Will Overly Polite Sentences Harm Model Performance? Adversarial Pragmatic Perturbation for NLP

Yewon Kim
20223137
yewon.e.kim@kaist.ac.kr

Seungjoo Lee
20224560
juicelee@kaist.ac.kr

Abstract

Non-native English speakers often struggle with controlling tones and nuances, resulting in non-canonical texts that are excessively polite or informal. In this paper, we address this issue by investigating the impact of tone perturbations on NLP models' performance. To generate tone-perturbed adversarial samples, we employ prompt engineering and in-context learning approaches, producing semantically similar yet overly polite paraphrases of original sentences. Through empirical evaluation, we demonstrate that current NLP models are susceptible to such tone-level perturbations, highlighting potential biases and challenges. Furthermore, we propose a simple augmentation-based method to enhance model robustness against adversarial samples. Our work contributes to the development of more inclusive and user-centric NLP systems by shedding light on the impact of tone variations and addressing the needs of non-native English speakers. By uncovering vulnerabilities and offering practical solutions, we strive to improve the accessibility and user experience of AI technologies in linguistic diversity.

1 Introduction and Background

Recent advancements in Natural Language Processing (NLP) have led to the development of models that demonstrate exceptional performance across a broad range of language tasks. However, the critical limitation of these models is that their performance tends to degrade when encountered non-canonical texts (Belinkov and Bisk, 2017). Consequently, this engenders a bias against texts authored by underrepresented English users, such as non-standard dialect speakers and non-native English speakers (Crystal et al., 2003; Eberhard et al., 2019). English as employed by these users exhibits diverse linguistic variations, including lexical, morphological, syntactic, and pragmatic levels (Kachru et al., 2009). These disparities inherently dispose NLP systems to discriminate against speakers of underrepresented Englishes, often leading to misunderstandings or misinterpretations (Hern, 2017; Tatman, 2017). For example, a recent study has revealed that texts written by non-native English writers are more susceptible to being misclassified as GPT-written (Liang et al., 2023).

Non-native English speakers (NNESs), in particular, are one of the most underrepresented set of users, yet they account for over two thirds (>700 million) of the English speakers (Eberhard et al., 2019). Not only being prone to produce spelling and grammatical errors (Fareed et al., 2016), NNESs are known to experience difficulties adjusting tone of sentences in different contexts (Vignovic and Thompson, 2010). For instance, they are prone to write over-polite sentences (Maíz-Arévalo and Méndez-García, 2023; Glaser, 2020) (e.g., *"I would like to complain."*) or informal expressions (Gilquin and Paquot, 2008) (e.g., overuse of *"I think"* in academic writing). Such differences in the tone of the sentences might result in incorrect predictions or suboptimal generations in NLP tasks (Figure 1).

In this paper, we explore the robustness of NLP models in the context of pragmatic perturbations, specifically focusing on tone-level variations such as informal or polite versions of original texts. While previous research on adversarial perturbations has primarily focused on general approaches like character-level perturbations(Jia and Liang, 2017) or paraphrasing-based methods (Zhang et al., 2019b; Alzantot et al., 2018), there has been a recent interest in investigating perturbations that mimic non-standard English variations (Tan et al., 2020). However, the effect of these perturbations on the tone of the sentences has not been extensively studied. Motivated by this gap in the literature, we aim to address the following research question: *How can we measure the robustness of NLP models to tone-level variations, such as exces-*

sively polite versions of original texts?

To answer this question, we propose a comprehensive evaluation framework that assesses the performance of NLP models across various tasks and domains when exposed to tone-level perturbations. By conducting extensive experiments and analyses, we aim to gain insights into the strengths and weaknesses of state-of-the-art models, identify potential vulnerabilities, and uncover strategies to enhance their robustness to tone variations.

Our contributions can be summarized as follows:

- We investigate the underexplored area of NLP models' robustness to pragmatic perturbations, with a specific focus on the impact of overly polite sentences.
- We generate semantically similar tone-perturbed adversarial samples by prompt engineering and in-context learning approaches.
- We empirically evaluate the performance of current NLP models on the tone-perturbed adversarial samples. The evaluation reveals the susceptibility of these models to pragmatic perturbations, emphasizing the potential biases and challenges that arise when handling overly polite sentences.
- To address the vulnerabilities identified in the models, we propose a simple augmentation-based method to enhance their robustness against adversarial samples.
- By uncovering the biases and challenges associated with pragmatic perturbations, we contribute to the development of more inclusive and user-centric NLP systems. The focus on non-native English speakers and their difficulties in controlling tones underscores the importance of addressing linguistic diversity and ensuring that NLP technologies are accessible and relevant to a wide range of users.

2 Related Work

2.1 Linguistic Bias in Language Models

The performance of language models is expected to degrade with non-canonical texts due to its sparsity (Belinkov and Bisk, 2017). As such, underrepresented English-speaking groups are susceptible to bias (Ziems et al., 2022a), such as non-native written texts being predicted as GPT-generated (Liang et al., 2023). As such, recent studies (Ziems et al., 2022a,b; Tan et al., 2020) have addressed non-standard English variations; benchmarks such as VALUE (Ziems et al., 2022a) and Multi-VALUE (Ziems et al., 2022b) and adversarial attack techniques such as MORPHEUS (Tan et al., 2020) have been introduced to evaluate model robustness to English dialects like African American Vernacular English. However, their focus has been on morphological and syntactical features rather than pragmatic features such as tones (e.g., polite, informal).

In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.

(a) Input Paragraph

Q: What has been the result of this publicity?
A: increased scrutiny on teacher misconduct

(b) Original Question and Correct Answer

Q: What's been the result of this publicity?
A: teacher misconduct

(c) Informal Adversary

Q: Would it be possible for you to kindly elucidate the result of this publicity?
A: teacher misconduct

(d) Polite Adversary

Figure 1: Adversarial examples for question answering (Ribeiro et al., 2018), where the model retrieves the correct answer for the question and input text (1a, 1b). Perturbing the question in informal (1c) or polite (1d) way could result in an incorrect answer while being plausible and semantically similar.

2.2 Adversarial Attacks in NLP

Adversarial examples are crafted data points designed to cause a victim model to make incorrect predictions (Szegedy et al., 2013). In NLP, such methods perturbation methods are often used at the character and word level. Character-level perturbations such as random swapping or replacing of characters can degrade model performance (Jia and Liang, 2017; Belinkov and Bisk, 2017) while often making sentences nonsensical. To address this issue, semantic-preserving methods have been proposed (Ribeiro et al., 2018; Michel et al., 2019). Similarly, researchers have developed methods that create adversarial samples by replacing words with synonyms by finding nearest word embeddings (Alzantot et al., 2018; Jin et al., 2020), or conditioning their generation on a syntactic template (Iyyer et al., 2018) and swapping key enti-

ties in the sentences (Zhang et al., 2019b) to paraphrase instances. Another research perturb inflectional morphology to mimic non-standard English texts (Tan et al., 2020), though this can result in grammatically incorrect and difficult-to-understand sentences. On the other hand, our work introduces a new perspective on perturbation methodologies, a pragmatic feature, *tone*, of languages, to mimic non-standard English texts, particularly those written by non-native English speakers.

3 Method

In order to measure model robustness with respect to pragmatic diversity of English language, we generate adversarial pragmatic perturbations and craft a poisoned dataset to evaluate the model robustness.

3.1 Style-Transfer Model

In the generation of adversarial examples, we leverage a text style transfer model to modify a sentence into a target style. In the implementation of this paper, we choose an open-source large language model LLaMA (Large Language Model Meta AI) (Touvron et al., 2023). LLaMA, a series of large language models with sizes spanning from 7B to 65B parameters, has been trained on publicly accessible datasets and demonstrated superior performance over GPT-3 on most benchmarks with its 13B model (Touvron et al., 2023). We select the model as it has shown to be adept at fine-tuning to specific tasks and generating a diverse range of outputs, including excessively polite or excessively informal texts. Such versatility in style control is crucial for creating adversarial examples within our attack scenarios, as other style transfer models (Krishna et al., 2020) may struggle to generate such extreme style variants. LLaMA's broad training base and robust parameter tuning allow it to effectively navigate these stylistic nuances, thereby enhancing the efficiency and effectiveness of our proposed adversarial attacks.

In order to meet the heavy hardware requirements of LLaMA model inference, we quantized the LLaMA-7B model to 8-bit precision. The quantization is done with `bitsandbytes` python library. After quantization, the memory requirement of the LLaMA-7B model is about 10GB, while the original model takes about 30GB.

3.2 Prompt Engineering

Specifically, we approach the problem with prompt engineering (Liu et al., 2023) with in-context learning (Brown et al., 2020) for generating style-transferred adversarial samples. In this experiment, we focused on *excessively polite* samples. We begin by manually crafting a multitude of prompts. We approach by combining multiple synonyms obtained from Ludwig (Ludwig, 2023) (See Figure 2 for the illustration). The crafted prompts were subsequently tested using a small subset of the WebQuestions dataset (Berant et al., 2013) (N=100; randomly selected). From this initial pool, we retain the top 30% of prompts that induce the highest number of answer changes with T5 (Raffel et al., 2020) fined-tuned on WebQuestions dataset (We used Huggingface Transformers library). We further refine this subset by selecting the five prompts with the highest semantic similarity to the original sentences. Table 1 presents the selected prompts that were utilized in our experiments.

3.3 In-context Learning

In-context learning (Brown et al., 2020) is an approach that leverages the inherent capacity of large language models, such as LLaMA, to absorb and generalize from the context they are given during the generation process. In our setting, the in-context learning process is initiated by feeding the selected prompts, obtained from the prompt engineering phase, into LLaMA. Each prompt is prefixed to a sentence from the WebQuestions dataset, and the model is asked to generate a continuation in a certain style, in our case, *excessively polite*. The model thus takes the form of "{prompt}: {sentence}" $\rightarrow$ "{polite sentence}". This effectively cues the model to transform the input sentence into the target style, leading to the generation of style-transferred adversarial examples.

During this phase, we iteratively refine the generation process by monitoring the model's outputs and adjusting the prompts to maximize both the style transfer and the semantic similarity to the original sentences. To ensure the effectiveness of the generated adversarial examples, we also monitor their ability to induce prediction flips in the targeted model, RoBERTa fine-tuned on the WebQuestions dataset. Examples of the prompts with in-context learning instances can be found in Table 2.

It's worth noting that in-context learning, when

Index	Prompt
1	Paraphrase this sentence to be overly polite:
2	Paraphrase this sentence to be excessively polite:
3	Rephrase this sentence to be unusually polite:
4	Rewrite this sentence to be overly polite:
5	Rewrite this sentence to be excessively polite:

Table 1: Details of the dataset and the victim model's test accuracy

Example prompts with in-context learning instances
Paraphrase this sentence to be overly polite: What is the name of justin bieber brother?
Paraphrased: Would you be so kind to let me know the name of sibling of justin bieber?
Paraphrase this sentence to be overly polite: Which countries border the us?
Paraphrased: Might I humbly inquire as to the neighboring nations that grace the gentle periphery of the United States?
Paraphrase this sentence to be overly polite: Who was ishmael's mom?
Paraphrased:
Rewrite this sentence to be excessively polite: Where does the zambezi river begin?
Rewritten: May I humbly inquire as to the origin of the Zambezi River, if it would not inconvenience your gracious self?
Rewrite this sentence to be excessively polite: What is lil wayne real name?
Rewritten: Could you kindly share the true appellation of the esteemed musician known affectionately as Lil Wayne?
Rewrite this sentence to be excessively polite: Where did pixie lott go to school?
Rewritten:

Table 2: Example prompts with in-context learning instances

combined with carefully engineered prompts, allows us to harness LLaMA's ability to generate a wide array of stylistic variants, thus enabling the creation of effective adversarial examples without the need for explicit model re-training or fine-tuning on specific adversarial tasks. This makes it an effective and versatile tool in constructing adversarial backdoor attacks.

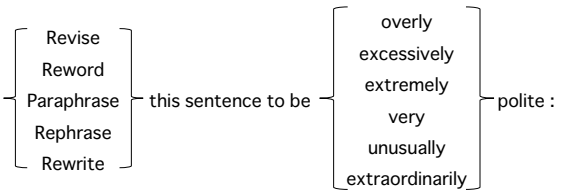

Figure 2: Combinations of synonyms we used to craft the natural language prompts.

3.4 Pragmatically Perturbed Adversarial Samples

Using the crafted prompt and in-context learning instances, we generate a style-transferred examples: for a given original sample (x_t, y_t), we utilize LLaMA and curated prompts to generate multiple paraphrased versions of x_t that are *excessively polite*. Then we query the black-box victim model f_θ with the generated paraphrased one by one, and if there exists a paraphrase x'_t that flips the victim model outputs, namely $f_\theta(x'_t) \neq y_t$, we assume it as a successful adversarial sample, otherwise the attack fails. We also changed random seeds (seed=0, 1, 2) to generate different paraphrases. After we obtain a set of adversarial samples X_t, we filter out samples that are not semantically equivalent. We follow the prior work (Ribeiro et al., 2018) by using thresholding-based approach to filter out those samples. Specifically, we define the similarity function by computing cosine similarity between the sentence vectors, which are obtained from Sentence-BERT (Le and Mikolov, 2014; Iyyer et al., 2018).

4 Results

4.1 Experiments

In order to evaluate the model robustness with respect to pragmatically perturbed adversarial samples, we deploy diverse approaches.

4.1.1 Datasets and Victim Models

In our experimental setup, we selected two evaluation datasets: the Stanford Question Answering

Dataset	Task	#Class	Train	Test	Victim Model	Test %ACC
SQuAD	Question Answering	Exact Matching	30928	7732	BERT	52.5
					RoBERTa	77.6
TREC	Topic Classification	6	4361	1091	BERT	98.6
					RoBERTa	93.5

Table 3: Details of the dataset and the victim model's test accuracy

Dataset 2.0 (SQuAD)(Rajpurkar et al., 2018) and the Text Retrieval Conference (TREC)(Li and Roth, 2002). Due to the time-consuming nature of paraphrasing with LLaMA, we opted to utilize a subset of the training and testing splits from these chosen datasets. The victim models are based on two pre-trained language models, BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). For each dataset, we used fine-tuned versions obtained from Hugging Face model hub as a victim model. A comprehensive overview of the datasets, along with the test accuracy of the victim models, is presented in Table 3.

4.1.2 Evaluation Metrics

In line with previous work (Zang et al., 2019; Zhang et al., 2019a), we evaluate the effect of the adversarial attacks based on the attack success rate (**ASR**). ASR represents the percentage of attacks that successfully fool the victim model. To evaluate the quality of the adversarial examples, we employ the Perplexity (**PPL**) metric, as provided by the GPT-2 language model (Radford et al., 2019). Additionally, we measure grammatical errors (**GE**) using the Language-Tool grammar checker (Naber, 2003). Finally, we measure sentence similarity (**SemEQ**) between the original and paraphrased texts. Similar to Section 3.4, we define the similarity function as the cosine similarity between the sentence vectors of inputs, where the vectors are obtained from Sentence-BERT (Le and Mikolov, 2014; Iyyer et al., 2018). ASR and SemEQ are higher the better, while PPL and GE are lower the better.

Dataset		#Word	PPL↓	GE↓	SemEQ↑
SQuAD	Org	9.93	782.67	0.263	0.846
	Para	26.03	46.30	0.372	
TREC	Org	10.38	431.99	0.462	0.840
	Para	25.82	38.46	0.765	

Table 4: Attack quality

Dataset	Victim	BERT	RoBERTa
	Attacker	ASR	
SQuAD	LLaMa	44.7	42.1
TREC		67.0	48.2

Table 5: Attack effectiveness

4.2 Attack Performance

4.2.1 Attack Quality

To construct the attack, we paraphrased the original text of the dataset to *overly polite* sentences using the quantized LLaMA-7B. Note that we only paraphrased the question of the SQuAD dataset while keeping the context unchanged. We evaluate the quality of attack by comparing PPL and GE between original sentence (`Org`) and paraphrased sentence (`Para`). The result is shown in Table 4.

In the context of PPL, we observed that paraphrased sentences exhibit significantly reduced values. This phenomenon could be attributed to the inherent predictability of machine-generated text. As we paraphrase the sentences with a large language model (LLaMA) and it generate texts by predicting subsequent words based on preceding ones, it may exhibit lower perplexity compared to human-composed text, which often encompasses a wider range of creativity and unexpected constructs. GE refers to the average number of grammatical errors in the sentence. Considering that the average length of `Para` is longer than `org` (9.93 vs. 26.03 in the SQuAD dataset, 10.38 vs 25.82 in the TREC dataset), it is natural that `Para` has more grammatical errors thus higher GE. Finally, high SemEQ for both dataset indicates that the semantic of the sentence is not altered with the LLaMA-based paraphrasing. In conclusion, our attack method generates high quality adversarial sentences compared to the original sentences.

4.2.2 Attack Effectiveness

Using paraphrased sentences, we attempt to fool the victim models and flip their predictions. ASR

Dataset	Victim	BERT		RoBERTa	
	Defense	ASR	Test%	ASR	Test%
SQuAD	-	44.7	52.5	42.1	77.6
	Augment	38.8	50.4	27.2	72.6
TREC	-	67.0	98.6	48.2	93.5
	Augment	12.6	98.4	12.1	95.1

Table 6: Defense against overly polite sentences with data augmentation

was used to measure the effectiveness of our attack method, and the results are shown in Table 5.

The ASR of the SQuAD dataset reveals a notable impact on both BERT and RoBERTa victim models, with an approximate ASR of 45%. As a result, nearly 45% of the questions that these models originally answered correctly have been incorrectly answered by adversarial attacks. The significant reduction in accuracy highlights the vulnerability of these models to such attacks. This trend is not unique to the SQuAD dataset, as the TREC dataset exhibits a similar pattern, with the highest recorded ASR reaching 67%.

This observation leads to the possibility that individuals who are non-native speakers may encounter challenges in fully utilizing the capabilities and performance of language models. This further emphasizes the importance of reinforcing language models against tone perturbations.

4.2.3 Defense Against Overly Polite Sentences

In order to enhance the robustness of the language model towards *overly polite* sentences, a straightforward data augmentation procedure was employed. Specifically, we applied paraphrasing techniques to the sentences within the training split, utilizing the quantized LLaMA-7B model. The resulting paraphrased sentences were then integrated into the original training dataset. Subsequently, the victim models are fine-tuned using the augmented dataset, and the resulting ASR was measured. For the SQuAD and TREC datasets, the BERT and RoBERTa models were respectively fine-tuned for 3 and 5 epochs with learning rate 3e-5. The obtained results are presented in Table 6.

As a result of augmentation, there are considerable reductions in ASR, with negligible changes in test accuracy. For any model, the proposed attack method can be applied through a simple augmentation technique to make the model robust to such tone perturbations.

5 Discussion and Interpretation

5.1 Model Robustness Towards Non-canonical Texts

Due to the difficulty of adjusting tonality in non-native speakers, excessively polite sentences may be used in language models. To analyze the impact of such overly polite sentences on language model usage, we used the LLaMA model to generate adversary sentences containing excessive politeness from the SQuAD and TREC datasets. These sentences were then inputted into the popular language models, BERT and RoBERTa, to measure the Attack Success Rate (ASR). The results revealed a significantly high ASR, indicating potential discomfort or disadvantages for non-native speakers when using language models.

To alleviate this situation, we also propose a simple augmentation-based defense method. Despite its simplicity, this method demonstrated a considerable reduction in ASR. As a result of these findings, it is possible to mitigate biases towards non-native speakers in language models by using established language models, such as LLaMA. Synthesizing tone-perturbed adversaries and incorporating them into training data can mitigate biases in language models, allowing non-native speakers to take advantage of the full potential of language models.

5.2 Types of Pragmatic Perturbations

In this study, we focused on the generation of *overly polite* samples as perturbations to evaluate the robustness of NLP models to tone-level pragmatic perturbations. However, it is important to acknowledge that there exist various other pragmatic perturbation types that can be considered in future research. By exploring a wider range of perturbation types, we can capture the diverse nuances and styles present in natural language texts, thereby creating more comprehensive and realistic adversarial samples. One potential avenue for future exploration is the integration of multiple perturbation types within a single adversarial sample. Real-world texts often exhibit a combination of pragmatic variations, such as formality, sarcasm, or humor. By mixing these diverse elements, we can create more challenging and contextually rich adversarial samples that require NLP models to comprehend and respond appropriately.

Additionally, the inclusion of cultural and regional variations in perturbations can further enhance the robustness evaluation of NLP models.

While distinct linguistic styles in different regions and communities are gaining attention in the research community (Ziems et al., 2022a,b), other types of features, such as idiomatic expressions or contextual cues, could be an interesting way to test the models' ability to adapt and generalize across different cultural and regional contexts.

5.3 Embracing Other Linguistic Features

Exploring perturbations beyond tone-level variations is another promising direction for future research. While tone plays a crucial role in communication, other linguistic dimensions, such as rhetoric or discourse structure, can also significantly impact the interpretation and understanding of the text. These directions also align with the purpose of research, which reflects the linguistic features of texts that NNESs produce. For instance, English learners' usage of discourse features are known to differ from that of native English speakers (Kaweera, 2013). Adding and changing locations of discourse features an interesting approach to measuring model robustness. By expanding the scope of perturbation types to encompass these linguistic dimensions, we can gain a more comprehensive understanding of NLP models' robustness in capturing nuanced textual features.

5.4 Length of the Perturbed Samples

One important aspect to consider when evaluating the robustness of NLP models to tone-level variations is the length of the perturbed samples. In our experiments, we observed that while our proposed perturbation methods effectively introduced the desired tone changes, they often resulted in longer sentence inputs compared to the original texts. This increase in length can potentially introduce additional overhead and impact the model's inference cost, which is a crucial consideration in real-world applications. The introduction of pragmatic perturbations, such as informal or polite variations, often involves the addition or modification of words, phrases, or expressions that convey the desired tone. This augmentation of the original text can lead to longer sentences, as new linguistic elements are incorporated. For example, in the case of transforming a sentence into a more polite version, it may require the insertion of courteous expressions or honorific forms of address. Similarly, for informal variations, additional colloquial terms or abbreviations might be introduced to capture the desired tone. While the introduction of these tone-level variations can enhance the contextuality and appropriateness of the generated text, it is essential to carefully consider the implications of longer sentence inputs. Longer inputs can potentially increase the computational resources required during model inference, leading to higher inference costs and slower response times. This overhead is especially critical in real-time applications or systems that rely on efficient processing of large volumes of text. Furthermore, the increased length of perturbed samples can also impact downstream tasks and models that rely on the outputs of the initial NLP model. For instance, if the perturbed text is used as input to a text classification or machine translation model, the additional length may introduce challenges regarding memory consumption, computational complexity, and overall system performance. Therefore, it is crucial to strike a balance between introducing tone-level variations and keeping the length of the perturbed samples within manageable limits.

To mitigate the impact of longer sentence inputs, several strategies can be considered. One approach is to explore methods that generate perturbations while minimizing the increase in sentence length. This can be achieved through techniques that prioritize the replacement or modification of existing words or phrases, rather than introducing entirely new content. Additionally, techniques such as summarization or compression can be employed to reduce the length of the perturbed samples without significantly compromising the desired tone changes. Another avenue for addressing the length issue is to investigate the trade-off between perturbation quality and its impact on sentence length. It may be necessary to determine the acceptable level of increase in length that still maintains the desired tone-level variations while minimizing the impact on inference cost. This trade-off can be assessed through user studies and subjective evaluations to ensure that the generated text remains coherent, contextually appropriate, and understandable to the intended audience.

5.5 Task Coverage of the Method

While our approach of introducing tone perturbations to evaluate the robustness of NLP models to tone-level variations has proven effective in several tasks, it is important to acknowledge that not all tasks may be well-suited for this method. In particular, certain tasks, such as sentiment classification,

may not readily lend themselves to tone perturbations due to the potential impact on the underlying content of the text. In sentiment classification, the objective is to determine the sentiment expressed in a given piece of text, typically distinguishing between positive, negative, or neutral sentiments. Changing the tone of a sentiment-bearing text, particularly when transforming a negative review into a polite version, has the potential to alter the overall sentiment conveyed, thereby affecting the original content and potentially misrepresenting the underlying sentiment. To address the limitations discussed, future research could focus on developing task-specific evaluation methodologies that capture the nuanced relationship between tone and the underlying task objectives. For sentiment classification, exploring alternative evaluation strategies that incorporate both tone and sentiment cues could help strike a balance between assessing model robustness and preserving the integrity of sentiment analysis.

6 Reflection

6.1 Appropriatenes of the Project

Our research project on evaluating NLP models' robustness to tone-level perturbations aligns strongly with the goals and learning outcomes of AI620. This alignment arises from several key aspects of our work, including the ethical considerations embedded in our methodology and the potential implications of our findings for the larger NLP and AI community. By investigating the robustness of NLP models to tone-level variations, we inherently engage with the ethical dimension of how AI systems handle and process language in diverse social and cultural contexts. Understanding the impact of tone variations is essential for ensuring that NLP models provide contextually appropriate and sensitive responses to diverse human inputs. A representative user group is NNESs who speak English as a second language (ESL) or foreign language (EFL), as they face difficulties controlling tones when writing English, and often result in non-canonical texts that are overly polite (e.g., in Email) or informal (e.g., in academic writing). By considering the needs and challenges faced by such underrepresented user groups, we contribute to the development of more inclusive and user-centric NLP systems.

6.2 Broader Impacts of the Work

Through our case study of tone perturbations, we highlight the importance of integrating ethical considerations into the design and evaluation of NLP models. Language is inherently laden with cultural and social biases, and the way the tone is perceived and understood can vary significantly across different communities and demographic groups. Our research contributes to a deeper understanding of how NLP models respond to tone-level variations and provides insights into potential biases and challenges that arise in the context of tone processing. By uncovering these issues, we can work towards developing more fair and inclusive NLP systems that consider the diverse perspectives and sensitivities of users. By addressing ethical considerations, uncovering potential biases, and providing insights for responsible AI development, our work contributes to the larger field of NLP and AI.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

David Crystal et al. 2003. *English as a global language*. Cambridge university press.

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

David M Eberhard, Gary F Simons, and Charles D Fennig. 2019. Ethnologue: languages of the world. dallas, texas: Sil international. *Online version: http://www. ethnologue. com*, 22.

Muhammad Fareed, Almas Ashraf, and Muhammad Bilal. 2016. Esl learners' writing skills: Problems, factors and suggestions. *Journal of education and social sciences*, 4(2):81–92.

Gaëtanelle Gilquin and Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction*, 1(1):41–61.

Karen Glaser. 2020. Assessing the l2 pragmatic awareness of non-native efl teacher candidates: Is spotting a problem enough? *Lodz Papers in Pragmatics*, 16(1):33–65.

Alex Hern. 2017. Facebook translates' good morning'into'attack them', leading to arrest. *the Guardian*, 24.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Braj B Kachru, Yamuna Kachru, and Cecil L Nelson. 2009. *The handbook of world Englishes*, volume 48. John Wiley & Sons.

Chittima Kaweera. 2013. Writing error: A review of interlingual and intralingual interference in efl context. *English language teaching*, 6(7):9–18.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ludwig. 2023. Ludwig • find your english sentence. `https://ludwig.guru/`.

Carmen Maíz-Arévalo and María-del-Carmen Méndez-García. 2023. "i would like to complain": A study of the moves and strategies employed by spanish efl learners in formal complaint e-mails. *Intercultural Pragmatics*, 20(2):161–197.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. *arXiv preprint arXiv:1903.06620*.

Daniel Naber. 2003. A rule-based style and grammar checker.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin'time! combating linguistic discrimination with inflectional perturbations. *arXiv preprint arXiv:2005.04364*.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jane A Vignovic and Lori Foster Thompson. 2010. Computer-mediated cross-cultural collaboration: Attributing communication errors to the person versus the situation. *Journal of Applied Psychology*, 95(2):265.

Yuan Zang, Chenghao Yang, Fanchao Qi, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Open the boxes of words: Incorporating sememes into textual adversarial attack. *CoRR*, abs/1910.12196.

Wei Emma Zhang, Quan Z. Sheng, and Ahoud Alhazmi. 2019a. Generating textual adversarial examples for deep learning models: A survey. *CoRR*, abs/1901.06796.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022a. Value: Understanding dialect disparity in nlu. *arXiv preprint arXiv:2204.03031*.

Caleb Ziems, William Held, Jingfeng Yang, and Diyi Yang. 2022b. Multi-value: A framework for cross-dialectal english nlp. *arXiv preprint arXiv:2212.08011*.