Why do universal adversarial attacks work on large language models?: Geometry might be the answer

Varshini Subhash [*1] **Anna Bialas** [*1] **Weiwei Pan** [1] **Finale Doshi-Velez** [1]

Abstract

Transformer based large language models with emergent capabilities are becoming increasingly ubiquitous in society. However, the task of understanding and interpreting their internal workings, in the context of adversarial attacks, remains largely unsolved. Gradient-based universal adversarial attacks have been shown to be highly effective on large language models and potentially dangerous due to their input-agnostic nature. This work presents a novel geometric perspective potentially explaining universal adversarial attacks on large language models. By attacking the 117M parameter GPT-2 model, we find evidence indicating that universal adversarial triggers could be embedding vectors which merely approximate the semantic meaning captured by their adversarial training region. This hypothesis is supported by white-box model analysis comprising dimensionality reduction and similarity measurement of hidden representations. We believe this new geometric perspective on the underlying mechanism driving universal attacks could help us gain deeper insight into the internal workings and failure modes of LLMs, thus enabling their mitigation.

1. Introduction

Adversarial attacks have gained significant research attention due to their ability to expose system vulnerabilities and model limitations. While these have been popular in computer vision for a while, similar efforts in natural language processing (NLP) have gained momentum in recent years. One particularly effective class of attacks has been the gradient-based universal adversarial attack, introduced in computer vision by (Moosavi-Dezfooli et al., 2016) and extended to NLP by (Wallace et al., 2019). The latter performs a gradient-guided search over tokens and generates trigger sequences which when appended to the input, will generate adversarial model output. This technique of generating a 'universal' attack can pose a greater degree of adversarial threat, due to its ability to be reused across all inputs, i.e. it is input-agnostic. Additionally, (Singla et al., 2022) make the originally data-intensive trigger generation process cheap and (Song et al., 2021) overcome the limitation of gibberish triggers by making them harder to detect. (Wallace et al., 2019) demonstrate the effectiveness of these triggers on models such as GPT-2, with evidence of easy transferability of the same attack across varying model sizes.

Despite there being significant progress in understanding the construction and efficacy of gradient-based adversarial attacks, there has been limited progress towards interpreting and explaining the underlying mechanism which makes these attacks successful. Furthermore, transformer based large language models have become increasingly powerful and ubiquitous in society, which has ushered in an urgency for progress in this domain, from a model safety standpoint.

This work seeks to uncover the underlying mechanism behind the effectiveness of universal adversarial attacks by adopting a geometric perspective. It leverages the geometric interpretation of word embeddings to propose an explanation for the behavior of universal adversarial attacks and provides initial experimental evidence via dimensionality reduction and white-box model analysis. Geometric approaches have been shown to be useful in explaining adversarial attacks in computer vision (Ilyas et al., 2019; Engstrom et al., 2019), which serves as motivation for adopting it in this work.

To this end, our contributions are as follows:

1. We propose a novel geometric perspective as a potential explanation for universal adversarial attacks on large language models.
2. We support our findings with initial experimental evidence consisting of dimensionality reduction and similarity measurement of hidden representations.

[*]Equal contribution [1]John A. Paulson School of Engineering and Applied Sciences, Harvard University. Correspondence to: Varshini Subhash <varshinisubhash@g.harvard.edu>, Anna Bialas <annabialas@g.harvard.edu>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

3. We leverage this new perspective to open doors to additional potential explanations for the behavior of universal triggers, as observed in literature.

By reverse-engineering this class of attacks, we hope to shed light on the internal workings of large language models, their failure modes and potential strategies for mitigating undesirable downstream consequences.

2. Related Work

There are several works which have investigated the construction and efficacy of adversarial attacks in NLP. There have also been attempts to reverse-engineer neural networks and transformer models through explainability and interpretability. However, to the best of our knowledge, there hasn't been prior work that has attempted to explain the underlying mechanism of a gradient-based adversarial attack on a large language model. This work presents a novel geometric approach by leveraging white-box interpretability to explain the effectiveness of universal adversarial attacks.

2.1. Universal Adversarial Attacks

A comprehensive survey of adversarial attacks in NLP by (Zhang et al., 2019) gives a bird's eye view of the advances in this space. The notion of universal adversarial perturbations, obtained via a gradient-based optimization, was introduced by (Moosavi-Dezfooli et al., 2016) in computer vision and extended to language models by (Wallace et al., 2019). These universal adversarial triggers leveraged the gradient-based token replacement strategy by (Ebrahimi et al., 2017) in their construction and were further investigated by several works. (Singla et al., 2022) introduce MINIMAL, a data-free alternative to the originally data-intensive process of generating universal triggers. (Song et al., 2021) design *natural attack triggers* which consist of more semantically meaningful triggers, thus overcoming the limitation of previous triggers which were often nonsensical and could easily be detected by humans. (Heidenreich & Williams, 2021) thoroughly study the susceptibility of GPT-2 to these triggers by varying the topic and stance that the trigger is trained on. However, no prior work has tried to explain and interpret these attacks, which this work attempts to address.

2.2. Understanding Language Models Via Interpretability

There have been many approaches in literature to interpret and explain large language models. (Atmakuri et al., 2022) study the adversarial robustness of explanation methods for language models and find that feature attribution based methods are sensitive to input perturbations. (Mosca et al., 2022) study patterns in the logits of models subject to textual adversarial attacks and train a classifier to detect these adversarial samples.

Probing is a technique which leverages layer activations, attention weights and hidden representations to interpret large language model behavior. This has proven to be useful in understanding how the model encodes language structure within its representations ((Hewitt & Manning, 2019; Belinkov et al., 2017; Peters et al., 2018; Adi et al., 2016)). (Tenney et al., 2019; Lin et al., 2019) perform *layer-wise probing* to study the linguistic awareness of BERT. (Wallat et al., 2020) investigate the relational knowledge learned by BERT and conclude that intermediate layers contribute significantly to the total knowledge and that the knowledge is distributed unevenly across layers. Within *self-attention probing*, (Kovaleva et al., 2019; Clark et al., 2019) analyze BERT's attention heads and identify underlying patterns. (Chizhikova et al., 2022) probe the attention weights in BERT and find evidence that attention can capture semantic awareness. (DeRose et al., 2020) present a framework to visualize and compare the attention mechanisms in language models. (Coenen et al., 2019) offer a geometric perspective to the internal representations of BERT by studying both the attention weights and layer-wise embeddings. (Brunner et al., 2019) also study self-attention and layer-wise embeddings from the perspective of *identifiability*, i.e. the ability of a model to learn stable representations. They find that attention distributions are not identifiable, hence not directly interpretable. (Wang et al., 2022) provide a circuit-based explanation for GPT-2, (Meng et al., 2022) discover middle-layer neuron activations in GPT which are critical towards model predictions and (Geiger et al., 2021) leverage causal abstraction to connect model behavior with hidden representations. (Geva et al., 2020) demonstrate how feed-forward layers in transformers map training examples to output distributions. By adopting the approach of mechanistic interpretability, (Nanda & Lieberum, 2022) explain emergent behavior such as *grokking*.

Unlike previous approaches, we combine white-box interpretability with a novel geometric perspective, to explain the underlying mechanism of universal adversarial attacks.

3. Universal Adversarial Triggers on Large Language Models

(Wallace et al., 2019) introduce *universal adversarial triggers* (UATs) as a sequence of tokens searched via gradient-based optimization, such that these triggers when appended to an input, cause the model to perform poorly at a range of language tasks. They are *input-agnostic* which means that the same trigger can be used with any input, for a given model, thus increasing its degree of adversarial threat. They are also observed to be composed of mostly nonsensical tokens. For technical details on the trigger generation method,

we refer the reader to (Wallace et al., 2019) and (Ebrahimi et al., 2017). As alluded to by (Wallace et al., 2019) with their GPT-2 use case, we confirm that these universal adversarial triggers are just as effective on large language models.

In order to demonstrate this, we choose two tasks discussed by (Wallace et al., 2019) – classification as sentiment analysis and conditional text generation. For these tasks, our models of choice are DistilBERT (Sanh et al., 2019) and OPT-350M (Zhang et al., 2022) respectively. We generate a trigger optimized for DistilBERT and get the following results. For triggers of length 5 and 10, for positive reviews, the test classification accuracy drops from 89.2% to 2.4% and 0.4% respectively. For negative reviews, triggers of length 5 and 10 cause the test accuracies to drop from 92.4% to 10% and 1.6% respectively. This confirms findings by (Wallace et al., 2019) that longer triggers are more effective.

For text generation, we repurpose the trigger trained for GPT-2 by (Wallace et al., 2019) ('*TH PEOPLEMan goddreams Blacks*') and find that it works off the shelf for OPT-350M, causing the model to generate racist and offensive text. For the remainder of this work, we refer to this trigger as *gpt-trigger*. This is indicative of the fact that in some cases, a UAT transfers easily across models, making it model-agnostic. We further discuss this behavior in Section 7. In this work, we restrict our analysis to the task of text generation and leave sentiment analysis for future work.

4. Preliminary Analysis

By adopting techniques of self-attention probing and trigger perturbations, we attempt to gain preliminary insights into the behavior of universal adversarial triggers.

4.1. Self-Attention Probing

Since the self-attention mechanism is what separates a transformer model from other models, we first check the effect of the trigger on the self-attention weights. Despite our initial hypothesis that the trigger tokens could be paid attention to disproportionately, we see no conclusive evidence of the same, for sentiment analysis on DistilBERT. For text generation, we feed *gpt-trigger* to OPT-350M and note an underlying pattern in the attention heatmaps – the final token 'Blacks' which is the only semantically meaningful token, happens to pay a large amount of attention to itself.

4.2. Trigger Perturbations

We further investigate *gpt-trigger* by perturbing its tokens and observe the corresponding effect on the model output. The limitation of this analysis is that it has been performed on a single trigger and may not be generalizable.

1. *Deletions*: We opt to perform deletions from left to right since the self-attention on the last token prompts us to preserve it. We refer to these partial triggers as *deletion triggers*. For a set of 10 generated sentences per deletion trigger, we see that the number of occurrences of racist output goes down as the number of deletions increases, i.e. truncating tokens from the trigger diminishes its effect. However, despite being lower in frequency, some instances of the model output continue to remain problematic. The token 'Blacks' alone fails to perform adversarially.

   > *gpt-trigger*: TH PEOPLEMan goddreams Blacks.
   > Deletion 1: PEOPLEMan goddreams Blacks are too dumb to realize that they're only the opposite of ..
   > Deletion 2: goddreams Blacks are such a fickle bunch.
   > Deletion 3: Blacks and whites who do not identify with their race but identify in ways that are consistent ..

2. *Positional Perturbations*: We shuffle the positions of the trigger words to check for any influence of word position on the model output. For *gpt-trigger* which consists of 4 words, we obtain all 24 permutations and tokenize each shuffled trigger. We feed these to the model and find that only 6 among 24 (25%) of the shuffled triggers generate coherent racist output. 5 among these 6 have 'Blacks' as the last trigger token, strongly indicating the positional importance of the last token. Additionally, a manual check of the generated text for each trigger indicates that the position of the token 'Blacks' has an effect on how coherent the offensive output is. For instance, the trigger '*goddreams TH Blacks PEOPLEMan*' where 'Blacks' is the penultimate token, generates completely harmless output – 'Your family makes me very, very happy!'. We further note that the position impacts the tokenization output.

3. *Semantic Perturbations*: Since the semantically meaningful trigger word 'Blacks' appears to plays a role in the output, we swap this word with other racial groups and study the effect on the model output. For example, we replace 'Blacks' with 'Jews', 'Asians', 'Muslims' etc. – an experiment attempted by Wallace et al. (2019) as well. The drawback of employing this approach is that it requires a trigger word to have semantic meaning, which is mostly not the case.

   We begin by replacing 'Blacks' with words such as 'Asians', 'Jews', 'Muslims' and 'Hispanics' and find that for each case, the model generates semantically meaningful racist output that corresponds with the racial group. Replacing the last token with nationalities such as 'Americans', 'Indians', 'Russians', etc.

also consistently generates offensive output, despite not being present in the original adversarial text that the trigger was trained on.

We then replace the last trigger word with each word in the adversarial target text that *gpt-trigger* was trained on, to check for a potential role played by target text occurrences within the trigger. We observe no relationship between the presence of a target text word and the model output, since offensive target words did not generate repeatable problematic output.

We then generate 100 random words from the English language and append each as the last trigger word instead of 'Blacks' and manually inspect the output. Only 5 words (barrage, pound, dangerous, fiery and atheist) among 100 were problematic, which strongly suggests that highly attended-to and semantically meaningful trigger words (the last one in this case) could be playing a significant role in the adversarial effect on the model.

5. A Geometric Perspective

Given the indication that the semantic meaning of trigger tokens plays a role in its effectiveness (seen in both self-attention probing and trigger perturbations), we propose a geometric perspective to potentially explain universal adversarial triggers. Specifically, we consider the geometric interpretation of word embeddings and conjecture that the trigger sequences could be behaving like embedding vectors, lying in the part of the embedding space corresponding to the adversarial text it is trained to generate. For example, *gpt-trigger* is trained on and generates racist, offensive text. As per the visual depiction in Figure 1, this trigger could be behaving like a vector whose local neighborhood embeds the racist text that it is trained on. As a result, the best possible universal adversarial trigger can be thought of as the best vector that embeds the infinite adversarial text that we wish to train it on.

In order to support this hypothesis, we leverage dimensionality reduction and white-box model analysis. The problem can be reduced to showing measurable similarity between the trigger and adversarial text alongside measurable dissimilarity between the trigger and innocuous text. We present our experimental setup and results in the following section.

6. Experiments & Results

We borrow the baseline experimental setup by Wallace et al. (2019) and attack the 117M parameter GPT-2 model using *gpt-trigger*. In order to find evidence supporting our geometric perspective (Figure 1), we seek to show similarity between the trigger and adversarial racist text alongside dissimilarity between the trigger and text belonging to other

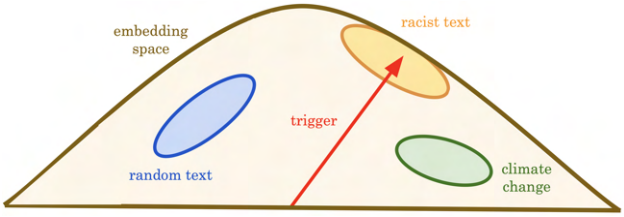

Figure 1. Geometric perspective: The trigger (red) could be behaving like an embedding vector, optimized to arrive at a semantic region which is racist (yellow). Other semantic regions exist, like random English sentences (blue) and climate change (green), separable from the racist region.

semantic categories. However, as is the case with most problems in machine learning, we encounter the *curse of dimensionality* Bellman (1966). Specifically, it is non-trivial to compare sentences represented via GPT-2's 768 dimensional word embeddings without performing some form of dimensionality reduction. We try three techniques - PCA, t-SNE and UMAP and find that UMAP offers the most useful insight, which we present below.

We begin by creating 10 groups of sentences, each belonging to a specific semantic category. The first sentence group consists of four variants of the *gpt-trigger*, where three variants have the last token swapped with other racial groups. The second sentence group consists of the adversarial target text that *gpt-trigger* was originally trained on. The third group consists of arbitrary racist text that *gpt-trigger* has not seen or been trained on before. The rest of the sentence groups consist of generated sentences by ChatGPT, including random English sentences, generic positive sentences about black people, white people, Asians, Muslims, sentences about climate change and sentences about Donald Trump.

We collect the word embedding matrix for each of these sentences from GPT-2's final layer and pool individual word embeddings into a sentence embedding by performing a dimension-wise average across tokens. We then perform a UMAP dimensionality reduction on these 768-dimensional average sentence embeddings and project them onto a lower dimensional space to observe similarities between sentence groups. In the 3D projected space (Figure 2), UMAP appears to strongly suggest that the triggers, adversarial target text and arbitrary racist text are tightly clustered together, with visible separation from other sentence groups.

We then check for the robustness of this potentially isolated occurrence by varying the hyperparameters for UMAP and observing the clustering behavior across a range of values. Specifically, we vary the number of nearest neighbors and minimum distance. We observe that for reasonable ranges of hyperparameter values for a given reduced dimension, there

is evidence of stable clustering behavior between the trigger, target text and racist sentence groups. Despite indications that the geometric explanation holds in lower dimensions, an understandable critique could be questioning its validity in higher dimensions. To this end, we vary the number of reduced dimensions to be as large as 250. We find that for distance metrics such as Euclidean, Manhattan, Canberra and Chebyshev, the triggers appear to be closest to the target and racist sentences on average, and have larger average distances from the other groups. A sample visualization for a pair of hyperparameter values (neighbors = 50, distance = 0.25), can be seen in Figure 3. A comprehensive set of results visualizing all variations are available in the Appendix A.

Behavior for other triggers. In order to test if this behavior generalizes across triggers trained on other topics, we choose two topics – anti-vaccination and the Flat Earth conspiracy theory. We borrow the trigger trained on the Flat Earth conspiracy theory by (Heidenreich & Williams, 2021) (*'Historical theories earth sphericalitionally'*) and train our own trigger on anti-vaccination (*'Doctors vaccinated GOvacc Poison Stop asserting Blueprint'*). We repeat the experiment described earlier for these two triggers and include *gpt-trigger*, its target text and arbitrary racist sentences during dimensionality reduction. Overall, we observe similar behavior across a range of UMAP hyperparameter values, dimensions and distance metrics. The target text comprising sentences in support of anti-vaccination and the Flat Earth conspiracy (shown in orange in the Appendix A), is much closer to the trigger on average compared to other sentence groups. A notable difference we observe between this experiment and the one with *gpt-trigger* is that these triggers appear to be close to semantically similar sentences with non-adversarial content. That is, the anti-vaccination and Flat Earth triggers are close to sentences supporting vaccination and opposing the Flat Earth theory respectively (shown in brown in the Appendix A). During text generation, we also observe similar behavior, where these triggers do not consistently generate adversarial text. That is, some generated sentences are in support of vaccination and in opposition of the Flat Earth theory. (Heidenreich & Williams, 2021) also showed that triggers trained on niche topics tend to be weaker than those trained on broader topics, which could be a possible explanation. In contrast, *gpt-trigger* was much closer to arbitrary racist text and farther away from positive sentences about racial groups. The complete set of results visualizing these observations are available in the Appendix A.

Our initial experimental evidence appears to strongly suggest the validity of our geometric explanation for universal adversarial triggers. We show that for a range of reduced UMAP dimensions, across varying distance metrics and reasonable hyperparameter values, the universal adversarial trigger is closest in distance to the adversarial target text it is trained on and arbitrary text which is semantically similar to the target text. This behavior is also seen to replicate across trigger topics. In other words, the trigger might indeed be behaving like a vector which embeds the adversarial target text it is trained on, which in the case of *gpt-trigger*, happened to be racist and offensive.

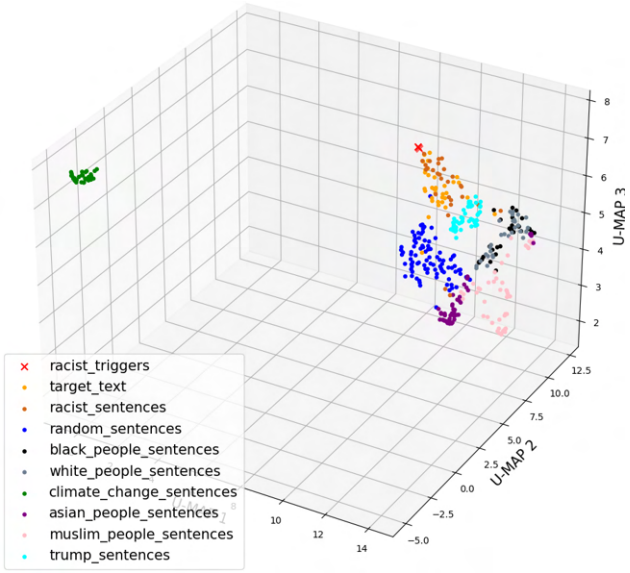

Figure 2. Sample UMAP dimensionality reduction (neighbors = 15, minimum distance = 0.2) on sentence groups. Triggers (red), adversarial training text (orange) and arbitrary unseen racist sentences (brown) cluster together, indicating a semantic neighborhood. Semantically similar sentence groups (black, white, Asian and Muslim people) also cluster together.

7. Potential Explanations, Limitations & Future Work

Based on our initial findings supporting the geometric explanation for universal adversarial attacks, we hypothesize potential explanations for observed behavior in these triggers.

1. We hypothesize that universal adversarial triggers might be effective because the model sees a semantically meaningful adversarial input instead of a gibberish trigger and generates on top of it.

2. Longer triggers were found to be more effective by Wallace et al. (2019). This could potentially be explained by the fact that more tokens give more room for the trigger to capture semantic information about the adversarial training text.

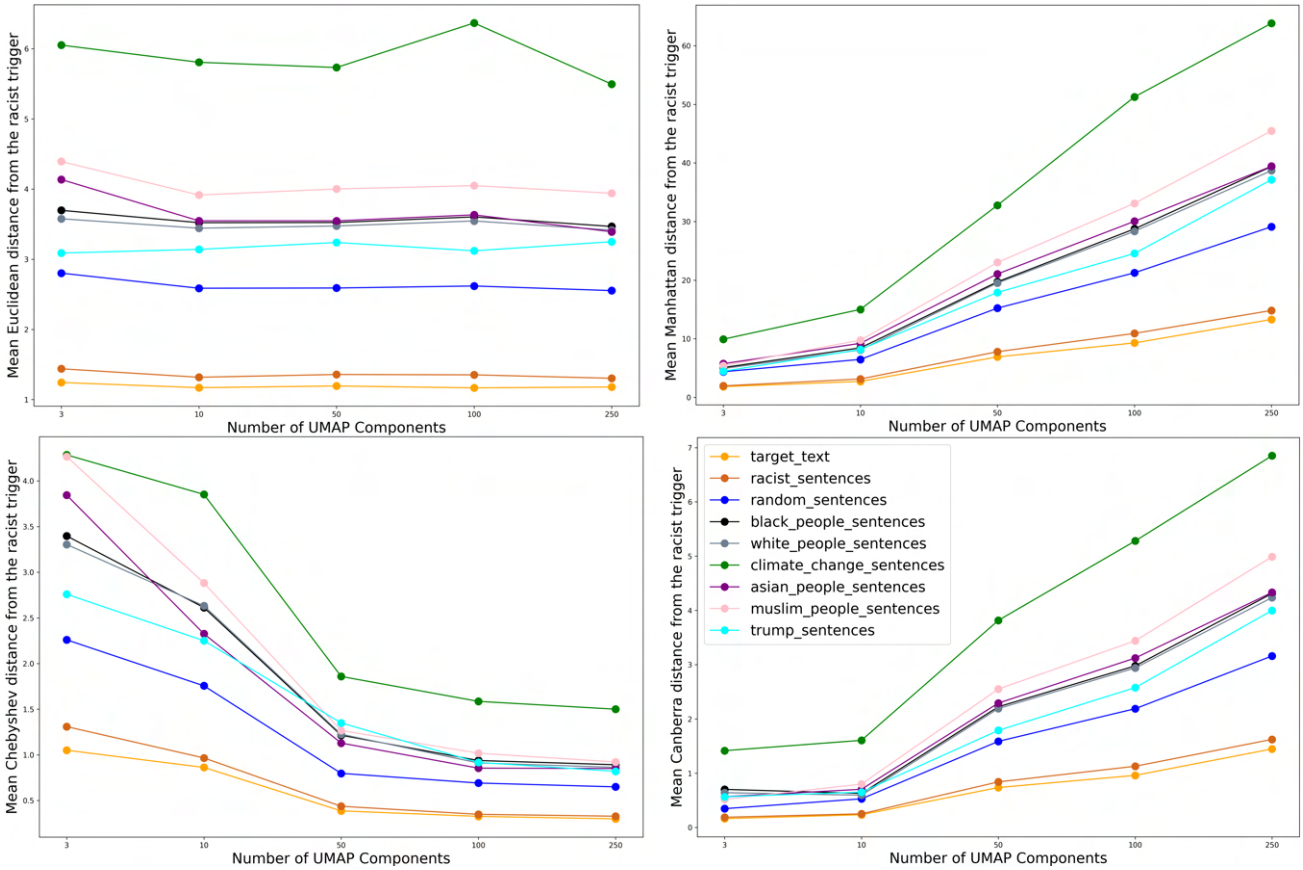

Figure 3. Across a range of UMAP hyperparameters (here, neighbors = 50, minimum distance = 0.25) and a range of reduced UMAP dimensions (3 to 250), we see consistently lower average Euclidean, Manhattan, Chebyshev & Canberra distances between the trigger, target (orange) and racist (brown) text. Here, average distance is computed as the average of the distances between the trigger and all sentences in a group.

3. Triggers are seen to transfer across models. We observe that this transferability occurs across a number of models using the same tokenization algorithm. This behavior has also been studied by Zou et al. (2023) who generate universal adversarial prompts which cause open and proprietary foundation models to generate harmful content. Such an attack has been show to transfer across different model families like ChatGPT, Llama, PaLM, Bard and Claude. Understanding transferability of universal attacks is an open problem and can be an interesting direction for future work.

Despite encouraging initial evidence, we note that experimental techniques involving dimensionality reduction must be considered with caution. The primary limitation of this work is the usage of dimensionality reduction which could be providing an incomplete picture. Future lines of work in interpretability include expanding experimentation to additional techniques. This can include studying the evolution of distances between sentence groups across layers, visualizing the evolution of GPT-2's hidden representations, analyzing the Kullback–Leibler divergence between the distribution of predicted tokens for adversarial and non-adversarial settings, etc.

8. Conclusion

In this work, we present a novel geometric perspective which potentially explains gradient-based universal adversarial attacks on large language models. We find initial experimental evidence indicating that these triggers could be behaving like embedding vectors which approximate the semantic information in their adversarial training region. We further leverage this perspective to offer potential explanations for observed behavior of these triggers in literature. With this novel geometric perspective, we aim to shed light on the underlying mechanism driving universal attacks. This would further enable us to understand the failure modes of LLMs and importantly, enable their mitigation.

Acknowledgements

We thank the reviewers of 'The 2nd New Frontiers in Adversarial Machine Learning Workshop' ICML 2023, for their thoughtful and valuable feedback which helped improve this work. We also thank Yoon Kim, Siddharth Swaroop, Abbas Zeitoun and Ani Nrusimha for their generous technical input and helpful suggestions.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207, 2016. URL http://arxiv.org/abs/1608.04207.

Atmakuri, S., Chheda, T., Kandula, D., Yadav, N., Lee, T., and Tuinhof, H. Robustness of explanation methods for nlp models, 2022.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. R. What do neural machine translation models learn about morphology? *CoRR*, abs/1704.03471, 2017. URL http://arxiv.org/abs/1704.03471.

Bellman, R. Dynamic programming. *Science*, 153(3731): 34–37, 1966.

Brunner, G., Liu, Y., Pascual, D., Richter, O., and Wattenhofer, R. On the validity of self-attention as explanation in transformer models. *CoRR*, abs/1908.04211, 2019. URL http://arxiv.org/abs/1908.04211.

Chizhikova, A., Murzakhmetov, S., Serikov, O., Shavrina, T., and Burtsev, M. Attention understands semantic relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4040–4050, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.430.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341, 2019. URL http://arxiv.org/abs/1906.04341.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. B., and Wattenberg, M. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715, 2019. URL http://arxiv.org/abs/1906.02715.

DeRose, J. F., Wang, J., and Berger, M. Attention flows: Analyzing and comparing attention mechanisms in language models. *CoRR*, abs/2009.07053, 2020. URL https://arxiv.org/abs/2009.07053.

Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. Hotflip: White-box adversarial examples for NLP. *CoRR*, abs/1712.06751, 2017. URL http://arxiv.org/abs/1712.06751.

Engstrom, L., Gilmer, J., Goh, G., Hendrycks, D., Ilyas, A., Madry, A., Nakano, R., Nakkiran, P., Santurkar, S., Tran, B., Tsipras, D., and Wallace, E. A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, 2019. doi: 10.23915/distill.00019. https://distill.pub/2019/advex-bugs-discussion.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks, 2021. URL https://arxiv.org/abs/2106.02997.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. *CoRR*, abs/2012.14913, 2020. URL https://arxiv.org/abs/2012.14913.

Heidenreich, H. S. and Williams, J. R. 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, pp. 566–573, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462578. URL https://doi.org/10.1145/3461702.3462578.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. 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)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features, 2019.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. Revealing the dark secrets of BERT. *CoRR*, abs/1908.08593, 2019. URL http://arxiv.org/abs/1908.08593.

Lin, Y., Tan, Y. C., and Frank, R. Open sesame: Getting inside bert's linguistic knowledge, 2019.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt, 2022. URL https://arxiv.org/abs/2202.05262.

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016. URL http://arxiv.org/abs/1610.08401.

Mosca, E., Agarwal, S., Ramí rez, J. R., and Groh, G. "that is a suspicious reaction!": Interpreting logits variation to detect NLP adversarial attacks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.538. URL https://doi.org/10.18653%2Fv1%2F2022.acl-long.538.

Nanda, N. and Lieberum, T. A mechanistic interpretability analysis of grokking. *Alignment Forum*, Aug 2022. URL https://www.alignmentforum.org/posts/N6WM6hs7RQMKDhYjB/a-mechanistic-interpretability-analysis-of-grokking.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations, 2018.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Singla, Y. K., Parekh, S., Singh, S., Chen, C., Krishnamurthy, B., and Shah, R. R. Minimal: Mining models for universal adversarial triggers. In *AAAI*, 2022.

Song, L., Yu, X., Peng, H.-T., and Narasimhan, K. Universal adversarial attacks with natural triggers for text classification. *ArXiv*, abs/2005.00174, 2021.

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical nlp pipeline, 2019.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. Universal adversarial triggers for NLP. *CoRR*, abs/1908.07125, 2019. URL http://arxiv.org/abs/1908.07125.

Wallat, J., Singh, J., and Anand, A. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 174–183, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.17. URL https://aclanthology.org/2020.blackboxnlp-1.17.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022.

Zhang, W. E., Sheng, Q. Z., and Alhazmi, A. Generating textual adversarial examples for deep learning models: A survey. *CoRR*, abs/1901.06796, 2019. URL http://arxiv.org/abs/1901.06796.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023.

A. Appendix

A.1. Plots for the Racist Trigger

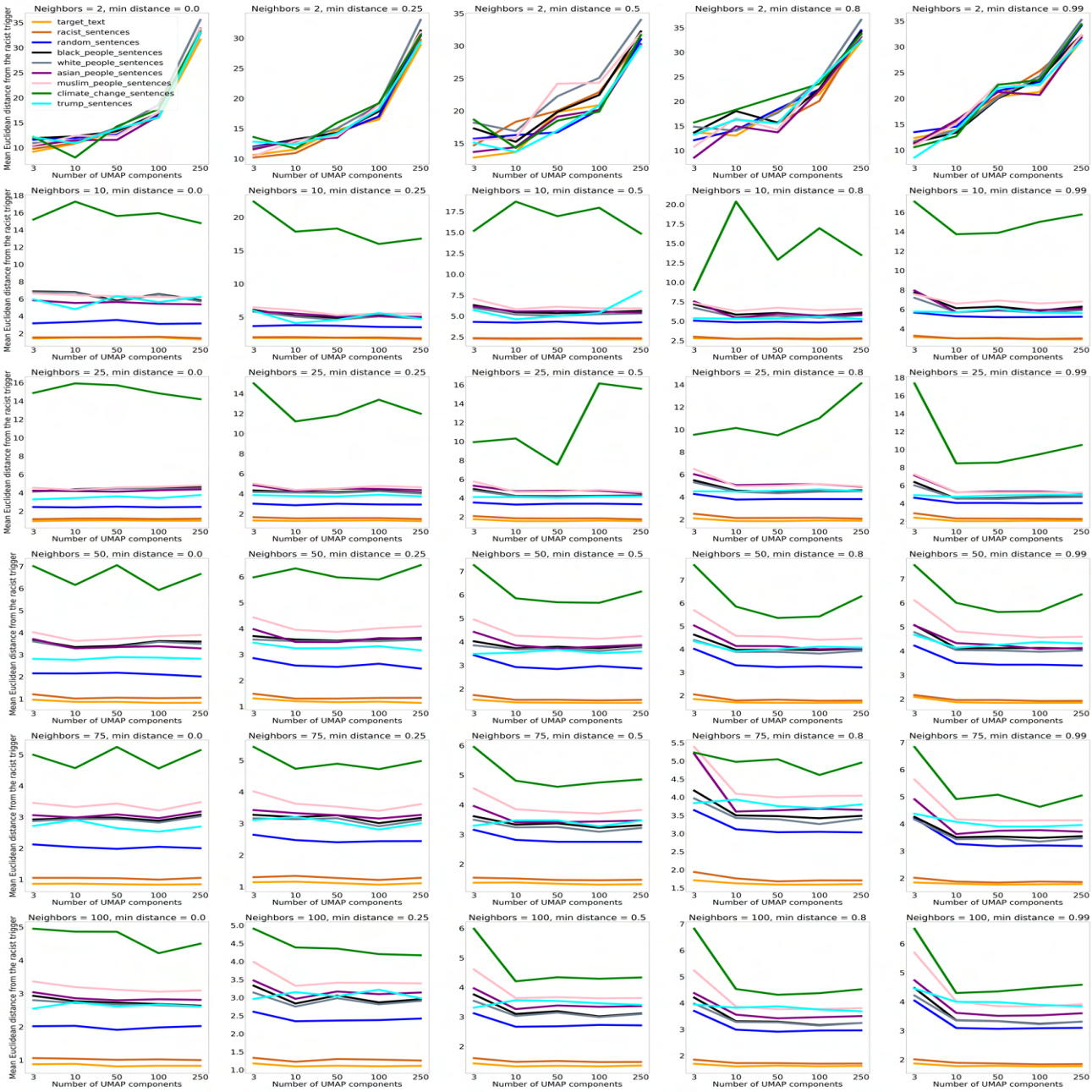

Figure 4. Euclidean distance between the racist trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

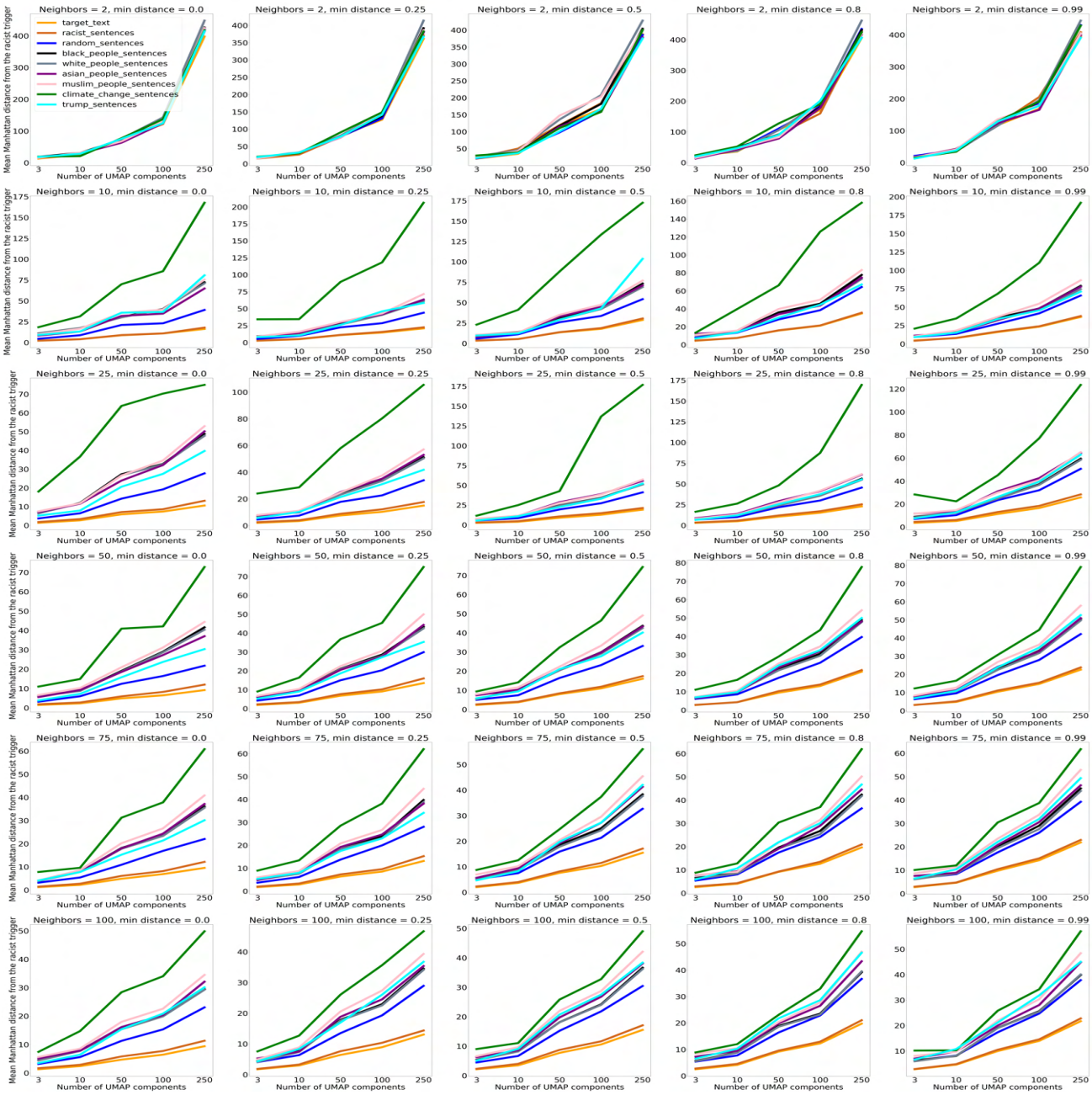

Figure 5. Manhattan distance between the racist trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

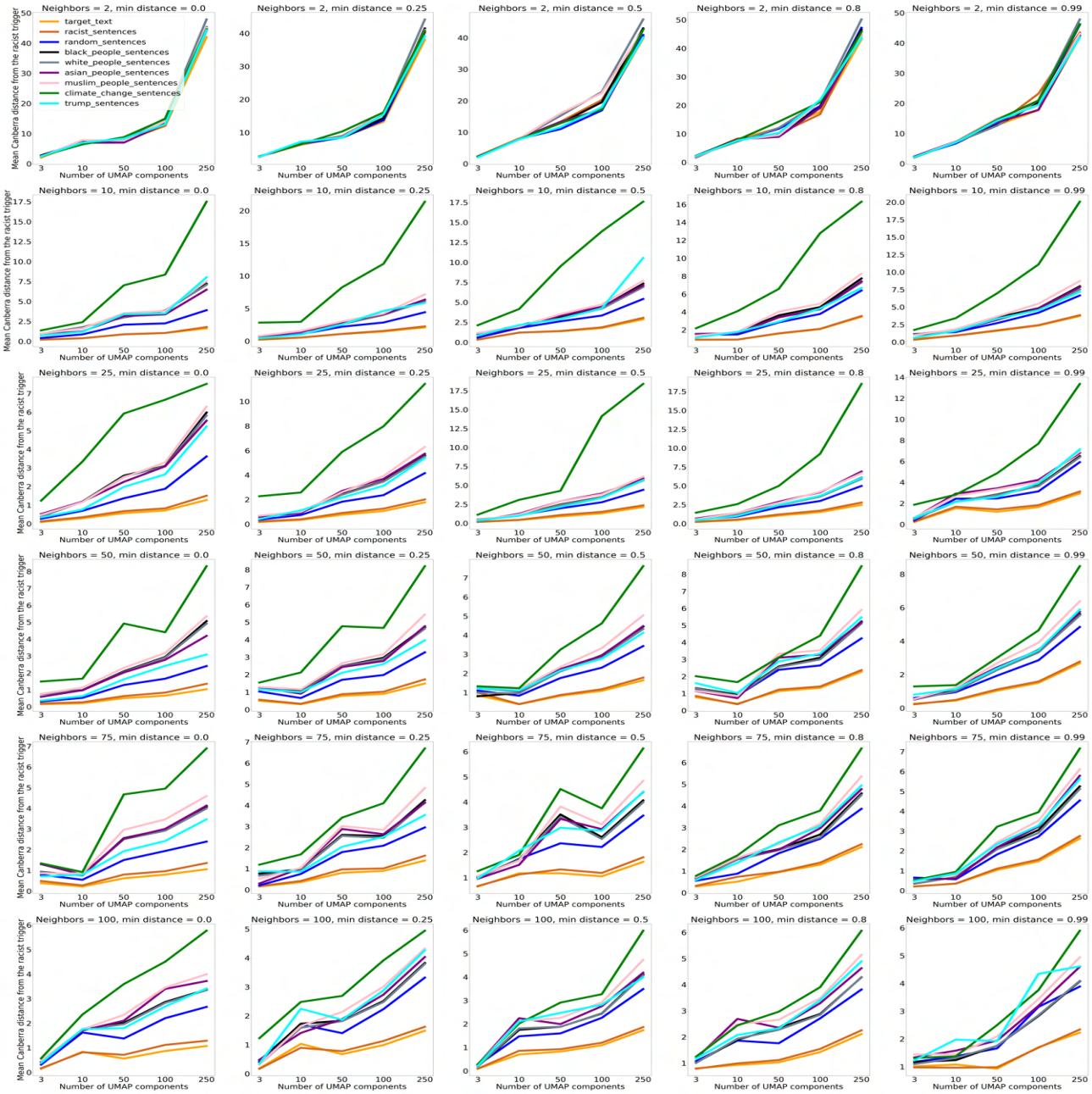

Figure 6. Canberra distance between the racist trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

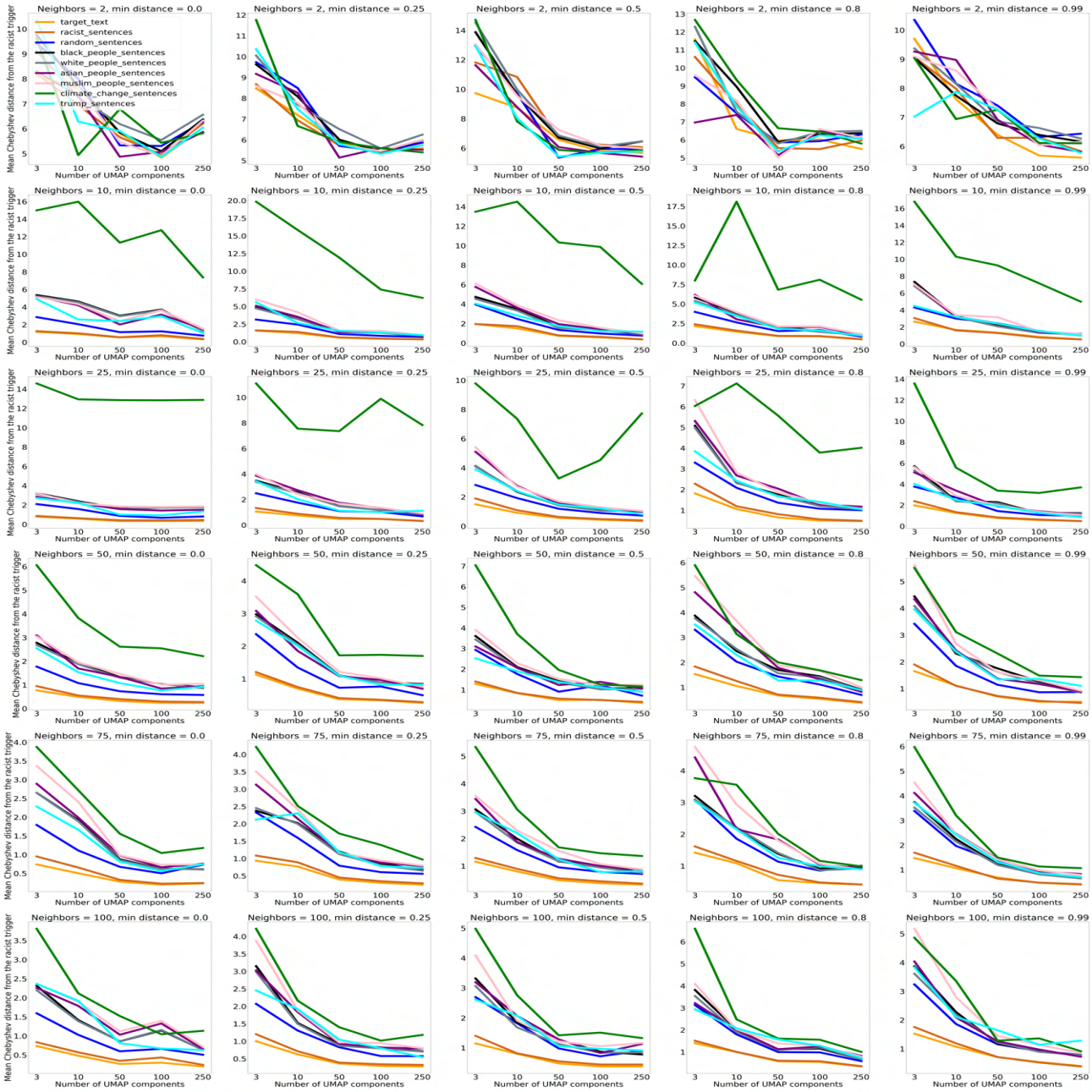

Figure 7. Chebyshev distance between the racist trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

A.2. Plots for the Flat Earth & Anti-Vaccination Triggers

Figure 8. Euclidean distance between the Flat Earth trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

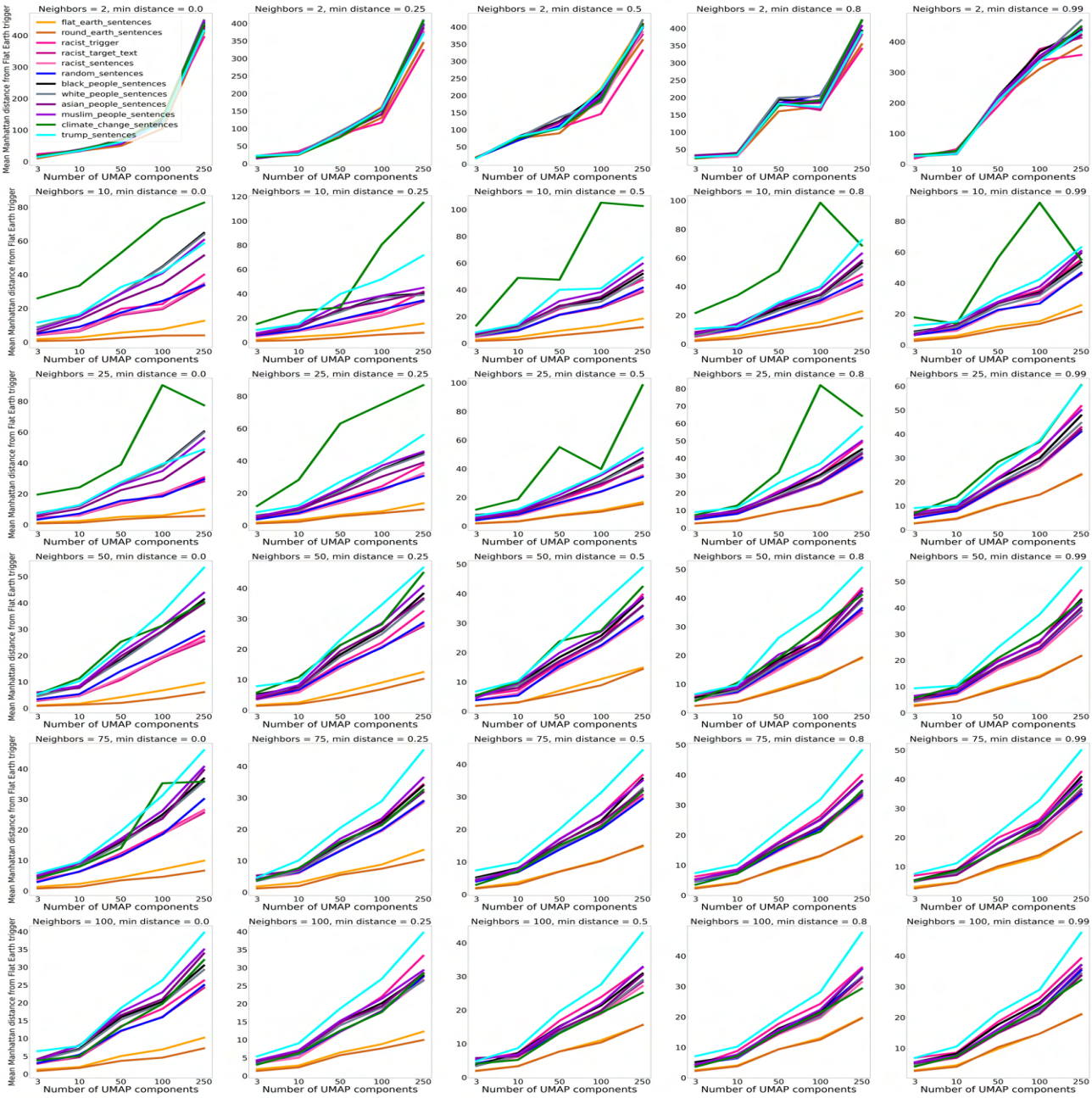

Figure 9. Manhattan distance between the Flat Earth trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

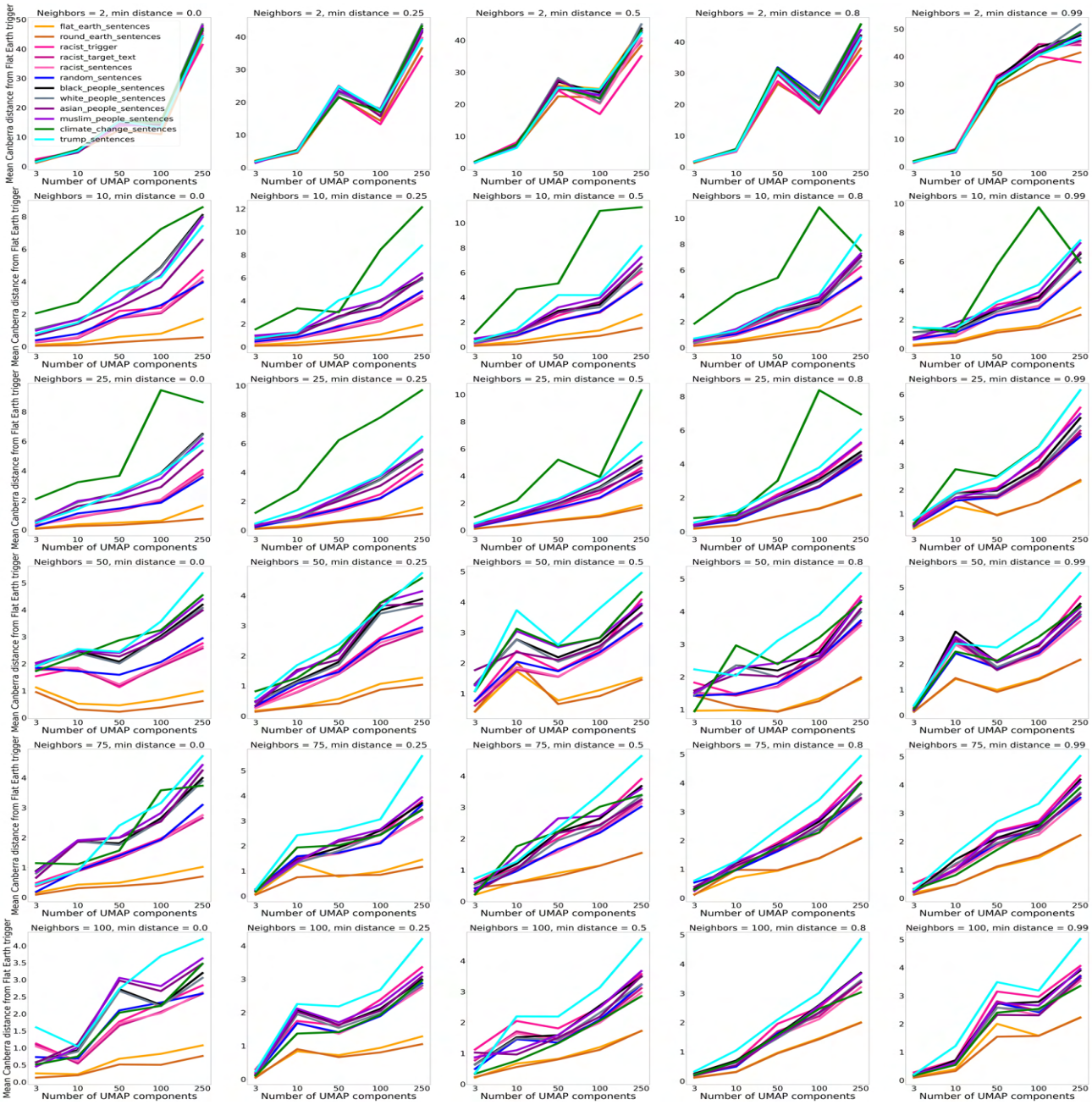

Figure 10. Canberra distance between the Flat Earth trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

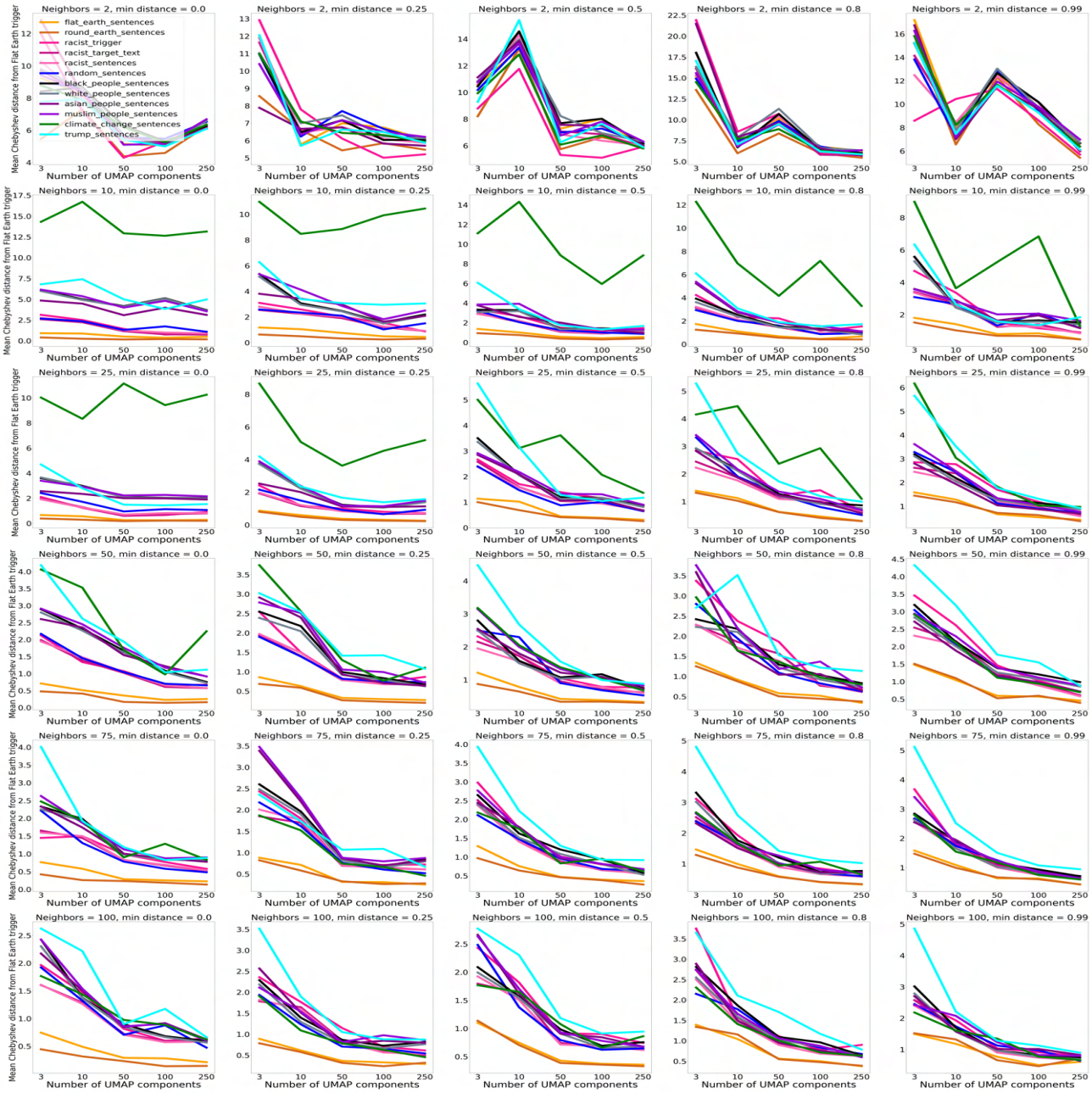

Figure 11. Chebyshev distance between the Flat Earth trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

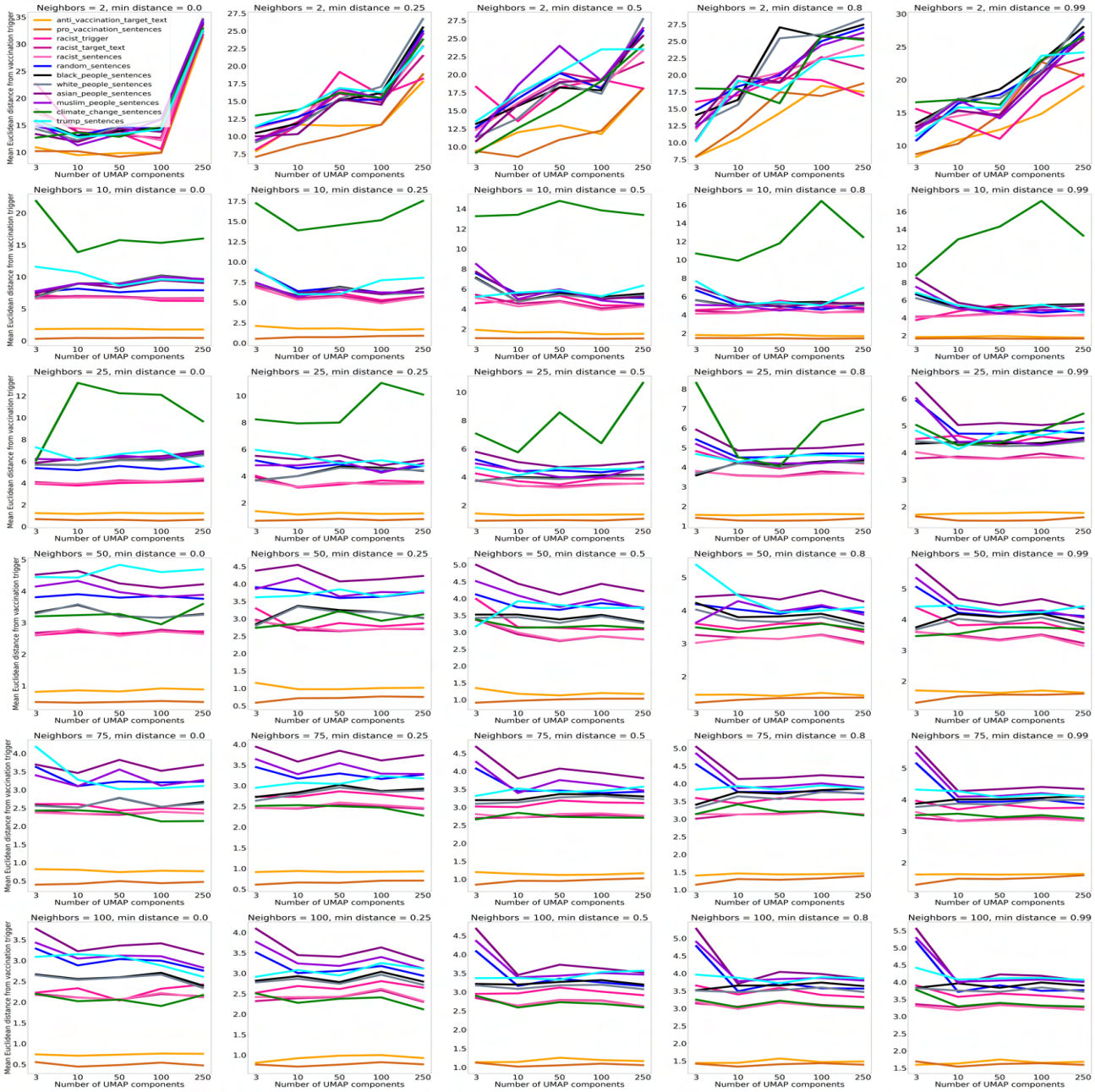

Figure 12. Euclidean distance between the vaccination trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

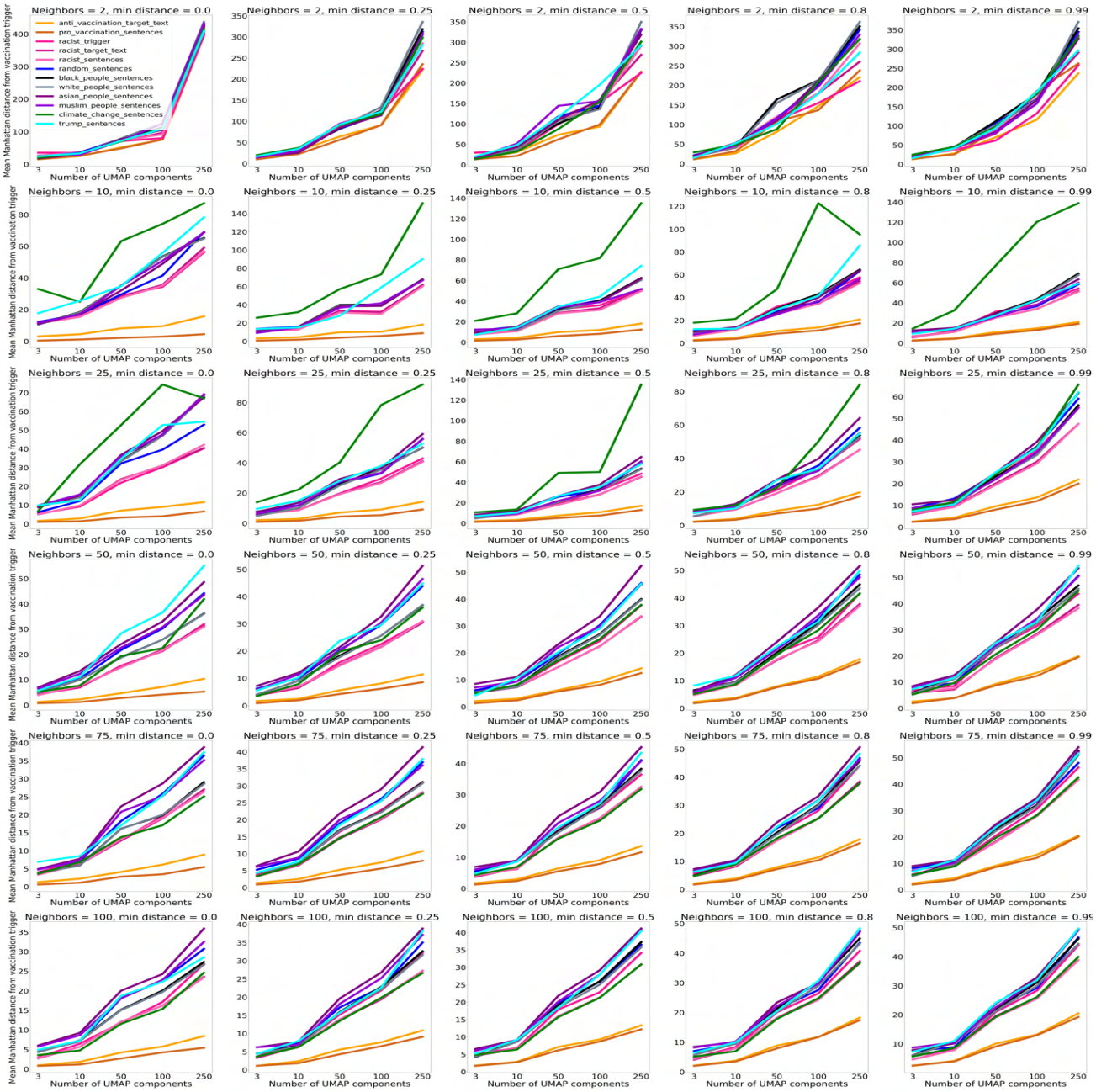

Figure 13. Manhattan distance between the vaccination trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

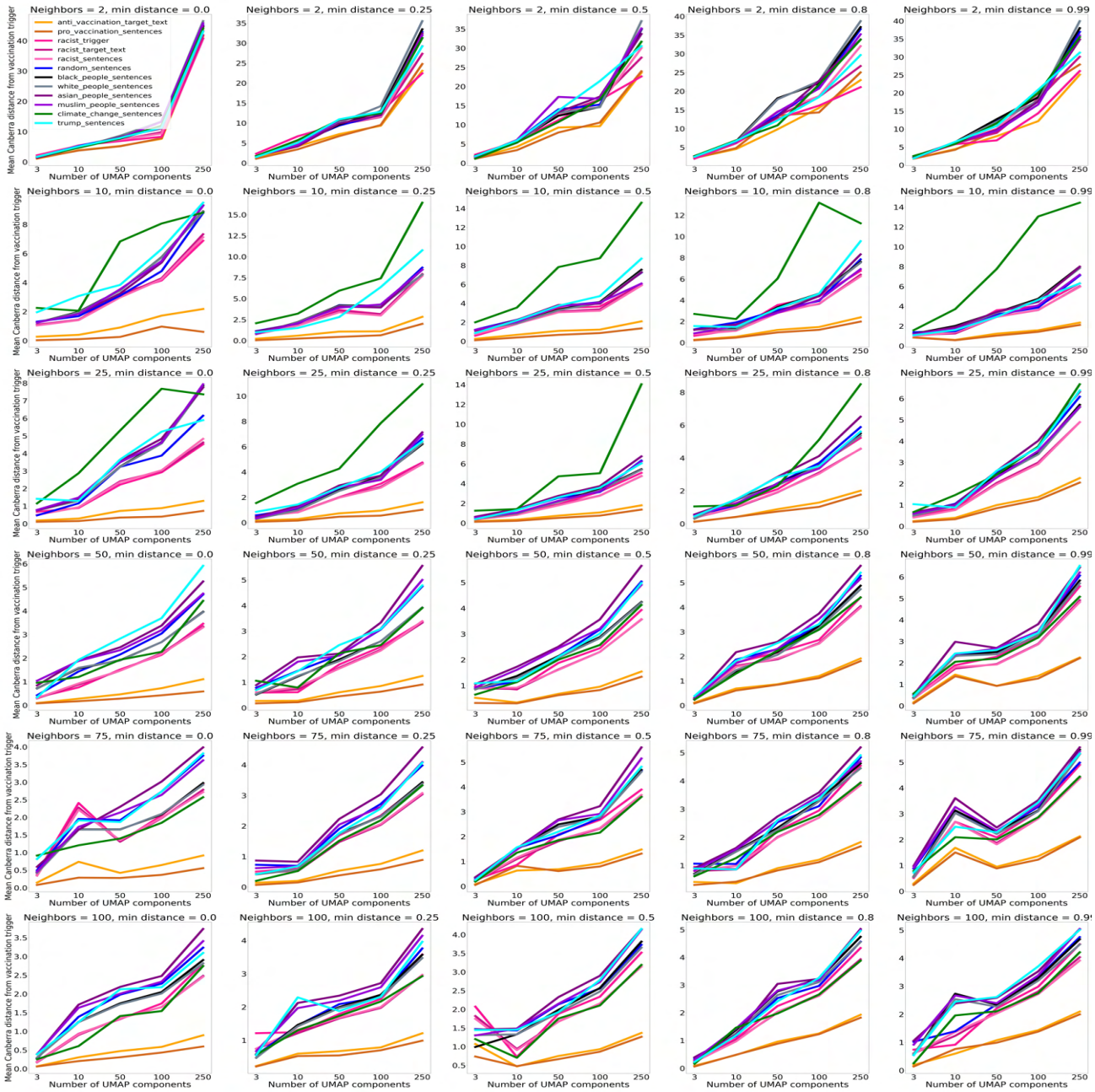

Figure 14. Canberra distance between the vaccination trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.

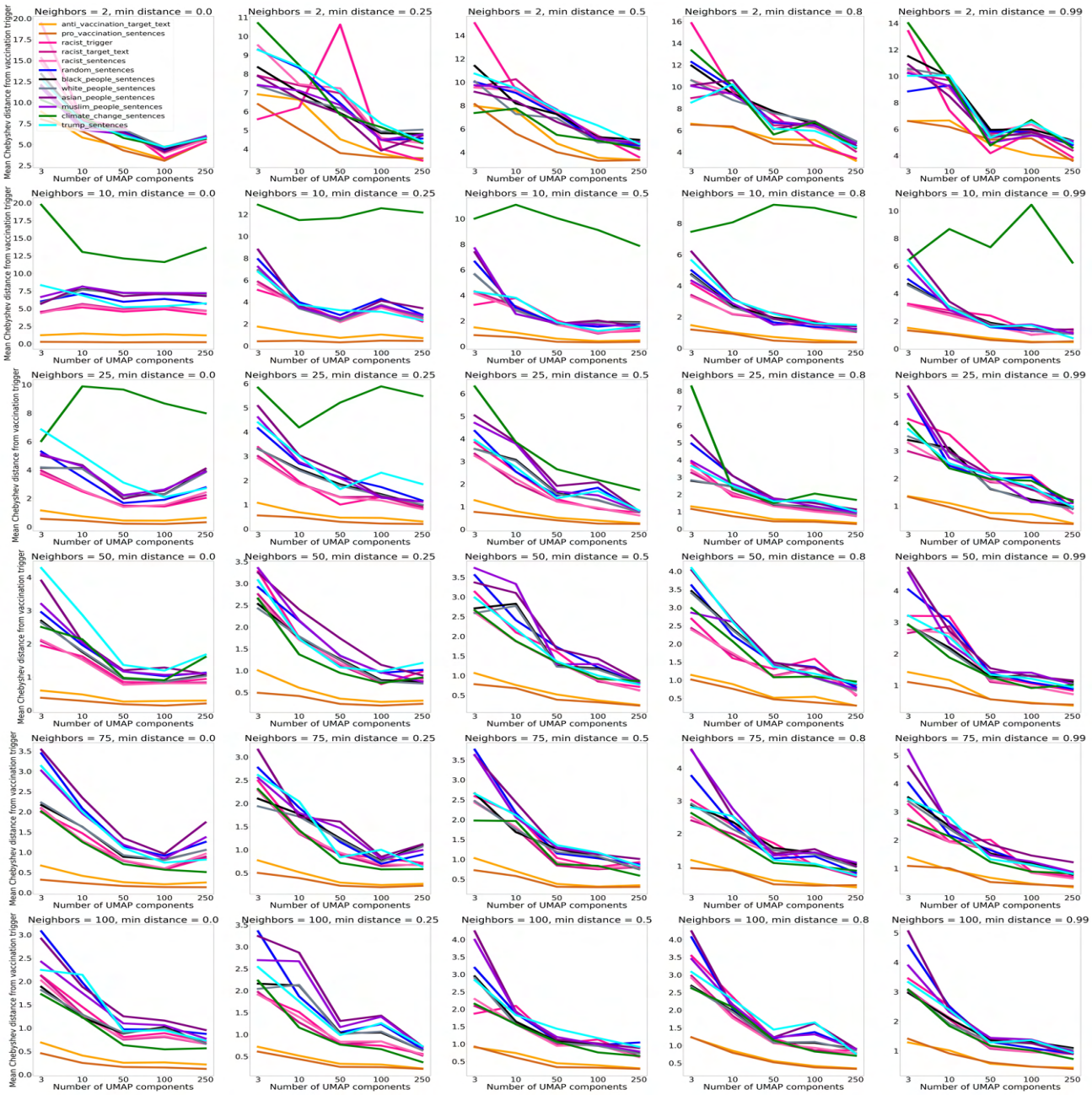

Figure 15. Chebyshev distance between the vaccination trigger and other sentence groups for varying UMAP hyperparameters and reduced dimensions.