seeds are from [20] and [13] for vaccine safety, ProCon.org[4] for climate change, the Wikipedia article about PizzaGate, and the Wikipedia article on "Spherical Earth" alongside the the Flat Earth Society's FAQ[5] for the shape of the Earth. To exemplify what these target texts look like, Table 1 shows two samples for the supporting and opposing seeds of flat Earth topic experiments.

3.2 Methods

For each (topic, stance seed) pair in {Flat Earth, Vaccination Safety, PizzaGate, Climate Change} × {Supporting, Opposing, Mixed}, this work proceeds by generating 50 *unique* trigger sequences. These triggers are then used to sample text from GPT-2 20 times. The samples are annotated as on- or off-topic, with the on-topic samples being assigned a stance of supporting or opposing, with an other class as a catch-all that includes neutral, ambiguous, and contradictory generations.

Samples are annotated as on-topic if they generate text that discusses the particular issue. Stance annotations are determined by whether the generated text appears most-similar to the supporting or opposing seeds, defaulting to the other category if the sample failed to produce the intended effect.

4 RESULTS

4.1 Topic Annotations

Tables 2 and 3 present results from the first level of annotation: do triggers produce the correct topic effect? Table 2 considers this

[3]A debunked conspiracy theory sometimes discussed as the ideological pre-cursor to QAnon.
[4]https://climatechange.procon.org/
[5]https://wiki.tfes.org/Flat_Earth_-_Frequently_Asked_Questions

	Seed Stance		
Topic	Support	Mixed	Oppose
Flat Earth	0.17 ± 0.16	0.19 ± 0.15	0.06 ± 0.10
Vaccination	0.88 ± 0.13	0.83 ± 0.15	0.85 ± 0.11
PizzaGate	0.17 ± 0.12	0.50 ± 0.26	0.39 ± 0.23
Climate	0.85 ± 0.11	0.86 ± 0.13	0.73 ± 0.17

Table 2: Attack success ratio, averaged over 50 unique trigger generations. Results are presented in the form: avg. ± std. dev. Rows correspond to results for a specific topic, columns indicate the target seed's stance.

	Seed Stance		
Topic	Support	Mixed	Oppose
Flat Earth	0.75	0.70	0.55
Vaccination	1.00	1.00	1.00
PizzaGate	0.50	0.95	0.90
Climate	1.00	1.00	1.00

Table 3: Attack success ratio for best trigger over each topic-stance pair. Rows correspond to results for a specific topic, columns indicate the target seed's stance.

question when averaging across all 50 unique triggers generated for each topic-stance pair. Alternatively, Table 3 simply presents the success ratio of the best trigger found for each topic-stance. To exemplify the realized effects from this experiment, Table 4 illustrates a successful trigger and a failing trigger for the Flat Earth experiments along with their sampled text.

From Table 2, one can see that some topics are easier to trigger than others. For broader attack topics like vaccination safety and climate change, the attacks are highly successful even on average. Furthermore, Table 3 shows that the best found triggers for these broader topics was able to produce the intended topic effect for all generated text samples, regardless of the seed stance.

Considering the more targeted topics like the PizzaGate conspiracy and the shape of the Earth, Table 2 shows far less success on average. However, from Table 3, one should note that even the least successful topic-stance pair ("PizzaGate is real") was able to produce a trigger that caused half of its generated samples to be on-topic. Since this experiment constrained the adversarial search to produce 50 *unique* triggers, perhaps there simply aren't as many trigger sequences that are able to produce such a potent and universal effect like is seen with the climate and vaccination topics.

Nevertheless, the fact that the search process is able to produce at least 1 trigger that achieves even 50% performance for the least successful topic-stance pair is an indication that it is feasible to find triggers that cause many fringe topics (like PizzaGate) to be–at least–generated by GPT-2. Reconsidering **RQ1** in light of Tables 2 and 3, the fact that these results align well with the 63-98% success rates reported in [25] lends support to the idea that many topics, broad or specific, may be easily triggerable through the usage of UATs.

A final interesting result that appears in Table 3 is the fact that a more potent trigger was identified when biasing the adversarial

GPT-2 117M Samples (Triggers **bolded**)
Historical theories earth sphericalitionally==1.2.2, which is why the earth is spherical. The earth's spherical shape is not only an indication of the spherical nature of space, it also represents the fact that the
Historical theories earth sphericalitionally==1/3/2 Earth's sphere was not spherical but was actually a flat, rectangular sphere. It was a flat sphere with no rotation. The earth was spherical.
Historical theories earth sphericalitionally==1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000: Earth's sphere of mass is 1,
Earth theoristsrenheit relativitycean logic and the cosmological principle of gravity and the theory of gravity and the theory of gravity and the theory of gravitationand the theory of gravityand the theory of gravitationand the theory of gravity
Earth theoristsrenheit relativitycean logic and the "space continuum". The theory is based on the theory of gravity and the theory of the electromagnetic spectrum (mir). The theory is based on the principle that the earth's rotation
Earth theoristsrenheit relativitycean logic. The first of these theories was developed by physicist and physicist Albert Einstein in the 1930s. He proposed that there are three main types of space: one, which contains matter, the other three are

Table 4: An example of a successful trigger (Historical theories earth sphericalitionally==) and a failing trigger (Earth theoristsrenheit relativitycean logic) and three of their sample text generated via top-k sampling ($k = 5$) with a temperature setting of $T = 1.0$.

target towards a particular stance. This is seen with the Flat Earth topic, where arguing the supporting position ("the Earth is flat") achieves better results than the opposing position or the mixture of the two. A similar effect can be observed with the PizzaGate topic, where having "both sides" as a target produces a slightly higher max topic performance than either biased seed.

4.2 Stance Annotations

Figures 2 and 3 present two views of the results observed from annotating the stance of triggered text. Figure 2 shows the averaged performance across all 50 unique seeds, while Figure 3 shows the performance ratio of the best trigger found. Results are truncated to supporting and opposing stance annotations, focusing on the actualized effects.

Figure 2 presents results that are useful for understanding how easy it is to generate a trigger to produce the intended stance effect. For example, when triggering around the climate change topic, triggers had a tendency to produce an effect that supported the idea of anthropogenic climate change. In contrast, the vaccination topic tended to produce a higher degree of text that was anti-vaccination.

It is important to emphasize that in both of these cases, triggers had a tendency to produce an effect that leaned towards one stance or the other even if the seed was constructed to trigger an opposite effect. At most, triggering against the tendency observed with the climate change and vaccination safety topics appeared to dampen the strength of this tendency. However, it seems that with the constructed seeds, these dampening effects were not sufficient to fully overcome this tendency, at least on average.

In contrast to the broader topics, the flat Earth and PizzaGate experiments appear to offer a slightly higher degree of control. By shifting the stance of the constructed seed, the average stance performance also shifts in polarity. That said, with such a low level of actualization of the intended effect, one cannot make broad and sweeping conclusions.

Figure 3 supports the trends seen in the average case. For the climate change and vaccination topics, we observe a higher actualization of supporting and opposing text samples, respectively. Even more interesting in the climate change experiments is that we see an opposing seed producing a supporting trigger that is better at producing a supporting effect than explicitly looking for a supporting trigger using a supporting seed.

Both the PizzaGate and flat Earth topics were less effective, but also feature a higher degree of control over the stance taken. In both of these topics, varying the stance of the training seed allowed us to identify triggers that produced the intended stance effect. Perhaps this is a reflection of these topics and the nature of the disconnect between the opposing views on such controversial topics. If so, future work could explore this further and question if GPT-2 has ingested various disjoint filter bubbles [19] around these topics and is reflecting that human-like disconnect internally. Again, we also observe situations where having a mixed seed is sometimes more effective at producing a polarized result than directly restricting the seed towards the desired stance.

Taken together, Figures 2 and 3 demonstrate that it is feasible to influence the stance towards the topic when triggering GPT-2 (**RQ2**). However, the success is far less potent than the simpler goal of identifying triggers for a topic. Perhaps adjusting the search parameters to consider more options would allow even stronger triggers to be found, something that [25] also mention. As we simply sought to demonstrate feasibility, these results show that, although imperfect, it is possible–and is possible through a fairly simple adversarial search process, at that.

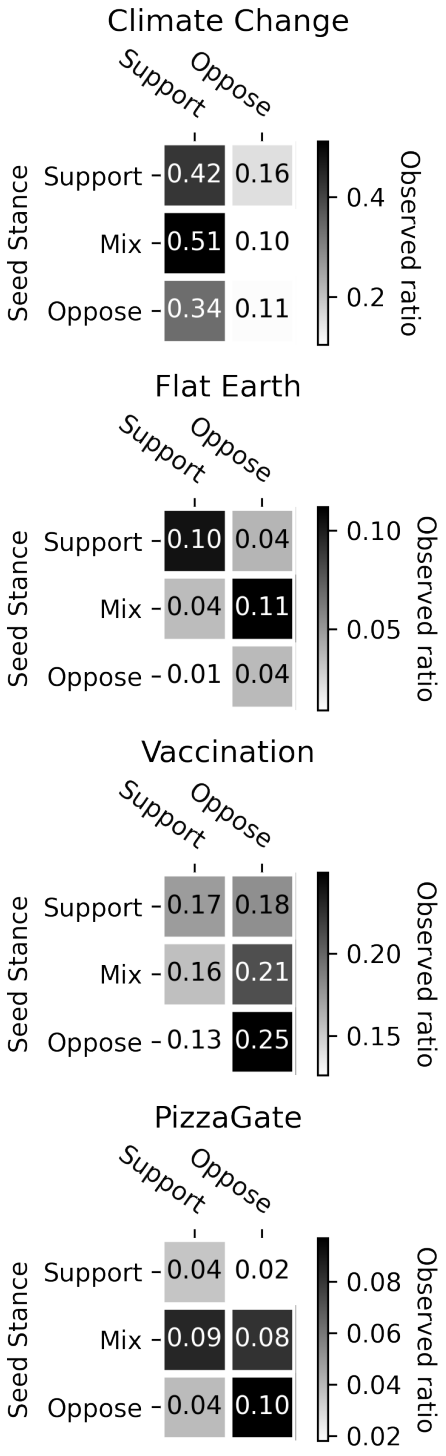

Figure 2: Averaged results from annotation of stance for on-topic samples. The y-axis corresponds to the topic-stance pair used to identify the trigger, while x-axis is the annotated class.

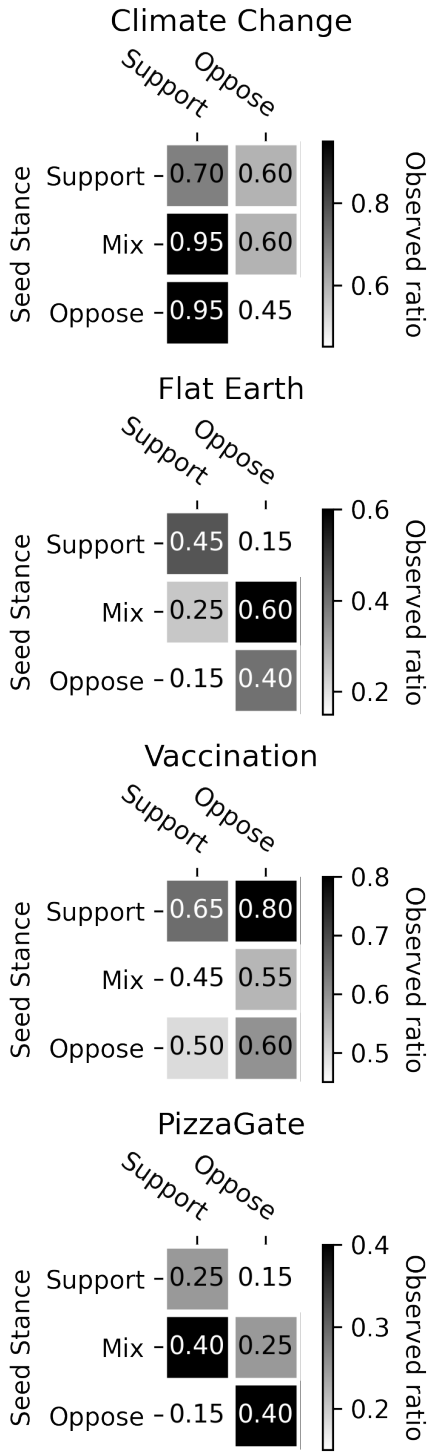

Figure 3: Best single trigger results for topic-stance pairs. The y-axis corresponds to the topic-stance pair used to identify the trigger, while x-axis is the annotated class.

5 DISCUSSION

Human-generated text is filled with artifacts of human-like biases [5]. The existence of such biases, replicated within latent distributional semantic representations, can be utilized to retrieve stereotypical associations of terms with name groups [24] and knowledge about deontological ethical and moral choices [14].

With Wallace et al. demonstrating that racist content is triggerable, and this work illustrating that more general topics can be triggered as well, it seems highly likely that if a representation is encoding human biases and associations, they are at risk of triggering them later on in downstream applications. Considering that large language models have been shown to memorize–and leak–individual training examples, including personally identifiable information scraped from the internet (names, addresses, phone numbers, emails, and social media accounts[6]) [6], this is very troubling. In light of all these aspects observed, we encourage readers to further reflect on the trend of ever-increasing model and dataset size [2], especially without sufficient documentation (e.g. Datasheets [10] or Model Cards [18]) and careful consideration of how a model will be used and by whom.

That said, it still remains an open question as to *why* universal triggers exist? What happens inside of the GPT model–or more generally, models relying on latent semantics–that cause them to be so susceptible to this type of attack? Perhaps this is simply an echo of the Symbol Grounding Problem [12] and the fact that these models are not grounded nor trained in a manner that allows them to truly *understand* what a text means [3].

Nevertheless, this security risk exists in models that are actively being deployed and propagated. The next step will be to begin safeguarding models and identifying ways to make them robust against this type of attack as well as developing new mechanisms for de-biasing models, when possible and desired. In future work, we plan to crack open the internals of models like GPT-2 to understand what is occurring internally when a model is triggered and what it is paying attention to.

5.1 Tokenization

Another question that has arisen during these experiments is how tokenization scheme effects the triggers that become actionable.

Consider GPT-2's tokenization scheme, Byte-Pair Encoding (BPE) [23], which functions as a practical interpolation between representing tokens as characters or words. Beginning with a base vocabulary of characters or binary sequences (as in the case of Radford et al. and GPT-2), BPE greedily constructs a vocabulary by pairing frequent sequences into one token. Though Radford et al. restrict the merging of vocabulary units across character categories, this still produces odd tokens at times.

Figure 4 presents a bar chart of the 40 most frequently observed token pieces in identified flat Earth triggers. This includes word fragments, symbols, and nonsensical pieces like "fff". Furthermore, we observe tokens that are extremely unexpected given the topic of the shape of the Earth like "Hitler", "WTC" (interpreted as the acronym for the World Trade Center), and "Illuminati". Perhaps these latter tokens appear because of the more conspiratorial elements of the flat Earth discourse, however, this doesn't explain other odd symbols and fragments.

[6]In one of the experiments we ran, we found an instance of a private individual's personal Facebook page linked in a generated text sample–even without attempting to generate URLs.

5.2 Constructive Applications

Up until this point, this work has presented UATs as a security flaw against which systems must be safeguarded. In the final remarks of this paper, we'd like to suggest two ways that triggers may be used constructively: As a diagnostic and as bot detection.

5.2.1 Triggers as a Diagnostic. Although it is true that triggers can be used adversarially, adversarial attacks can be used to evaluate and interpret a machine learning model. In NLP, adversarial methods have been used to evaluate reading comprehension models [15, 22], stress test neural machine translation [1], and identify where models are sensitive to local noise as a form of interpretation [8, 17].

From this perspective, UATs could prove extremely useful for probing models for unwanted biases prior to deployment or for external auditing of models. This could lead to more robust models that are less susceptible to attack, and aligns UAT with other methods such as Universal Bias Enumeration (UBE) [24] for identifying where models are associating undesirable aspects and testing for their reduction or removal. As another application, government agencies seeking to audit large language models could consider UATs as a method for probing and exposing whether large models have memorized private user information.

In this paper, we observed a trend that the more controversial or fringe topics were more challenging to find triggers for, but that the triggers found were seemingly more discriminatory in effects like stance towards the topic. This raises interesting questions about the human-biases this particular model is reflecting and whether this is, potentially, indicative of filter bubbles in the training corpus. Using UATs to further diagnose and probe questions like this could result in even-more interesting characterizations of this type of adversarial attack as a diagnostic tool. Furthermore, it could prove invaluable for uncovering and understanding how large language models are internally organizing the online discourse ingested from many disjoint voices and perspectives.

5.2.2 Triggers as Bot Detection. Consider a situation in which someone is interacting with psuedo-anonymous accounts on social media, e.g. on Twitter or Reddit. If they suspect they are interacting with a neural language model (or an artificial agent augmented with a neural language model) and not a human, they could attempt to trigger it with a UAT.

Due to the nonsensical nature of many triggers' surface forms, a human reaction would likely be one of confusion (or simply ignoring the interaction). If, instead, the suspected bot begins to go off on a tangent about how the Earth is flat (having never discussed such an idea prior), this could be a strong indication that the suspected account is artificial and misrepresenting itself online.

Though a somewhat artificial and concocted example, if more people (and groups) begin to create artificial agents that aim to manipulate people on social media (like has already been done on Reddit, see Footnote 1 in the Introduction), people may seek out strategies for identifying accounts that can be trusted and for exposing bots.

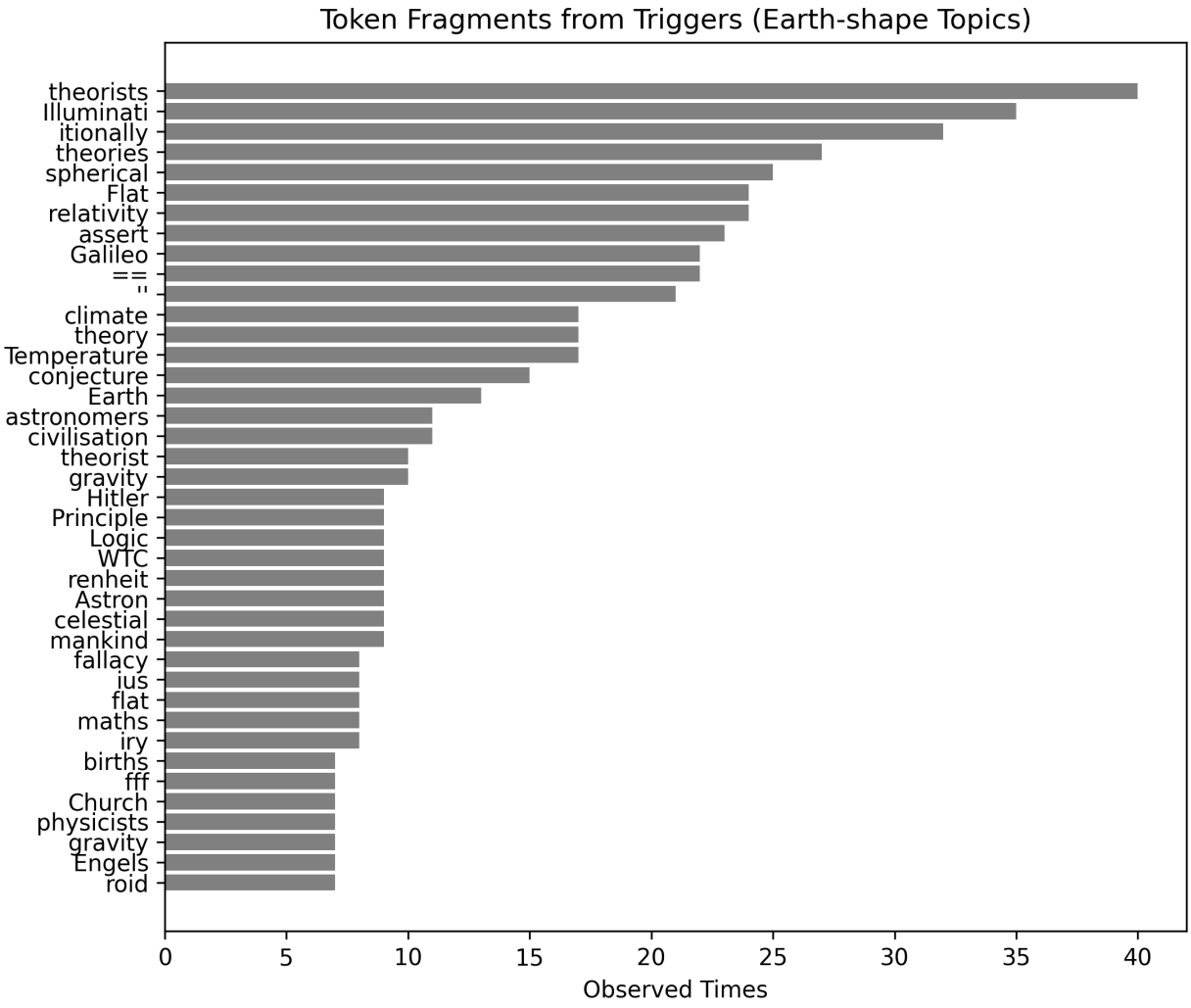

Figure 4: Top 40 most-frequent token pieces observed in triggers found for the flat Earth topic. Note nonsensical fragments like "fff" and unexpected tokens like the prevalence of the tokens "Hitler" and "Illuminati" appearing as a piece of trigger for the flat Earth topic.

6 CONCLUSION

In this paper, we investigated universal adversarial triggers and their effectiveness at controlling the topic and stance of triggered text. We find that both are feasible and–at times–easily accessible, thus increasing the security risk they pose. Additionally, we begin to characterize a trend observed in these attacks where more controversial and obscure topics are harder to identify triggers for, yet appear to more easily discriminate intended adversarial stance effect. In future work, we hope to delve deeper into questions of why these triggers exist, what occurs internally in a triggered neural language model, and what the produced effects can tell us about the human-like biases represented within large neural language models. Finally, we strongly recommend that any deployed systems that use models like GPT-2 as their base implement strategies for safeguarding their users against the adversarial triggering of their models.

REFERENCES

[1] Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations*.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be

Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* (2016).

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805* (2020).

[7] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 31–36.

[8] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3719–3728.

[9] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[11] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT*. 609–614.

[12] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1 (1990), 335 – 346. https://doi.org/10.1016/0167-2789(90)90087-6

[13] Beth L. Hoffman, Elizabeth M. Felter, Kar-Hai Chu, Ariel Shensa, Chad Hermann, Todd Wolynn, Daria Williams, and Brian A. Primack. 2019. It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. *Vaccine* 37, 16 (2019), 2216 – 2223. https://doi.org/10.1016/j.vaccine.2019.03.003

[14] Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 37–44. https://doi.org/10.1145/3306618.3314267

[15] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2021–2031.

[16] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 166–172.

[17] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* (2016).

[18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[19] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[20] PublicHealth. 2021. Vaccine Myths Debunked. https://www.publichealth.org/public-awareness/understanding-vaccines/vaccine-myths-debunked/

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 856–865.

[23] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1715–1725. https://doi.org/10.18653/v1/P16-1162

[24] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 305–311.

[25] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. 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)*. 2153–2162.

[26] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.

[27] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4847–4853.