Weakly Supervised Detection of Hallucinations in LLM Activations

Miriam Rateike
Saarland University
IBM Research Africa
Nairobi, Kenya

Celia Cintas
IBM Research Africa
Nairobi, Kenya

John Wamburu
IBM Research Africa
Nairobi, Kenya

Tanya Akumu
IBM Research Africa
Nairobi, Kenya

Skyler Speakman
IBM Research Africa
Nairobi, Kenya

Abstract

We propose an auditing method to identify whether a large language model (LLM) encodes patterns such as hallucinations in its internal states, which may propagate to downstream tasks. We introduce a weakly supervised auditing technique using a subset scanning approach to detect anomalous patterns in LLM activations from pre-trained models. Importantly, our method does not need knowledge of the type of patterns *a-priori*. Instead, it relies on a reference dataset devoid of anomalies during testing. Further, our approach enables the identification of pivotal nodes responsible for encoding these patterns, which may offer crucial insights for fine-tuning specific sub-networks for bias mitigation. We introduce two new scanning methods to handle LLM activations for anomalous sentences that may deviate from the expected distribution in either direction. Our results confirm prior findings of BERT's limited internal capacity for encoding hallucinations, while OPT appears capable of encoding hallucination information internally. Importantly, our scanning approach, without prior exposure to false statements, performs comparably to a fully supervised out-of-distribution classifier.

1 Introduction

The rapid proliferation of Large Language Models (LLMs) has transformed the landscape of natural language processing, empowering applications ranging from chatbots and dialogue systems [57] to content generation [45, 46]. However, as these models become an integral part of our communication, there are concerns about the potential biases (e.g., hallucinations [1], toxicity, stereotypes) embedded within their outputs [14, 33, 42, 55, 59]. These subtle and implicit biases can reinforce stereotypes, marginalize certain groups, and perpetuate inequalities [7, 21]. Auditing LLMs for bias is thus essential for upholding ethical standards, reducing harm, and an inclusive deployment.

Despite advances in bias detection and mitigation strategies for LLMs in recent years, a large corpus of prior work has focused on word-level representations [5, 9, 49]. However, in recent years, sentence-level representations from models such as GPT [45, 46] have become prevalent for text sequence encoding. Moreover, prior research has primarily operated under the assumption that the particular type of bias, such as hallucinations, is known *a priori*, i.e., during the training phase [3, 32]. However, this assumption can potentially restrict these methods' practical applicability and generalizability

[1] The term "hallucinations" refers to factual errors [3].

Socially Responsible Language Modelling Research (SoLaR) Workshop at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

since they rely on access to labeled data, which may necessitate a resource-intensive data collection process. Gaining access to labeled data can be particularly challenging when the occurrence of a bias is infrequent and given that language constantly evolves, which inherently leads to the continuous change of biases and their linguistic expressions. Moreover, exposure to harmful content can lead to emotional and psychological stress for content moderators/labelers [38].

In this work, we propose an auditing approach to bias detection in LLMs sentence embeddings when the bias is *not known a-priori*. Our goal is to determine if a pre-trained LLM has internalized harmful anomalous patterns (e.g., hallucinations) by examining its internal states (node activations). Following prior work [3, 24], our underlying assumption is that an LLM capable of predicting or generating anomalous content will exhibit detectable indicators of this tendency within its internal states.

Our method extends prior work on anomalous subset scanning for neural networks [13, 26, 27] by scanning pre-trained LLM activations. Our method operates without the need for training data containing labeled anomalous content (e.g., false/hallucinated statements). Unlike classifiers, we do not require a training phase for a particular bias type. Instead, during testing, we rely on a reference dataset assumed to contain "normal" (or safe) content devoid of anomalies. When testing, we input a test dataset, possibly containing bias, to a pre-trained LLM model under audit. Our method examines the LLM's hidden states (activations) and identifies a subset of input data (e.g., sentences and nodes) as anomalous. We extend the prior work [13, 26, 27] to address that LLM embeddings for anomalous sentences may deviate in either direction from the expected distribution. If anomalous sentences, e.g., those containing stereotypes, are detected from the activations, it suggests the model encodes those anomalous patterns. Conversely, the absence of anomalies detected in the activations of anomalous sentences may indicate the model's potential robustness to those patterns.[2]

1.1 Related Work

Auditing LLM Outputs. Prior work on detecting anomalies such as stereotypes, toxicity, or hallucinations in LLM models has concentrated on analyzing the model's generated content such as the percentage of anomalous options preferred or chosen [41, 29]. Other work has explored the propagation of bias to downstream tasks, including coreference resolution [58], sentiment analysis [42], topic modeling [20], and prediction models [18]. However, the effectiveness of these approaches is heavily reliant on the quality of pre-trained downstream models. A different line of work has examined bias in the activations of LLMs, using principal component analysis [4, 32, 58], clustering [4], or training detection classifiers on the latent space [3, 4, 8, 22]. Other work has studied distance metrics between word pair representations [8, 4]. However, this approach has shown inconsistency detection results within contextual scenarios [22, 29, 35]. Furthermore, these approaches assume the availability of fully labeled training data and require predefined anomalous patterns. Few prior work has addressed the identification of unknown biases in LLMs, particularly in the context of unbiased sentence classification [52]. In this work, our goal is to detect whether an LLM encodes anomalies (e.g., hallucinations) within its hidden states. We work under the assumption that only "normal" (e.g., true) data is available, while the presence of anomalous (e.g., false) data remains undisclosed.

DeepScan In the context of analyzing data using a pre-trained network, deep subset scanning (DeepScan) [13] has been used to detect anomalous samples in various computer vision and audio tasks, including creativity sample characterization [12], audio adversarial attacks in inner layers of autoencoders [2], patch-based attacks in flow networks [26] and skin condition classification [27]. In this work, we extend prior work by scanning pre-trained LLM activations and introducing two novel methods to effectively identify anomalous sentences deviating from the expected activation distribution in either direction.

2 DeepScan for LLMs

This section introduces adaptations and extensions to deep subset scanning (DeepScan) [13, 26] for auditing LLM activations. Specifically, we detail the adaptation of previous deep scanning approaches to search for the "most anomalous" subset of node activations and input sentences within the inner layers of a pre-trained LLM network. We assume two datasets: a reference dataset $\mathcal{B}$ containing B "normal" (e.g., factually true) sentences, and an independent test dataset $\mathcal{T}$ containing M sentences,

[2]Note that we cannot confirm a null hypothesis (absence of anomalies).

which may be either "normal" or "anomalous" (e.g., factually false). A sentence can represent any continuous text span and is not limited to a traditional linguistic sentence [25].

Problem Formulation. Consider an LLM, such as BERT [25], which can generate activations through its encoder[3]. Assume we have M test sentences represented by a vector of activations $Z^l = [Z_1^l, \ldots, Z_M^l]$ generated by the LLM at layer l, with each sentence activation having dimension J corresponding to the set of nodes $O^l = \{O_1^l, \ldots, O_J^l\}$. Now, let $Z_S \subseteq Z$ and $O_S \subseteq O$, then we define a subset over sentences and nodes as $S = Z_S \times O_S$. Our goal is to identify the subset of activations containing the most anomalous (e.g., hallucinated) content based on a scoring function $F(S)$ that yields the anomaly score of a subset S: $S^\star = \arg\max_S F(S)$.

Detecting anomalies in activations typically requires parametric assumptions for the scoring function (e.g., Gaussian, Poisson). However, given that the distribution of activations in specific layers can be highly skewed, we adopt a non-parametric approach, following prior work [11, 13, 27, 36, 37]. This approach, known as non-parametric scan statistics (NPSS), makes minimal assumptions about the underlying distribution of node activations. For this, we first derive p-values from the activations and then perform a scan over these p-values to quantify the difference or shift in the activation distribution for each dimension (node) compared to the reference distribution. We detail this now.

Empirical p-values. In line with prior work [13], we utilize the activations from the ("normal") reference data $\mathcal{B}$ to compute empirical p-values for the activations from the ("normal" or "anomalous") test data $\mathcal{T}$. For a given test activation $z_{mj}^{\mathcal{T}l}$ (corresponding to sentence m, layer l and node j), we calculate its empirical p-value by first sorting the set of activations from the reference data $\{z_{bj}^{\mathcal{B}l}\}_{b=1}^{B}$ corresponding to layer l and node j across all reference sentences $b = 1 \ldots B$ and then determining the rank of the test activation within that sorted list of reference activations. Subsequently, we normalize these positions to generate p-values within $[0, 1]$:

$$p_{mj}^l = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(z_{bj}^{\mathcal{B}l} \geq z_{mj}^{\mathcal{T}l})}{1 + B}. \tag{1}$$

Here, $\mathbf{1}(\cdot)$ is the indicator function. For ties, we consider both the rank on the left (`pmin`) and right sides (`pmax`). This gives us a range $[\texttt{pmin}, \texttt{pmax}]$. To obtain a single p-value within the range, we perform uniform sampling. We compute p-values for a given layer l for each node j and sentence m of the test activations.

There are different methods for computing these empirical p-values. In left-tail (right-tail) p-values, the focus is on extreme values on the left (right) side of the reference activation distribution, indicating smaller (larger) values compared to the reference dataset. For two-tailed p-values, the focus is on extreme values on both (left and right) sides. Prior research concentrated on one-tailed p-values [13, 27] on the left side of the activation distribution. However, in our case, we observe deviations on both sides of the distribution and introduce two novel methods to incorporate the extreme values from both ends, as we detail below.

Uniform Distribution of p-values Under Null Hypothesis When the null hypothesis is true, it implies that any observed data point has an equal chance of falling anywhere within the distribution of possible values under the null hypothesis [39, 48]. Therefore, when we calculate empirical p-values by determining how extreme our observed data is relative to this null distribution, each potential outcome is equally likely. This uniformity in probabilities across the distribution ensures that p-values for samples confirming the null hypothesis follow a uniform distribution, as they are essentially measuring the randomness of the data in a manner consistent with the null hypothesis's assumptions. Thus if the test dataset were to contain only "normal" sentences (null hypothesis), the p-values for test activations at layer l would exhibit for each node j a uniform distribution across sentences. When anomalous sentences are introduced, and the LLM activations encode these anomalies, we hypothesize a departure from this uniform p-value distribution, particularly for certain nodes j.

Scoring Function. To test whether the p-value distributions diverge from a uniform distribution, we employ a scoring function based on a test statistic, denoted as $F(S) = \max_\alpha F_\alpha(S)$, where α represents a significance level, and $F_\alpha(S)$ is defined by a suitable goodness-of-fit statistic.

[3] Or decoder for decoder-only models like GPT [45, 46] or OPT [56].

In the following explanation, we closely follow [13]. The general form of the scoring function is:

$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \phi(\alpha, N_{\alpha}(S), N(S)) \tag{2}$$

where $N(S)$ represents the number of empirical p-values contained in subset S, $N_{\alpha}(S)$ is the number of p-values less than (significance level) α contained in subset S, $\alpha \in (0, 1)$ is a significance level and ϕ is a goodness-of-fit statistics. To identify a subset S that presents the strongest indication of significantly exceeding the expected activation distribution under the null hypothesis ("normal" or clean data). This is expressed by the condition $N_{\alpha}(S) > \alpha N(S)$, where α denotes the chosen significance level. In our experiments, we run a grid search over $\alpha \in [0.05, 0.5]$ in steps of 0.05.

Higher Criticism Test Statistic While there are several established goodness-of-fit statistics available for use in NPSS [37], in this work, we utilize the Higher Criticism (HC) test statistic [17]:

$$\phi(\alpha, N_{\alpha}(S), N(S)) = \frac{|N_{\alpha}(S) - N(S)\alpha|}{\sqrt{N(S)\alpha(1-\alpha)}} \tag{3}$$

This could be understood as the test statistic for a Wald test assessing the number of significant p-values, where N_{α} follows a binomial distribution with parameters N_{α} and α. Due to its normalization by the standard deviation of N_{α}, HC tends to yield smaller subsets characterized by wider ranges of p-values. This occurs because such subsets yield larger values in the numerator while generating smaller values in the denominator. In our case, small subsets are particularly preferable in scenarios where the quantity of anomalous data within the test dataset is small, as shown in our experiments (Section 3), where the test dataset comprises only 10-20% anomalous data.

Efficient Search Algorithm. To overcome the computational challenge posed by maximizing a scoring function across all possible data sample and node subsets, we employ Fast Generalized Subset Scanning (FGSS) as previously utilized in similar research [13, 27]. This approach significantly reduces the number of subsets under consideration from $O(2^E)$ to $O(E)$ within each optimization step, where E represents the number of elements currently being optimized, such as images or nodes (see Appendix C, Algorithm 2). This efficiency is based on the application of the LTSS property [43], which involves sorting each element based on its priority, defined as the proportion of p-values below a threshold α. FGSS assures convergence to a local optimum. The algorithm returns anomalous subset $S^{\star}$ defined by a set of nodes $O_{S^{\star}}$ and a subset of sentences $Z_{S^{\star}}$ from the test dataset, collectively defining the most anomalous pattern as a group.

Using results from both tails. We observe that LLM embeddings for anomalous data may shift from the expected reference distribution in both directions. To identify subsets marked as anomalous due to shifts to the left or right for different nodes, we introduce two novel methods to aggregate scanning results. The first approach involves aggregating results obtained from scanning left-tail and right-tail p-values. We identify subset of sentences $Z_{S^{\star}_R}$ by scanning over right-tailed p-values and $Z_{S^{\star}_L}$ by scanning over left-tailed p-values. We then combine these results through union: $Z^{\text{Union}}_{S^{\star}} := Z_{S^{\star}_R} \cup Z_{S^{\star}_L}$. The second approach is an iterative method that combines results from scanning two-sided p-values. It aggregates the top-k subsets returned by the scanning, where after each iteration $i = 1 \dots k$, the found subset of sentences $Z_{S^{\star}_i}$ is removed from the test dataset, and the subsequent scanning is performed on the reduced test set. The final subset is the union over all identified subsets: $Z^{\text{top}-k}_{S^{\star}} := Z_{S^{\star}_1} \cup Z_{S^{\star}_2} \cup \dots Z_{S^{\star}_k}$.

3 Experimental Setup and Results

This section presents experimental results for bias detection in LLMs using our two proposed scanning methods. We focus on hallucination detection and analyze the subset of input sentences returned. For additional results on toxicity and stereotype detection, refer to Appendix E.

Data and LLM Models. We use an English-language dataset Hallucinations [3] ('Cities' topic) containing factually true (e.g., "Nakuru is a city in Kenya.") and false statements (e.g., "Surrey is a city in Kenya."). We use a test dataset of 800 samples comprising 10% anomalous data, reflecting the real-world scarcity of such data. We sample test data 10 times with replacement from a larger pool of data and report mean and standard deviation. For further details regarding the dataset selection

LLM	Layer	Clf	Precision	Recall	Size
BERT	10	clf+	**0.099 (0.006)**	**0.73 (0.046)**	0.737 (0.01)
		scan2	**0.114 (0.022)**	0.631 (0.105)	**0.563 (0.075)**
		scanLR	0.091 (0.017)	0.61 (0.062)	0.685 (0.075)
	12	clf+	**0.128 (0.017)**	**0.428 (0.059)**	**0.335 (0.012)**
		scan2	0.085 (0.011)	**0.392 (0.062)**	0.459 (0.018)
		scanLR	0.065 (0.029)	**0.465 (0.246)**	0.695 (0.074)
OPT	20	clf+	0.33 (0.018)	**0.742 (0.041)**	0.225 (0.01)
		scan2	0.16 (0.057)	**0.734 (0.2)**	0.479 (0.065)
		scanLR	**0.45 (0.034)**	0.605 (0.055)	**0.134 (0.01)**
	24	clf+	0.274 (0.015)	**0.752 (0.034)**	0.275 (0.013)
		scan2	0.149 (0.031)	0.685 (0.096)	0.471 (0.094)
		scanLR	**0.693 (0.272)**	0.418 (0.146)	**0.076 (0.042)**

Table 1: Comparison of our weakly supervised scan methods unions: left- and right-tailed p-value scans (`scanLR`), and top-3 two-tailed p-value scans (`scan2`), and the supervised out-of-distribution classifier baseline (`clf+`) on auditing BERT and OPT. Performance and relative subset (Size) reported as mean (std) across 10 random test datasets with 10% anomalous data. Best (significant) bold.

and preprocessing, see Appendix A. We audit two pre-trained LLMs: the BERT base (uncased) model [16] with a 12-layer encoder with 768 nodes per layer, where focus on activations for the `[CLS]` token, and the Facebook OPT 6.7 model [56] with a 32-layer decoder with 4096 nodes per layer. Note that our method can audit any LLM that provides activations. For details, see Appendix B.

DeepScan Extensions and Baseline. We scan over left-, right-, and two-tailed p-values. Building upon prior work [3], we analyze the activations from layers closer to the output as they are suspected of encoding higher-level information. For further details, see Appendix C. We report results for the introduced scanning extensions: the outcomes derived from the combined subset of left and right p-value scans (`scanLR`), along with the results from the top-3 scan over the two-sided p-values (`scan2`). To evaluate our detection power, we compare to a supervised classifier (`clf+`) that aims to predict whether a sentence is true or false based on the LLM activation [3]. It is trained for each LLM layer on an approximately balanced dataset containing 1739 false sentences and is tested on an out-of-distribution held-out dataset (e.g., other topics) of the same task. For details, see Appendix D.

Results. We present results in Table 1. After identifying the subset, we assume access to test labels and report precision, and recall. Precision measures the ability to avoid false positives, calculated as the ratio of correctly identified anomalous samples to the total samples flagged as anomalous. Recall quantifies the ability to find all anomalous sentences, measured as the ratio of correctly identified anomalous samples to the total actual anomalous instances. We also report the size of the subgroup of sentences returned by the scanner or the number of sentences predicted as false by the classifier.

We first audit BERT. Across activations from both layer 10 and (last) layer 12, we observe low precision across all methods. These results indicate that BERT has a limited capacity to represent hallucinations within its internal state effectively, confirming prior work [3]. Subsequently, we audit OPT, a more potent model designed to match GPT's capabilities [56]. We consistently observe higher precision in detecting false statements across methods and test sets. From an auditing perspective, these findings indicate that OPT does indeed encode hallucination information within its internal state. This is consistent with prior research [3], which found layer 20 to be the most predictive for their classifier. Yet, our scanning approach (`scanLR`) excels at layer 24.

Comparing methods, `scan2` achieves similar precision levels to the classifier for BERT. At layer 10, the classifier shows highest recall while flagging $\sim 74\%$ of sentences as false, despite an expected rate of 10%, indicating a high False Positive Rate. In the case of OPT, `scanLR` exhibits higher precision than the supervised baseline for layer 20, with a subset size ($\sim 13\%$) closely matching the expected 10%. Nevertheless, `scan2` and `clf+` achieve the highest recall rates and return larger subsets for both layers. In summary, method effectiveness varies with test dataset and layers, without a clear dominant method. Rather, we observe a trade-off between precision and recall, with one method excelling in precision at the cost of recall, and vice versa. We also observe a strong connection between subset

size and recall, where larger subsets tend to yield higher recall but often at the cost of decreased precision. In conclusion, our method—with no prior exposure to false statements—exhibits similar performance to an out-of-distribution classifier trained on larger amounts of anomalous samples when both are assessed using a dataset comprising just 10% anomalous data (80 samples).

4 Summary and Discussion

We have introduced a weakly supervised auditing technique to identify, whether a pre-trained LLM is encoding patterns such as hallucinations within its internal states. We are interested in this problem because if an LLM is encoding these patterns internally, it can potentially impact downstream tasks and one may be able to deploy bias mitigation strategies. Our method employs subset scanning across various neural network layers in pre-trained LLMs, without the need for for prior knowledge of the specific patterns or access to labeled false statements.

During validation on a hallucination dataset, our approach achieved performance similar to, and sometimes surpassing a baseline fully supervised out-of-distribution classifier. Importantly, our approach only requires access to samples labeled as "normal" (true) eliminating the need for anomalous pattern data, which can be costly and ethically challenging to obtain (especially for other types of anomalies, such as toxicity and stereotypes). This makes our method highly suitable for real-world applications. Nonetheless, recent research [31] has raised doubts about the generalizability of prior methods [3, 10] in detecting hallucinations, which we intend to explore further. Our work makes assumptions about the background dataset, assuming it *only* contains "normal" statements tailored to the problem at hand. We plan to extend our work to more realistic assumptions by including small amounts of anomalous data in the reference dataset and composing it of various data sources.

Finally, similar to work using metrics such as cosine-similarity [32], our method only has positive predictive ability: it can be used to detect the presence of anomalies but not their absence. To understand how much of our experimental results can be attributed to our method's detection power and how much is due to the LLM not encoding the anomalous pattern, we have compared it to a supervised baseline. Our results indeed show that our method, which requires no training and no prior exposure to false statements, performs comparably to the supervised baseline.

5 Outlook: Informing Fine-tuning

We briefly discuss the potential expansion of our work. Unsupervised pre-training and task-specific fine-tuning have become the standard approach for various LLM tasks [19, 23, 50]. Despite their impressive achievements, these methods face challenges in generalization performance on downstream tasks [25, 30, 44] and suffer from catastrophic forgetting [1, 34, 54]. To address these issues, sub-network optimization approaches have emerged as a promising method to enhance stability and reduce overfitting without requiring full network retraining [51, 54].

We believe that our method can build on the advancement of this line of research for bias mitigation strategies. As detailed in § 2, our method allows identifying the subset of nodes $O_{S^\star}$ that are most responsible for the anomalous patterns found. This means, for each layer, we are able to identify the nodes that align with the most anomalous subset of sentences, that is those nodes for which the empirical p-values of their activations deviate from the uniform distribution such that they are flagged anomalous. This suggests that these nodes are pivotal in identifying anomalous patterns within the data, which could guide the efficient fine-tuning of sub-networks for bias mitigation strategies. We show initial findings in Appendix E.

Acknowledgments

The authors thank Edward McFowland III for helpful feedback and discussions.

References

[1] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021.

[2] Victor Akinwande, Celia Cintas, Skyler Speakman, and Srihari Sridharan. Identifying audio adversarial examples via anomalous pattern detection. *arXiv preprint arXiv:2002.05463*, 2020.

[3] Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*, 2023.

[4] Christine Basta, Marta R Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, 2019.

[5] Christine Basta, Marta R Costa-Jussa, and Noe Casas. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384, 2021.

[6] Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

[7] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[9] Shikha Bordia and Samuel Bowman. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, 2019.

[10] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[11] Feng Chen and Daniel B Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1166–1175, 2014.

[12] Celia Cintas, Payel Das, Brian Quanz, Girmaw Abebe Tadesse, Skyler Speakman, and Pin-Yu Chen. Towards creativity characterization of generative models via group-based subset scanning. In *International Joint Conference on Artificial Intelligence*, 2022.

[13] Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, and Edward McFowland. Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 876–882, 2021.

[14] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

[15] Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building stereotype repositories with llms and community engagement for scale and depth. *Cross-Cultural Considerations in NLP@ EACL*, page 84, 2023.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[17] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. 2004.

[18] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020.

[19] Muhammad Usman Hadi, R Qureshi, A Shah, M Irfan, A Zafar, MB Shaikh, N Akhtar, J Wu, and S Mirjalili. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*, 2023.

[20] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[21] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2018.

[22] Alexander Henlein and Alexander Mehler. What do toothbrushes do in the kitchen? how transformers think our world is structured. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5791–5807, 2022.

[23] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023.

[24] Mohsen Jamali, Ziv M Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. *arXiv preprint arXiv:2309.01660*, 2023.

[25] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[26] Hannah Kim, Celia Cintas, Girmaw Abebe Tadesse, and Skyler Speakman. Spatially constrained adversarial attack detection and localization in the representation space of optical flow networks. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 965–973. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

[27] Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush Varshney. Out-of-distribution detection in dermatology using input perturbation and subset scanning. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

[28] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for bert sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, 2021.

[29] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, 2019.

[30] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, 2020.

[31] BA Levinstein and Daniel A Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *arXiv preprint arXiv:2307.00175*, 2023.

[32] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, 2020.

[33] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[34] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. Variational information bottleneck for effective low-resource fine-tuning. *arXiv preprint arXiv:2106.05469*, 2021.

[35] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, 2019.

[36] Edward McFowland, Skyler Speakman, and Daniel B Neill. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1):1533–1561, 2013.

[37] Edward McFowland III, Sriram Somanchi, and Daniel B Neill. Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv preprint arXiv:1803.09159*, 2018.

[38] Milagros Miceli and Julian Posada. The data-production dispositif. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–37, 2022.

[39] Duncan J Murdoch, Yu-Ling Tsai, and James Adcock. P-values are random variables. *The American Statistician*, 62(3):242–245, 2008.

[40] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, 2021.

[41] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 1953–1967. Association for Computational Linguistics (ACL), 2020.

[42] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 554–565, 2023.

[43] Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012.

[44] Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

[45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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

[47] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*. Association for Computational Linguistics, 2019.

[48] John A Rice. *Mathematical statistics and data analysis*. Thomson Brooks/Cole, 2007.

[49] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32, 2019.

[50] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, pages 1–11, 2023.

[51] Shoujie Tong, Heming Xia, Damai Dai, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. Bi-drop: Generalizable fine-tuning for pre-trained language models via adaptive subnetwork optimization. *arXiv preprint arXiv:2305.14760*, 2023.

[52] PA Utama, NS Moosavi, and I Gurevych. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 7597–7610. Association for Computational Linguistics, 2020.

[53] Andrew Wang, Mohit Sudhakar, and Yangfeng Ji. Simple text detoxification by identifying a linear toxic subspace in language model embeddings. *arXiv preprint arXiv:2112.08346*, 2021.

[54] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, 2021.

[55] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*, 2023.

[56] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[57] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.

[58] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 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, 2019.

[59] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*, pages 12–2, 2023.