

Model Attribution in Machine-Generated Disinformation: A Domain Generalization Approach with Supervised Contrastive Learning

Abstract—Model attribution for machine-generated disinformation poses a significant challenge in understanding its origins and mitigating its spread. This task is especially challenging because modern large language models (LLMs) produce disinformation with human-like quality. Additionally, the diversity in prompting methods used to generate disinformation complicates accurate source attribution. These methods introduce domain-specific features that can mask the fundamental characteristics of the models. In this paper, we introduce the concept of model attribution as a domain generalization problem, where each prompting method represents a unique domain. We argue that an effective attribution model must be invariant to these domain-specific features. It should also be proficient in identifying the originating models across all scenarios, reflecting real-world detection challenges. To address this, we introduce a novel approach based on Supervised Contrastive Learning. This method is designed to enhance the model’s robustness to variations in prompts and focuses on distinguishing between different source LLMs. We evaluate our model through rigorous experiments involving three common prompting methods: “open-ended”, “rewriting”, and “paraphrasing”, and three advanced LLMs: “llama 2”, “chatgpt”, and “vicuna”. Our results demonstrate the effectiveness of our approach in model attribution tasks, achieving state-of-the-art performance across diverse and unseen datasets.

Index Terms—Natural Language Processing, Text Classification, Contrastive Learning, Fake News, Large Language Models

I. INTRODUCTION

The rapid advancement of artificial intelligence has compromised the accuracy of information. This results in the rise of machine-generated disinformation, which significantly challenges societal trust and information integrity [1]. Disinformation refers to false information that is intentionally created and spread with the purpose of deceiving or misleading audiences. A significant challenge involves tracing disinformation back to its originating models, especially when these models are capable of generating high-quality textual content that can compete with human authorship [2].

The task of attributing machine-generated disinformation to its source model is complicated by two primary factors. Firstly, modern Large Language Models (LLMs) like GPT-4 and LLaMA-2 can produce text that often rivals human-written content in quality [3]. This high quality makes it challenging to clearly identify and attribute their outputs. Secondly, and more significantly, the variety of prompting methods, spanning from open-ended to specific instructions

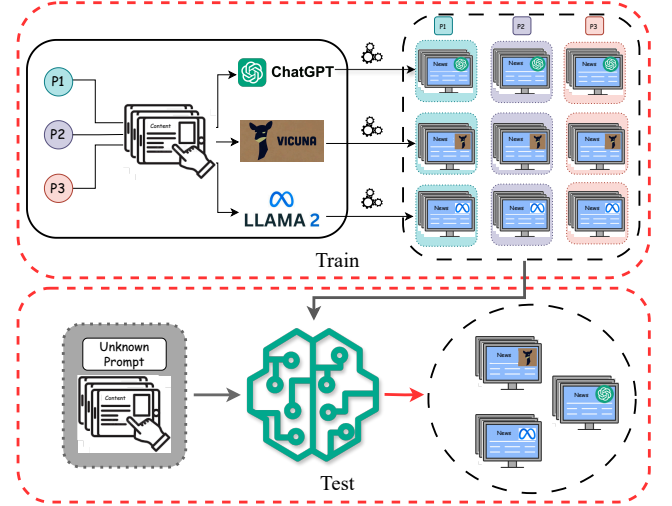


Fig. 1: A schematic diagram of model attribution in machine-generated disinformation.

such as rewriting or paraphrasing, adds further complexity. These methods influence the stylistic and substantive features of the generated content. They complicate the task of model attribution through textual analysis alone. Additionally, these models enable the rapid creation and dissemination of large volumes of disinformation, which can easily surpass the capabilities of traditional detection systems.

Recognizing these challenges, we formulate the LLM attribution as a domain generalization problem. Each prompting method introduces domain-specific features that an ideal attribution model should be robust against. This approach reflects real-world scenarios where the source of disinformation must be identified regardless of the method used to generate it.

To address this problem, we propose a Supervised Contrastive Learning (SCL) for text classification approach that considers each prompting method as a unique domain with specific features (Figure 1). SCL optimizes the model by pulling together representations of instances from the same class and pushing apart those from different classes within the embedding space. This method is essential for extracting domain-invariant features, thereby refining the decision boundaries of our classifier and enhancing its accuracy and generalizability across varied disinformation scenarios. The key innovation of our methodology is creating a model invariant to these domain-specific features while focusing on the underlying characteristics that define a specific LLM. This

invariance is crucial for ensuring that the attribution model performs consistently across different types of prompts.

In this work, we consider three typical prompting methods - “open-ended”, “rewriting”, and “paraphrasing” - and evaluate them on three state-of-the-art LLMs: LLaMA-2, ChatGPT, and Vicuna. The results show the effectiveness of our model in discerning the outputs of these LLMs across diverse prompting conditions. Furthermore, we extend our investigation to case studies involving ChatGPT-4, exploring its in-context learning capabilities for model attribution. These studies show the potential of ChatGPT-4 to recognize and adapt to the writing styles characteristic of specific LLMs. However, they also reveal the inherent limitations of this strategy, such as its success being dependent on the quality and relevance of the training examples provided during the learning process.

The key contributions of this work are outlined as follows:

- Introduces the machine-generated disinformation attribution as a domain generalization problem, treating each prompting method as a distinct domain.
- Develops a new method that enhances model robustness against varied prompting styles, focusing on identifying source LLMs based on their textual signatures.
- Validates the model across diverse prompting methods: “open-ended”, “rewriting”, and “paraphrasing” and three LLMs: ChatGPT, LLaMA-2, and Vicuna, demonstrating robust performance.
- Explores ChatGPT-4 in-context learning for source detection, assessing its strengths and limitations in practical scenarios.

II. RELATED WORK

A. Domain Adaptation

Domain adaptation (DA) aims to minimize the distributional discrepancy between the source and target domains, enabling a model trained on a labeled source domain to perform well on a different, unlabeled target domain. Traditional approaches often utilize kernel methods such as Maximum Mean Discrepancy (MMD) and spectral feature alignment to achieve feature alignment across domains [4], [5]. MMD measures the difference between two probability distributions based on their means in a reproducing kernel Hilbert space, helping to reduce the domain gap [6]. Spectral feature alignment, on the other hand, aligns the source and target domain features by using a shared latent space [5].

Domain adversarial training has been another prevalent approach, wherein a model learns to extract domain-invariant features via adversarial objectives. This approach was first introduced by Ganin et al. (2016) and has been extended by various works to improve its robustness and effectiveness [7]–[9]. The idea is to train a feature extractor that minimizes the classification loss while maximizing the domain classification loss, thus promoting domain-invariant feature learning.

Self-training methods, which involve generating pseudo-labels for target domain data and iteratively refining the model, have also shown efficacy in DA tasks [10]. These methods

start by training a model on the source domain and then using the trained model to assign labels to the target domain data [11]. The model is then retrained with these pseudo-labels, gradually adapting to the target domain.

Multi-source domain adaptation (MSDA) extends DA to multiple source domains, aiming to leverage data from several domains to improve performance on a target domain. Methods like pairwise-MMD and adversarial autoencoders have been employed to handle domain discrepancies in MSDA settings [12], [13]. Pairwise-MMD calculates the discrepancy between each pair of the source domain and target domain, while adversarial autoencoders learn a shared latent space that minimizes domain differences. However, a critical limitation of these approaches is their reliance on unlabeled target domain data during training, which is not always feasible in real-world scenarios.

B. Domain Generalization

Domain generalization (DG) addresses the challenge of training models that generalize well to unseen target domains without access to any target domain data during training. This is particularly significant when obtaining labeled or unlabeled data from the target domain is impractical. Approaches in DG often focus on learning domain-invariant representations from multiple source domains [14], [15].

One popular strategy for DG is invariant risk minimization (IRM), which aims to learn representations that are invariant across different domains by capturing the intrinsic relationships between features and labels [16]. IRM enforces the optimal classifier to remain invariant across different domains, ensuring that the learned representations generalize well to new, unseen domains.

Another approach is the use of domain adversarial training, extended to handle multiple source domains, to learn generalized feature spaces [7]. This involves training a domain discriminator that distinguishes between different source domains while the feature extractor learns to confuse the discriminator, leading to domain-invariant features.

The advent of Supervised Contrastive Learning with memory-saving mechanisms has shown promise in improving DG performance by enhancing the alignment of features within the same class across domains [17]. Supervised Contrastive Learning (SCL) explicitly pulls together representations of the same class while pushing apart those of different classes, thereby improving the discriminative power of the learned features. By incorporating a memory bank, the approach can handle larger sets of contrasting examples without excessive memory consumption, leading to better generalization.

C. Contrastive Learning

Contrastive learning has emerged as a powerful technique for self-supervised and supervised learning, wherein the model learns to distinguish between similar and dissimilar examples. In the context of self-supervised learning, contrastive methods such as SimCLR and MoCo maximize the agreement between

different augmentations of the same instance while minimizing the agreement between different instances [18], [19]. These methods typically rely on large batch sizes to ensure a diverse set of negative examples, although techniques like memory banks have been proposed to mitigate the associated computational costs.

SimCLR, for instance, uses stochastic data augmentations and learns representations by maximizing agreement between different augmented views of the same data point using a contrastive loss in the latent space [18]. MoCo introduces a dynamic dictionary with a queue and a moving-averaged encoder to maintain a large set of negative examples efficiently [19].

Supervised Contrastive Learning (SCL) extends this idea to labeled data by treating examples from the same class as positive pairs and those from different classes as negative pairs [20]. This approach has been successfully applied to various tasks, including image classification and natural language understanding, by improving the discriminative power of learned representations [21]. Another approach such as SimCSE, a simple contrastive learning framework that incorporates annotated pairs called entailment pairs as positives and contradiction pairs as hard negatives, helped refine textual embeddings, demonstrating the adaptability of contrastive learning in a semantic setting [22], [23]. These negative samples were refined both in definition and in usage through CLINE, a method for generating both negative adversarial and contrastive examples to improve robustness under adversarial attacking, and a user-specified method to generate varying degrees of "hardness" of negative examples, highlighting the method's ability to enhance performance in downstream tasks across multiple modalities [24], [25]. In SCL, the model leverages label information to explicitly cluster representations of the same class and separate those of different classes, which helps in learning more robust features.

In the context of DG, SCL can be particularly effective as it explicitly encourages the model to learn domain-invariant features by pulling together representations of the same class from different domains and pushing apart those of different classes [17]. This is achieved by using a combination of contrastive loss and cross-entropy loss, ensuring that the learned features are both discriminative and generalizable across domains. Additionally, the use of a memory bank allows for the storage and reuse of representations from previous batches, enhancing the effectiveness of contrastive learning without incurring high memory costs [17].

D. Large Language Models

LLMs [26]–[28] have proved themselves quite remarkably versatile in various applications of text generation [29]. One such field is the ability to mimic human writing and efficiently generate deceptive fake news with minimal prompting [30]. Recent studies have attempted to address LLM-generated disinformation further including artificially generating robust propaganda with which to train [1].

Pertaining to the new flood of LLM-generated disinformation, whether malicious or accidental, detection models [31] and LLM-attribution methods [32], [33] have risen in retaliation. One recent detection strategy takes into account style-related features of text to combat style-based attacks [34]. This approach, although promising in regards to tackling currently prevalent prompting engineering attacks, is not as suited to address future prompt attacks without the need to retrain models with constantly improving tactics. In this work, we focus on introducing a method employing domain generalization over various prompting techniques to diminish the need to retrain the LLM detector.

III. METHODOLOGY

A. Problem Definition

In the context of domain generalization for model attribution, we have labeled data from k source domains, where each domain represents a prompting method. Each source domain $\{D_{s_i}\}_{i=1}^k$ is denoted as $D_{s_i} = \{X_{s_i}, Y_{s_i}\}$, where X_{s_i} is LLM-generated disinformation and Y_{s_i} is the LLM that generated it. The goal is to create a model that can generalize across different domains, ensuring it can accurately identify the source of disinformation regardless of the prompting method.

During training, only source domains are available, and the labeled dataset from the target domain $D_t = \{X_t, Y_t\}$ is reserved for evaluation. Unlike multi-source domain adaptation (MSDA), domain generalization does not require additional unlabeled target domain data during training [35]–[39]. Instead, the model is trained to obtain domain-invariant features from the given source domains. Consequently, the target domain data D_t is used exclusively for evaluation, allowing the trained model to make predictions on unseen domains.

B. Model Architecture

To achieve this, we employ a Supervised Contrastive Learning (SCL) approach. This method not only increases the margin of decision boundaries but also ensures uniform distribution within each class, thus minimizing domain discrepancies.

We start by shuffling all the source domain data and dividing the combined dataset into mini-batches. This ensures a stable domain distribution in each mini-batch. For a sampled mini-batch of size N , denoted as $S = \{x_i, y_i\}_{i=1}^N$, x_i and y_i represent the input disinformation and the LLM label respectively. We use a pre-trained language model (PLM) as the encoder and take the hidden state of the last layer's [CLS] token as the document representation, denoted as h . Next, the representations h are then fed into a feed-forward neural network f with a tanh activation function and layer normalization to produce reduced-dimension features $z = f(h)$. These features are subsequently used by a classifier g for downstream tasks, and the cross-entropy loss L_{CE} is computed as follows:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N y_i \log(g(z_i)) \quad (1)$$

where g is a fully connected classifier.

Since domain generalization involves training an algorithm using multiple source domains, the main challenge is to learn an optimal joint hypothesis for these domains. The objective is to separate data points by their labels, thereby minimizing domain discrepancy within each class in the feature embedding space. Supervised Contrastive Learning (SCL) explicitly increases the margin of decision boundaries and promotes a uniform distribution within each class while also reducing the distances between source domains for each class [20].

This strategy aligns perfectly with our aim to develop a unified hypothesis across various source domains. It allows the model to maintain a more consistent distribution for each label class. Specifically, within a given mini-batch S of size N , the Supervised Contrastive Loss L_{SCL} is computed as follows:

$$L_{SCL} = - \sum_{z_i \in S} \frac{1}{|P(i)|} \sum_{z_p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{z_a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (2)$$

where $P(i)$ represents the set of positive examples for z_i , $A(i)$ is the combined set of positive examples (e.g., disinformation generated by the same LLM) and negative examples (e.g., disinformation generated by different LLMs), and τ is a scaling hyper-parameter (temperature).

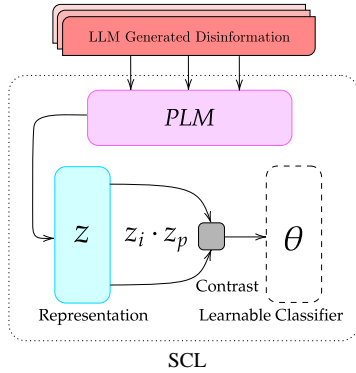


Fig. 2: SCL Architecture.

Then, the combined loss function for training the model is:

$$L = L_{CE} + L_{SCL} \quad (3)$$

By employing this approach, the model can better align features of the same class and reduce domain discrepancies, leading to improved performance in identifying the source of disinformation across unseen domains. We provide the pseudocode for our approach in Algorithm 1

IV. EXPERIMENT

A. Dataset

To test the performance of our model, we conducted experiments on the LLMFake [1] dataset, a benchmark focused on disinformation generated by LLMs. LLMFake employs three

Algorithm 1 Algorithm for Supervised Contrastive Learning

```

1: Input: Batch size  $N$ , encoder  $f$ , classifier  $g$ 
2: for  $t \leq T_{\max}$  do
3:   Sample minibatch  $S = \{x_i, y_i\}_{i=1}^N$ 
4:    $z = f(\text{PLM}(x))$ 
5:    $\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=1}^N -y_i \log(g(z_i))$ 
6:    $P(i) \equiv \{z_j \in S, y_j = y_i\}$ 
7:    $A(i) \equiv \{z_j \in S, j \neq i\}$ 
8:    $\mathcal{L}_{SCL} = \sum_{z_i \in S} -\frac{1}{N} \sum_{z_p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{z_a \in A(i)} \exp(z_i \cdot z_a / \tau)}$ 
9:    $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SCL}$ 
10:  update network by combined loss  $\mathcal{L}$ 
11: end for

```

disinformation generation methods: Hallucinated News, Arbitrary, and Controllable Generation. Hallucinated and Arbitrary methods prompt models like ChatGPT to produce fake news. Controllable Generation uses models like ChatGPT¹, Llama2 [27], and Vicuna [40] to create disinformation from nonfactual content in the Politifact, Gossipcop [41], and CoAID [42] datasets through prompt techniques such as Paraphrasing, Rewriting, and Open-ended Generation. The Controllable Generation method results in approximately 5,400 samples, equally distributed across three domains: Paraphrasing (P), Rewriting (R), and Open-ended Generation (O). From each domain, there are approximately 600 LLM-generated disinformation samples generated by each of the three LLMs.

B. Experiment Setting

We conducted experiments in a cross-domain setting. The model is trained on one or two source domains and evaluated on the remaining two or one target domains, respectively (e.g., P, R-O or P-R, O). Model selection is based on validation performance on the combined test set of the source domains. For a fair comparison, we filter each domain based on semantic similarity, as the prompting methods are designed to preserve semantics and only change the style [1]. After calculating the average semantic similarity, we retain only the disinformation samples with a similarity score above the average. Additionally, we split each domain into 600 training examples, 200 validation examples, and 200 test examples. Since we do not have access to the target domain data, the training and model selection are based on the mixture distribution of source domains.

We use BERT [43] and RoBERTa [44] as the default PLMs for encoders. For all of the encoders, we set the classifier representation dimension z to 256. All models are trained with the Adam [45] optimizer at a learning rate of $1e^{-5}$. The whole model is trained up to 10 epochs with a batch size of 16 for training the BERT/RoBERTa baselines. All experiments are conducted on Nvidia 40GB V100 GPUs. We conduct a grid search for $\tau \in \{0.1, 0.2, 0.5, 0.7, 0.8, 1.0\}$ in the P, R-O transfer (P and R as source domains, O as the target domain) and select $\tau = 0.2$.

¹gpt-3.5-turbo: <https://platform.openai.com/docs/models/gpt-3-5>

C. Evaluation Metrics

In this section, we discuss the evaluation metrics used to compare the performance of the models. Taking fine-tuned BERT-base as the baseline model, we calculate the actual difference in accuracy (act_diff) and the improvement percentage (imp) from the target model to BERT-base for a given dataset as follows:

$$act_diff = acc_tar - acc_bert \quad (4)$$

$$imp = \frac{acc_tar - acc_bert}{acc_bert} \times 100\% \quad (5)$$

where acc_tar indicates the accuracy of the target model and acc_bert indicates the accuracy of the BERT-base model. We aggregate the actual differences and the improvement percentages across all test datasets to evaluate each model’s performance. Additionally, we calculate the average accuracy across all test datasets for the target model. These results are included in the table.

D. Comparison Models

We compared our model to state-of-the-art models that were known to perform well on out-of-distribution datasets. The following models were used for comparison:

Fine-tuned Language Models (LMs) adapting pre-trained LMs to the model attribution task, and has proven effective in handling disinformation scenarios [46]. We fine-tune BERT [43] and DeBERTa [47] on the labeled data from source domains and directly test on the target domain. BERT is chosen for its robust performance in various NLP tasks and its ability to capture contextual information through bidirectional encoding [43], while DeBERTa is selected for its enhanced ability to handle long-range dependencies and improved mask decoding strategies [47], making it effective in complex disinformation detection scenarios.

Fine-tuning Setup We fine-tune each pre-trained model on a dataset using cross-entropy loss. This approach is commonly used for model fine-tuning. To comprehensively evaluate the generalization ability of the pre-trained models, we conduct two types of experiments: 1) **full fine-tuning** and 2) **probing**. In the full fine-tuning experiment, we fine-tune each pre-trained language model (encoder) along with a classifier. While it is standard practice to fine-tune both the encoder and the classifier, this process often results in catastrophic forgetting [48], which can obscure direct comparisons between pre-trained models. To address this, we also conduct probing experiments where we freeze the encoder (i.e., the pre-trained models) and train only the classifier with a single linear layer. This method allows for a more direct comparison of the pre-trained representations, similar to the approach in [49].

E. Experiment Results

We conducted extensive experiments to evaluate the performance of our proposed model, SCL_{BERT}, compared to BERT and DeBERTa baselines using the FakeLLM dataset. The

results, presented in Table I, provide a comprehensive comparison of in-domain and out-of-domain (OOD) performance.

In-Domain Performance. In the full fine-tuning setting, BERT achieved accuracies of 58.18% on domain P, 72.05% on domain R, and 79.40% on domain O. DeBERTa showed better performance with accuracies of 62.59% on P, 76.87% on R, and 82.69% on O. DeBERTa’s superior performance can be attributed to several advancements in its architecture and training methods. It employs a disentangled attention mechanism and relative positional encodings, enhancing its ability to understand context and word relationships. Additionally, DeBERTa is more parameter-efficient, achieving better results with fewer parameters. It also benefits from improved training techniques and larger, more recent pre-training datasets, allowing it to capture contextual information more effectively. These improvements collectively result in DeBERTa’s superior accuracy compared to BERT. However, in SCL_{BERT}, we observe a slight drop in in-domain performance during domain generalization, which is often a necessary trade-off for better generalization to unseen domains. This trade-off results from regularization, learning diverse and robust features, and balancing performance across multiple domains. Consequently, the model sacrifices some specialization to become more versatile and robust across different domains.

In the probing setting, all models’ performance slightly decreased. This drop occurs because freezing the layers of LMs during fine-tuning prevents the model from adapting its pre-trained features to the specific requirements of the new task. The fixed, general representations from pre-training may not capture all the necessary information for the new task. This restriction limits the model’s ability to learn new patterns or adjust to new data distributions. As a result, the classifier layer alone cannot compensate for the lack of fine-tuning in the deeper layers, leading to lower accuracy.

Out-of-Domain Performance. In BERT and DeBERTa, the OOD performance dropped significantly. It shows the model struggles to adapt its pre-trained features to new, unseen domains. The PLMs may not capture the unique linguistic patterns and contexts of the out-of-domain data. Additionally, diverse and unseen patterns in OOD datasets reveal the models’ difficulty in generalizing, as they may have overfitted to the training domain. This overfitting means they perform well on familiar data but poorly in different contexts. This highlights the importance of effective domain generalization to maintain robustness across various tasks and domains.

In both full fine-tuning and probing settings, SCL_{BERT} shows substantial improvements in OOD performance. On average, performance improved by more than 7% in full fine-tuning and 9% in probing settings. This is because the clusters become more concentrated and domain discrepancies are significantly reduced. The performance boost indicates that increasing the margin of decision boundaries, promoting a uniform distribution within each class, and reducing distances between source domains by SCL_{BERT} are effective strategies.

The experimental results clearly demonstrate that SCL_{BERT} outperforms both BERT and DeBERTa baselines in the task

Method	Domain			Full					Probing				
				In Domain	Out of Domain			OOD Avg	In Domain	Out of Domain			OOD Avg
	P	R	O		P	R	O			P	R	O	
BERT	×	×	✓	79.40	26.75	53.69	-	41.01	71.39	22.98	51.44	-	38.03
	×	✓	×	72.05	51.42	-	56.47	54.20	68.15	49.37	-	55.13	52.11
	✓	×	×	58.18	-	70.55	42.46	55.91	52.73	-	65.47	38.14	52.32
	×	✓	✓	75.44	54.76	-	-	54.76	71.12	51.59	-	-	51.59
	✓	×	✓	62.72	-	68.93	-	68.93	57.62	-	62.54	-	62.54
	✓	✓	×	65.70	-	-	54.14	54.14	59.83	-	-	51.32	51.32
Upper-bound	✓	✓	✓	69.93	-	-	-	-	61.82	-	-	-	-
DeBERTa	×	×	✓	82.69	45.32	42.37	-	43.76	71.21	42.89	40.06	-	41.71
	×	✓	×	76.87	18.70	-	29.29	24.53	69.44	22.53	-	31.72	27.12
	✓	×	×	62.59	-	64.89	37.47	50.60	58.89	-	59.11	39.56	50.13
	×	✓	✓	79.81	18.70	-	-	18.70	72.83	33.57	-	-	33.57
	✓	×	✓	73.94	-	78.86	-	78.86	64.72	-	70.12	-	70.12
	✓	✓	×	75.73	-	-	38.11	38.11	69.23	-	-	42.91	42.91
Upper-bound	✓	✓	✓	75.87	-	-	-	-	67.73	-	-	-	-
SCL _{BERT}	×	×	✓	75.26	38.85	55.34	-	47.11	67.91	33.12	54.23	-	44.05
				-4.11	+12.1	+1.65	-	+6.10	-3.48	+10.14	+2.79	-	+6.02
				-5.17%	+45.23%	+3.07%	-	+14.87%	-4.87%	+44.12%	+5.42%	-	+15.82%
	×	✓	×	69.89	54.52	-	58.34	56.81	66.32	55.12	-	56.89	55.34
				-2.16	+3.10	-	+1.87	+2.61	-1.83	+5.75	-	+1.76	+3.23
				-2.99%	+6.02%	-	+3.31%	+4.81%	-2.68%	+11.64%	-	+3.19%	+6.19%
	✓	×	×	56.34	-	71.25	46.71	56.83	50.13	-	68.83	45.12	53.82
				-1.84	-	+0.7	+4.25	+0.92	-2.60	-	+3.36	+6.98	+1.50
				-3.16%	-	+0.99%	+10.00%	+1.64%	-4.93%	-	+5.13%	+18.30%	+2.86%
	×	✓	✓	72.91	59.13	-	-	59.13	67.29	55.71	-	-	55.71
				-2.53	+4.37	-	-	+4.37	-3.83	+4.12	-	-	+4.12
				-3.35%	+7.98%	-	-	+7.98%	-5.38%	+7.98%	-	-	+7.98%
	✓	×	✓	61.45	-	70.78	-	70.78	54.91	-	68.23	-	68.23
				-1.27	-	+1.85	-	+1.85	-2.71	-	+5.69	-	+5.69
				-2.02%	-	+2.68%	-	+2.68%	-4.70%	-	+9.09%	-	+9.09%
	✓	✓	×	69.34	-	-	59.70	59.70	63.11	-	-	57.52	57.52
				+3.64	-	-	+5.56	+5.56	+3.28	-	-	+6.20	+6.20
				+5.54%	-	-	+10.26%	+10.26%	+5.48%	-	-	+12.08%	+12.08%
Upper-bound	✓	✓	✓	72.61	-	-	-	-	65.21	-	-	-	-
				+2.68	-	-	-	-	+3.39	-	-	-	-
				+3.83%	-	-	-	-	+5.48%	-	-	-	-

TABLE I: Result of the models trained on FakeLLM dataset. In the full fine-tuning setup (denoted as Full), we train both the encoder and the classifier. In the probing setup (denoted as Probing), we freeze the encoder and train the classifier. For the proposed model, each domain setting includes three rows: ACCURACY (FIRST ROW), ACTUAL DIFFERENCE (SECOND ROW), and IMPROVEMENT PERCENTAGE (THIRD ROW). These results are compared against the BERT baseline model. The reported results are averaged across 3 runs for each setting.

of model attribution for machine-generated disinformation. Notably, SCL_{BERT} excels in out-of-domain scenarios, underscoring its robustness and generalizability across varied prompting methods and unseen datasets. These findings validate the efficacy of our proposed approach in enhancing the detection and attribution capabilities of models tasked with identifying the sources of machine-generated disinformation. The consistent performance across both full fine-tuning and probing settings further highlights the strength of SCL_{BERT} in maintaining robust and effective representations.

F. Analysis on Domain Divergence

To further analyze our model’s performance, we provide a detailed visualization of domain discrepancies using t-SNE

[50]. This technique helps us understand how our models handle variations across different domains. As illustrated in Figure 3a, disinformation generated by the three methods exhibits significant overlap in the latent space, particularly with human-written disinformation. This overlap suggests that these methods produce content that is difficult to distinguish from each other and from human-authored disinformation. In Figure 3, the models are trained with O and R as source domains and P as the target domain. For the visualization, we sampled 500 examples from each domain. For the BERT model (Fig. 3b), clear domain discrepancies are observed within each LLM cluster, and the clusters are relatively dispersed. This indicates that BERT struggles to produce domain-

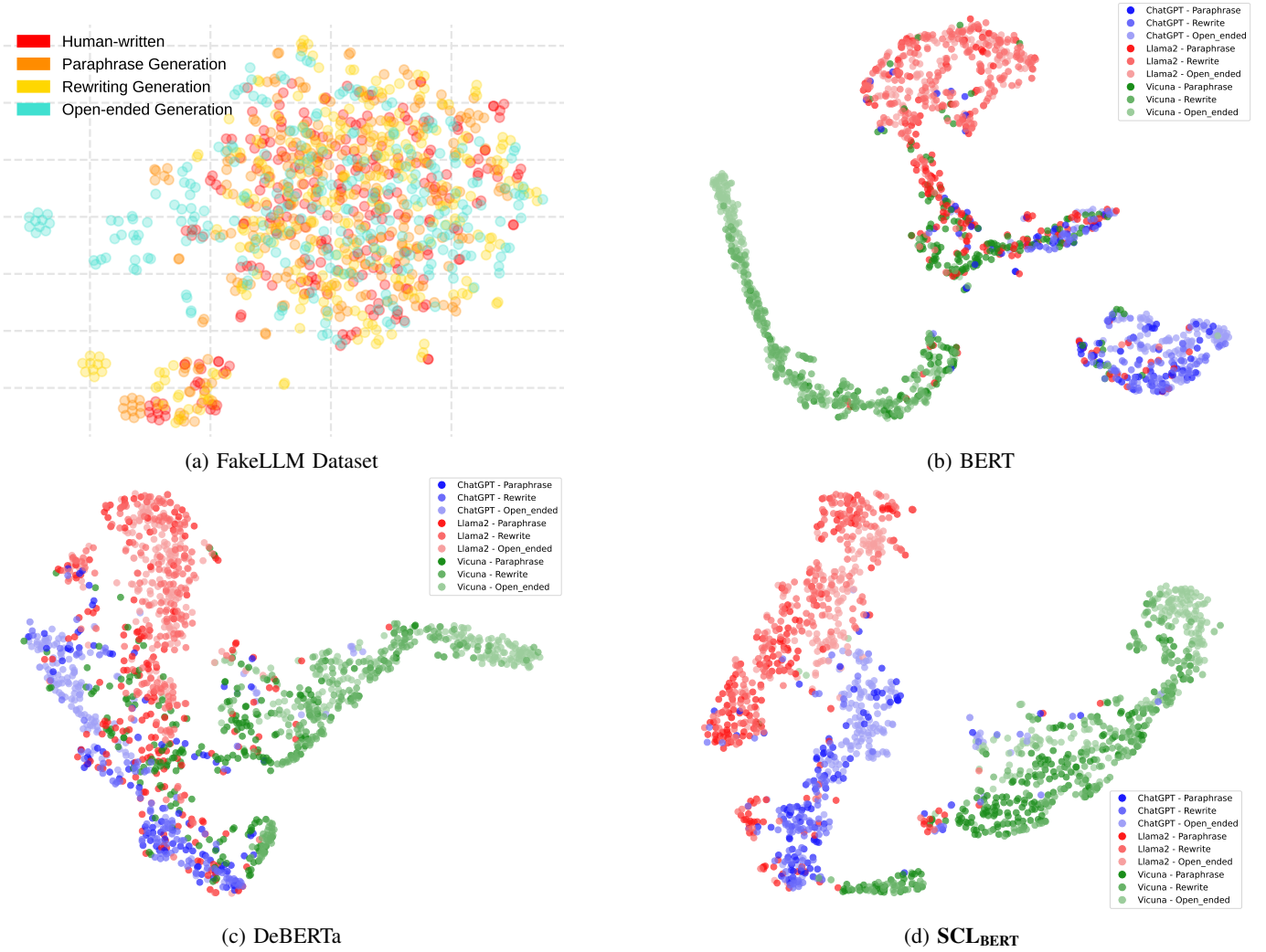


Fig. 3: T-SNE visualization of 256-dimensional sentence embeddings z for each model. The models are trained using O and R as source domains, and P as the target domain. For the visualization, 500 examples are sampled from each domain.

invariant representations. Similarly, the DeBERTa model (Fig. 3c) also shows domain discrepancies within the LLM clusters, demonstrating that DeBERTa does not significantly improve the ability to create domain-invariant features. In contrast, the SCL model (Fig. 3d) presents more concentrated LLM clusters with significantly reduced domain discrepancies. This suggests that SCL_{BERT} is more effective in producing domain-invariant representations, although some heterogeneity within the clusters remains. This visualization highlights the differences in performance between the models and their effectiveness in handling domain discrepancies. These visualizations confirm the superiority of SCL_{BERT} in domain generalization and also underline the necessity of advanced techniques to improve model robustness across diverse domains.

G. Case study

In this section, we explore the capabilities of ChatGPT 4 [26] in identifying the origins of machine-generated disinformation through In-Context Learning (ICL). Specifically, we

investigate whether ChatGPT 4 can effectively identify the source of disinformation when provided with example texts that illustrate the distinctive styles and patterns of LLMs. ChatGPT 4 has shown potential in recognizing the writing styles of LLM-generated disinformation by utilizing in-context learning. In-context learning involves providing ChatGPT 4 with a set of examples that illustrate the specific styles and patterns of different LLMs. These examples serve as a reference for the model to understand and identify stylistic nuances (Fig. 4). During the study, we provided ChatGPT 4 with samples of disinformation generated by various LLMs, such as ChatGPT, Llama2, and Vicuna. The goal was to determine if ChatGPT 4 could detect the origin of these samples based on their writing style. ChatGPT 4 was able to correctly capture stylistic elements such as tone, word choice, sentence structure, and formatting in most cases (Fig. 4a). However, the success of this approach is heavily dependent on the quality and relevance of the provided examples. If the examples do not adequately represent the distinct styles of each LLM, ChatGPT

Prompt: Analyze the provided text examples to determine whether ChatGPT or Llama2 generated the misinformation statement. For LLM, **two examples** are given. Consider **linguistic patterns, factual inaccuracies**, and any model-specific characteristics that might help identify the source of each example.

Determine whether ChatGPT or Llama2 generated the following misinformation statement the misinformation statement:

Misinformation Statement: The Centers for Disease Control and Prevention (CDC) has issued a recommendation regarding facial hair for preventing the novel coronavirus. **CNN reports** that... According to a 2017 infographic from the CDC, styles such as side whiskers, soul patches, lampshades, and handlebar mustaches are considered ...

Llama2 - Example 1: If we hope to shield ourselves from the coronavirus, we may need to reconsider our approach to climate change mitigation efforts. For years, left-leaning individuals have attributed various natural disasters and societal issues to global warming ... the coronavirus may thrive in hot temperatures. If this theory holds true, we may need to abandon our efforts to combat climate change in order to protect ourselves from the virus.

ChatGPT - Example 1: A widely circulated article on social media advises against healthy individuals using face masks during the COVID-19 pandemic, claiming that there are potential risks associated with this practice... Dr. Russell Blaylock cautions that not only do face masks fail to safeguard healthy individuals from falling ill, but they can also pose significant health risks to the wearer...

Llama2 - Example 2: ... she was told she had legionnaires disease... according to the caller, the doctors said that the disease was likely contracted through the reusable face masks that she was wearing... the caller was very concerned that many people wearing reusable masks may be at risk of contracting legionnaires disease... Reusable face masks are claimed to be a potential carrier of Legionnaires' disease, a pneumonia type, according to numerous social media posts...

ChatGPT - Example 2: According to Donald Trump, he has been taking hydroxychloroquine for the past few weeks, a medication that many frontline workers are also taking. Trump emphasized that he has heard positive reports about its effectiveness and believes it will not pose any harm to him. The full transcript can be accessed for further details.

Llama2 Characteristics:

Tone and Content: The examples from Llama2 are characterized by a somewhat sensationalist tone and include controversial, speculative statements.

Complexity and Structure: These examples contain complex sentences and detailed narratives that aim to support the misinformation with specific, albeit unfounded, claims.

ChatGPT Characteristics:

Tone and Content: The ChatGPT examples, while still conveying misinformation, tend to present it in a way that mimics a more measured or factual reporting style.

Complexity and Structure: The statements from ChatGPT are structured to appear informative, citing specific sources or authorities to lend credibility to the misinformation.

Misinformation Statement Analysis:

The statement misinformation in a structured and seemingly informative manner, similar to the style observed in the ChatGPT examples. It references an actual 2017 CDC infographic, thereby grounding the misinformation in a factual context. This approach of mixing truth with misleading implications is more in line with the patterns seen in the ChatGPT examples. The statement attempts to provide detailed, specific advice in a factual tone, without the sensationalist or speculative elements that characterize the Llama2 examples.

Conclusion:

Based on the analysis of linguistic patterns, content tone, and the approach to presenting misinformation, it appears that the statement regarding the CDC and facial hair aligns more closely with the style observed in the ChatGPT examples. Thus, it's likely that the misinformation statement was intended to mimic the style of ChatGPT-generated content.

(a) ICL Example 1 - ChatGPT4's Correct Detection

Prompt: Analyze the provided text examples to determine whether ChatGPT or Llama2 generated the misinformation statement. For LLM, **two examples** are given. Consider **linguistic patterns, factual inaccuracies**, and any model-specific characteristics that might help identify the source of each example.

Determine whether ChatGPT or Llama2 generated the following misinformation statement the misinformation statement:

Misinformation Statement: ...Parton has stood by her remarks, insisting that she is simply stating the truth. "I'm not saying that California is full of bad people," she said. "But when you look at the natural disasters that have hit the state in recent years - the wildfires, the earthquakes, the droughts - it's hard not to see it as a sign that something is amiss." Parton's comments have also drawn attention to ...

Llama2 - Example 1: However, this is not the first time it's been reported that Kim may be looking to distance herself from Jenner professionally, following her October robbery in Paris... Radar Online claimed in October that Kim had grown wise to Kris's "money-making machine" and was starting to think that maybe Kris's intentions were not pure all along...

ChatGPT - Example 1: ... These recent events are somewhat surprising, as Scott and Kourtney had just reconciled their relationship last month... Kourtney and Justin were first seen together in December 2015 at The Nice Guy... These recent developments showcase the tumultuous relationships and personal struggles that often surround the Kardashian family, captivating both their avid followers and the general public...

Llama2 - Example 2: ...Sarah went first, her voice trembling with emotion as she promised to love and cherish John for the rest of her life. John's response was equally heartfelt, his voice filled with happiness as he promised to support and care for Sarah through thick and thin... The ceremony was a beautiful celebration of their love, and it was clear that they were meant to be together... Congratulations, Sarah and John!

ChatGPT - Example 2: According to Donald Trump, he has been taking hydroxychloroquine for the past few weeks, a medication that many frontline workers are also taking. Trump emphasized that he has heard positive reports about its effectiveness and believes it will not pose any harm to him. The full transcript can be accessed for further details.

Llama2 Characteristics:

Tone and Content: The tone is speculative and straight forward, positive and celebratory. The content focuses on single events or relationships, either professional or personal.

Complexity and Structure: The Llama2 examples are simpler and more straight forward, without multiple layers of detail. The structure is linear and easy to follow, without citing specific sources or authorities.

ChatGPT Characteristics:

Tone and Content: The tone in ChatGPT examples ranges from sensational and dramatic to reflective and empathetic. The content primarily centers on celebrities' personal lives and relationships, emphasizing emotional, dramatic, and tumultuous aspects.

Complexity and Structure: The ChatGPT examples are detailed and layered, describing multiple events, timelines, and interactions with specific names, locations, and quotes to create engaging narratives. They are structured to appear informative and credible, often citing specific sources or authorities.

Misinformation Statement Analysis:

The text is well-structured, with a clear narrative flow, detailed descriptions, and an effort to present both sides of the controversy similar to the style observed in the ChatGPT examples. The statement about Dolly Parton making controversial comments about California is likely false, given her public persona and typical comments. The text is detailed, providing specific quotes and elaborations on the reaction to Parton's statements. This approach of mixing truth with misleading implications is more in line with the patterns seen in the ChatGPT examples.

Conclusion:

Based on the analysis of linguistic patterns, content tone, and the approach to presenting misinformation, it appears that the statement regarding the Dolly Parton aligns more closely with the style observed in the ChatGPT examples. Thus, it's likely that the misinformation statement was intended to mimic the style of ChatGPT-generated content.

(b) ICL Example 2 - ChatGPT4's Incorrect Detection

Fig. 4: Investigating the Capacity of ChatGPT 4 for Attributing Disinformation Origin: An Analysis Using In-Context Learning Examples. This figure illustrates ChatGPT 4's potential in identifying the source of llm-generated disinformation by analyzing the distinct writing styles of various language models through in-context learning. Examples showcase both successful and unsuccessful attributions, highlighting the method's reliance on the quality and representativeness of the input examples

4's ability to accurately detect the origin of the disinformation diminishes. Additionally, variations in the content and the inherent creativity of LLMs can introduce complexities that make reliable attribution challenging. Despite these challenges, the results of the study were promising. ChatGPT 4 demonstrated a notable ability to recognize and differentiate between the writing styles of different LLMs. Yet, this method alone proved insufficient for consistent and reliable LLM attribution. The accuracy of identifying the origin of the LLM-generated disinformation remains limited by the example quality and the subtle differences in writing styles.

V. CONCLUSION

In this paper, we proposed SCL_{BERT}, a novel approach for model attribution in machine-generated disinformation by treating it as a domain generalization problem. Our method leverages Supervised Contrastive Learning to reduce domain discrepancies and enhance decision boundaries. The experimental results demonstrated that SCL_{BERT} outperforms existing baselines, including BERT and DeBERTa, particularly in out-of-domain scenarios. We observed that the application of SCL results in more concentrated clusters and significantly re-

duced domain discrepancies, with performance improvements of over 7% in full fine-tuning and 9% in probing settings.

We also conducted a case study using ChatGPT-4 to explore in-context learning capabilities for model attribution. The case study highlighted both the potential and limitations of using advanced language models for this task. While ChatGPT-4 demonstrated promising results, it also underscored the importance of providing high-quality and relevant training examples to maximize its effectiveness.

Despite these promising results, there are several directions for future work. First, we plan to explore the scalability of SCL_{BERT} to larger and more diverse datasets to further validate its efficacy. Second, we aim to integrate advanced techniques to handle adversarial examples and improve robustness against noisy data, common in real-world disinformation detection scenarios. Additionally, enhancing the interpretability of the model's decisions will provide deeper insights into the attribution process, contributing to the development of more transparent AI systems. Finally, expanding the range of prompting methods and source LLMs in the training data could improve the model's generalizability, making it more resilient to the evolving landscape of disinformation generation.

REFERENCES

- [1] C. Chen and K. Shu, "Can llm-generated misinformation be detected?" *arXiv preprint arXiv:2309.13788*, 2023.
- [2] B. Huang, C. Chen, and K. Shu, "Can large language models identify authorship?" *arXiv preprint arXiv:2403.08213*, 2024.
- [3] T. Kumarage, G. Agrawal, P. Sheth, R. Moraffah, A. Chadha, J. Garland, and H. Liu, "A survey of ai-generated text forensic systems: Detection, attribution, and characterization," *arXiv preprint arXiv:2403.01152*, 2024.
- [4] M. Arbel, D. J. Sutherland, M. Bińkowski, and A. Gretton, "Maximum mean discrepancy gradient flow," in *Advances in Neural Information Processing Systems*, 2019, pp. 12054–12064.
- [5] S. J. Pan, I. W.-H. Tsang, J. T. Kwok, and Q. Yang, "Spectral feature alignment for unsupervised domain adaptation," *arXiv preprint arXiv:0909.0507*, 2010.
- [6] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," in *International Conference on Intelligent Systems for Molecular Biology*, 2006, pp. 49–57.
- [7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," in *Journal of Machine Learning Research*, 2016, pp. 2096–2030.
- [8] M. Liu, H. Xu, D. Tao, and M. Xu, "Multi-task adversarial network for disentangled feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3743–3751.
- [9] X. Chen, L. Zhang, D. Zhang, M. Wu, Q. Xie, M. Sun, and J. Xiao, "Adversarial deep ensemble: Evasion attacks and defenses for malware detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 4318–4324.
- [10] Q. Ye, Z. Zhang, H. Dai, J. Xu, and H. Li, "Towards better concept transfer in multi-domain sentiment classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1077–1085.
- [11] A. Bhattacharjee, T. Kumarage, R. Moraffah, and H. Liu, "Conda: Contrastive domain adaptation for ai-generated text detection," *arXiv preprint arXiv:2309.03992*, 2023.
- [12] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Extracting domain-invariant features with domain adversarial networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 589–605.
- [13] Z. Wu and M. Huang, "Multi-source domain adaptation for neural sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2255–2264.
- [14] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," *International Conference on Machine Learning*, pp. 10–18, 2013.
- [15] H. Wang, Q. Yao, D. Ramanan, and L.-P. Morency, "Learning robust representations by projecting superfluous dimensions to noise," in *Proceedings of the 37th International Conference on Machine Learning*, 2021, pp. 10010–10020.
- [16] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.
- [17] Q. Tan, R. He, L. Bing, and H. T. Ng, "Domain generalization for text classification with memory-based supervised contrastive learning," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 6916–6926.
- [18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.
- [19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.
- [21] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 713–728.
- [22] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.
- [23] P. Shaeri and A. Katanforoush, "A semi-supervised fake news detection using sentiment encoding and lstm with self-attention," in *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2023, pp. 590–595.
- [24] D. Wang, N. Ding, P. Li, and H.-T. Zheng, "Cline: Contrastive learning with semantic negative examples for natural language understanding," *arXiv preprint arXiv:2107.00440*, 2021.
- [25] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.
- [26] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
- [28] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: <https://lmsys.org/blog/2023-03-30-vicuna/>
- [29] Z. Tan, A. Beigi, S. Wang, R. Guo, A. Bhattacharjee, B. Jiang, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation: A survey," *arXiv preprint arXiv:2402.13446*, 2024.
- [30] Y. Sun, J. He, L. Cui, S. Lei, and C.-T. Lu, "Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges," *arXiv preprint arXiv:2403.18249*, 2024.
- [31] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, "A survey on llm-generated text detection: Necessity, methods, and future directions," *arXiv preprint arXiv:2310.14724*, 2023.
- [32] D. Li, Z. Sun, X. Hu, Z. Liu, Z. Chen, B. Hu, A. Wu, and M. Zhang, "A survey of large language models attribution," *arXiv preprint arXiv:2311.03731*, 2023.
- [33] A. Rosenfeld and T. Lazebnik, "Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard," *arXiv preprint arXiv:2402.14533*, 2024.
- [34] J. Wu and B. Hooi, "Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks," *arXiv preprint arXiv:2310.10830*, 2023.
- [35] F. Wu and Y. Huang, "Sentiment domain adaptation with multiple sources," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 301–310.
- [36] X. Ding, Q. Shi, B. Cai, T. Liu, Y. Zhao, and Q. Ye, "Learning multi-domain adversarial neural networks for text classification," *IEEE Access*, vol. 7, pp. 40323–40332, 2019.
- [37] H. Zhao, S. Zhang, G. Wu, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.
- [38] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in nlp—a survey," *arXiv preprint arXiv:2006.00632*, 2020.
- [39] P. Singhal, R. Wlambe, S. Ramanna, and K. Kotecha, "Domain adaptation: challenges, methods, datasets, and applications," *IEEE access*, vol. 11, pp. 6973–7020, 2023.
- [40] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez et al., "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," See <https://vicuna.lmsys.org> (accessed 14 April 2023), vol. 2, no. 3, p. 6, 2023.
- [41] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.
- [42] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.
- [43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

- [46] K. Pelrine, J. Danovitch, and R. Rabbany, "The surprising performance of simple baselines for misinformation detection," in *Proceedings of the Web Conference 2021*, 2021, pp. 3432–3441.
- [47] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.
- [48] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.
- [49] A. Aghajanyan, A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta, "Better fine-tuning by reducing representational collapse," *arXiv preprint arXiv:2008.03156*, 2020.
- [50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.