



**A METRIC FOR ROBUST MODEL EVALUATION THROUGH TEXT
PERTURBATIONS**

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Global Management and Innovation technologies, Master's thesis

2024

Sergei Smirnov

Examiner(s): Professor Leonid Chechurin

Ildar Idrisov, Phd (Tech.)

ABSTRACT

This thesis introduces a new approach to improve robustness of text generative models by introduction of noise or text perturbations in prompts. A metric is developed to evaluate how well large language models perform when faced with noisy data. The efficiency of the approach is demonstrated experimentally by measuring models' accuracy on controlled perturbations in 1200 questions from CommonsenseQA dataset. The result opens a new promising direction to evaluate and improve AI based solutions quality.

Lappeenranta–Lahti University of Technology LUT

Your school: LUT School of Engineering Sciences

Your degree programme: Industrial Engineering and Management

Smirnov Sergei

A METRIC FOR ROBUST MODEL EVALUATION THROUGH TEXT PERTURBATIONS

Master's thesis

2024

49 pages, 11 figures, 1 tables and 0 appendices

Examiner(s): Professor Leonid Chechurin and Ildar Idrisov, Phd (Tech.)

Keywords: large language models, model robustness, text perturbations, evaluation metrics

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

Oma schoolisi: LUTin insinööritieteiden tiedekunta

Oma koulutusohjelmasi: Tuotantotalous

Mikäli yhteistutkinto-ohjelma, Yhteistyöyliopisto: Yliopiston nimi

Smirnov Sergei

MITTARI VAHVAN MALLIN ARVIOINTIIN TEKSTIN HÄIRIÖIDEN KAUTTA

Koulutusohjelmasi ja tutkielmasi: Insinööritieteiden DI-tutkielma

2024

49 sivua, 11 kuvaa, 1 taulukkoa ja 0 liitettä

Tarkastaja(t): Professori Leonid Chechurin ja Ildar Idrisov, Phd (Tech.)

Avainsanat: suuret kielimallit, mallien robustius, tekstihäiriöt, arviointimetriikat

Table of contents

1 Introduction	5
1.1 What is Robustness? Main definitions and research questions	6
1.2 The Need for a Robustness Evaluation	9
2 Background	10
2.1 Audio dithering	11
2.2 Textual Data and LLMs	12
3 Methodology	12
3.1 Experimental Setup	12
3.2 Evaluation Protocol	13
3.3 Tools and Technologies	14
3.4 Expected Outcomes	14
4 Literature review	14
4.1 Papers search	15
4.1.1 Scopus database papers search	15
4.1.2 Web of Science database papers search	17
4.1.3 Other databases papers search	18
4.2 Papers selection	18
4.3 Papers review	19
4.4 Summary of reviewed insights	32
5 Proposed approach	33
5.1 Dataset	34
5.2 Types noise	35
5.2.1 Orthographic Noise	36
5.2.2 Semantic Noise	37
5.3 ProposedMetric calculation	37
6 Experiments	38
7 Limitations	42
8 Future work	42
9 Discussion	42
10 Conclusion	43
References	44

1 Introduction

Large language models (LLMs) such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) stand at the cutting edge of natural language processing (NLP) technology. These models possess the capability to generate text that is often indistinguishable from that written by humans, leading to their widespread adoption in a variety of applications. These range from automated content creation and chatbot development to sophisticated translation services (Devlin et al., 2018; Radford et al., 2018).

Despite their impressive capabilities, LLMs are not without their flaws. Issues such as inconsistency in text generation, lack of contextual depth and factual inaccuracies frequently occur, presenting significant challenges. (Bender et al., 2021). These problems undermine the reliability of LLMs, particularly in critical applications that require high levels of accuracy and dependability, such as legal document automation, educational content generation, and healthcare communication systems.

To illustrate the concept of model robustness (refer to definitions in Section 1.1 below), let's consider image generation models. Although the presented work focuses on text generative models, images are used for better illustration of idea. This highlights that small perturbations in the prompt can significantly influence the model output, a characteristic shared by both text and image models. For instance, if we provide a prompt with a typo like 'make image of unikorn,' most people would still understand the intended request. However, AI models don't possess the same robustness as humans (Fig. 1). While humans can infer the correct interpretation, AI models may struggle with such variations.



Figure 1. Two image generative models were prompted with 'make image of unikon.' Stable Diffusion Realistic Vision is on the left, DALL-E 3 is on the right. DALL-E successfully understood and responded to the prompt, while Stable Diffusion did not.

The central idea of the work is that what works well for image AI based design should also work for AI based text design. But quantitative assessment method is needed to evaluate the quality of generated text. This brings the central research question of the thesis:

Research Question: How to evaluate the robustness of transformer-based text generative models over the class(es) of prompt deviations in the domain of common-sense reasoning, as measured by accuracy?

1.1 What is Robustness? Main definitions and research questions

Several concepts in this work can be interpreted differently by different fields or schools of thought, making it important to standardize and clearly define them. To achieve clarity and precision, we will rely on the rigorous definitions provided by mathematics. By adopting mathematical definitions, we ensure that our terminology is universally understood and unambiguous.

Metric (norm)

Mathematically, a metric (or norm) is a mapping from a space to real positive numbers set. It's defined as a mapping $\|\cdot\|: V \rightarrow R$ for a space V , satisfying the following properties for all elements $x, y \in V$ and scalar $a \in R$:

1. Non-negativity: $\|x\| \geq 0$
2. Definiteness: $\|x\| = 0 \Leftrightarrow x = 0$
3. Homogeneity: $\|ax\| = |a| \|x\|$
4. Triangle Inequality: $\|x + y\| \leq \|x\| + \|y\|$.

Having introduced the concept of norm it becomes possible to measure “the size” or “the distance between” the elements of the space. Thus, the norm concept is chosen to evaluate how much the noise affects the performance of large language models (LLMs). By adding noise to the data and using the norm to calculate the 'distance' between the original and noisy data, we can see how well the models handle these changes. This helps to understand and compare the robustness of different models in dealing with real-world, imperfect data.

Accuracy

Accuracy in the context of LLMs refers to the closeness of the generated outputs to the true or expected values. Formally, accuracy can be defined as the proportion of correct predictions to the total number of predictions. If N is the total number of predictions and $N_{correct}$ is the number of correct predictions, then

$$Accuracy = \frac{N_{correct}}{N}.$$

Accuracy measures how well a model performs under normal conditions, but it doesn't account for how models perform with noisy or distorted data. In this thesis, we develop a new metric that combines accuracy with robustness against noise. By adding controlled noise to the data and observing changes in accuracy, we can evaluate how resilient the models are. This helps in understanding their reliability in real-world situations where data is often imperfect. Thus, combining accuracy with robustness gives a more comprehensive evaluation of the model's performance.

Uncertainty

Uncertainty quantifies the degree of confidence in the predictions made by an LLM. It is often expressed as a probability distribution over possible outcomes. Mathematically, if vector $p = (p_1, p_2, \dots, p_N)$ represents the probability distribution of n possible outcomes, uncertainty can be measured using entropy

$$H(p) = - \sum_{i=1}^N p_i \log(p_i), \text{ where } 0 \leq H(p) \leq \log(N).$$

High uncertainty indicates that the model is less confident about its predictions, especially when dealing with noisy or perturbed data. By evaluating how uncertainty changes when noise is added, we can measure the model's stability and reliability. A robust model should maintain low uncertainty even with noisy inputs, reflecting consistent confidence in its predictions.

Noise

Noise in the context of data and models refers to random variations or disturbances that obscure the true signal. If y is the observed value, $f(x)$ is the true underlying function, and ϵ is the noise, then:

$$y = f(x) + \epsilon$$

In this context, the noise is additive, as it is simply added to the true signal. However, it is important to note that noise can also be multiplicative, where the disturbances scale the signal.

In this work, noise is viewed positively and is used synonymously with text perturbations, referring to deliberate modifications or disturbances introduced to test the robustness of LLMs.

Robustness

Robustness in machine learning refers to the ability of a model to maintain its performance

when faced with uncertainties or adversarial conditions (Braiek and Khomh, 2024). This includes handling noisy data, distribution shifts, and adversarial attacks. A robust model should generalize well from its training data to new, unseen data, and provide reliable predictions even when dealing with unforeseen inputs or circumstances.

In a broader mathematical context, robustness can also refer to the stability and resilience of a system or model under various perturbations. For instance, in control theory, robustness is concerned with a system's ability to function correctly despite external disturbances or model inaccuracies (Zames and Francis, 1983).

Formally, let f be the function represented by the model, x be the input, and X be a set of perturbations or adversarial conditions. The model is robust if

$$\sup \|f(x) - f(x + \delta x)\| \leq \epsilon, \text{ for } \delta x \in X.$$

By analyzing how well these models maintain their accuracy and stability when subjected to various perturbations, we can better assess their practical utility. This also shows the development of more resilient models that can handle the unpredictable nature of real-world data.

1.2 The Need for a Robustness Evaluation

The quality of generated text is crucial for the effectiveness of LLMs in real-world applications. Texts produced by models need to be grammatically correct and smooth, but also meaningful, consistent, and accurate in terms of the information they convey. Enhancing text generation quality boosts user trust and expands the range of potential applications, including text-sensitive fields such as education, law, and medicine. (See et al., 2019).

The development of a robust evaluation metric specifically tailored for assessing the performance of LLMs in the presence of noise-added data becomes crucial under these

circumstances. Traditional metrics often fail to capture the nuances necessary to evaluate how well these models maintain their performance when faced with noisy or adversarial inputs. This gap in evaluation can lead to an overestimation of a model's capabilities in real-world scenarios (Jia and Liang, 2017).

Recent studies indicate that LLMs can exhibit varying levels of sensitivity to different types of noise, such as typographical errors, syntactical inconsistencies, or semantic ambiguities (Smith et al., 2022). For instance, while some models may handle typographical errors with minimal impact on their output quality, they might struggle significantly with semantic noise, which misleads the models about the intended meaning of the text.

Therefore, the thesis proposes an approach to leverage the benefits of noise usage in a metric (*ProposedMetric*) that evaluates LLM robustness by introducing various forms of noise into the data during testing. The metric aims to provide a more accurate assessment of model performance, focusing on the model's ability to process and understand text effectively, despite the presence of errors or distortions. Such a metric will not only aid in benchmarking current models but also guide future developments in model training, ultimately leading to the creation of more reliable and effective LLMs.

2 Background

In the rapidly evolving field of natural language processing (NLP), large language models (LLMs) have demonstrated significant advancements in text generation capabilities. However, alongside improvements in text quality, an equally crucial aspect is the robustness of these models. In conjunction with improving text quality, there is an equally important aspect known as model robustness. Machine learning robustness pertains to a model's capacity to sustain its performance when encountering uncertainties or adversarial conditions (Brown, Curtis and Goodwin, 2021). This includes managing noisy data, distribution shifts, and adversarial attacks, among other obstacles. A robust model ought to generalize effectively and furnish dependable predictions even in the presence of

unforeseen inputs or circumstances. In numerous instances, there exists a trade-off between model robustness and accuracy (Yang et al., 2020). While aiming for the best accuracy on a specific dataset might seem attractive, it could lead to overfitting or a failure to adapt to new data. Therefore, it's important to assess models from a robustness perspective too. With the rise of generative text models relying on well-written prompts, there's a need for a new benchmark to evaluate model robustness based on text variations.

2.1 Audio dithering

Historically, the concept of noise addition has played a pivotal role in enhancing the robustness and quality of signal processing in fields such as audio and image processing. Techniques like dithering in audio processing involve adding low-level noise to reduce quantization distortions (Stuart and Craven, 2019). Similarly, noise injection in image processing helps combat overfitting in neural networks, making them more resilient to minor variations in input data. These techniques highlight how controlled noise addition can significantly improve the ability to adapt to unknown data without compromising the utility of the original signal.

Audio dithering reduces quantization distortion by adding low-level noise to the signal before quantization. Mathematically, the dithered signal $u_d(t)$ is given by:

$$u_d(t) = u_0(t) + N(t)$$

where $u_0(t)$ is the original signal and $N(t)$ is the noise. The quantized signal $u_q(t)$ is:

$$u_q(t) = Q(u_d(t)) = Q(u_0(t) + N(t))$$

Noise $N(t)$ typically has a zero mean, ensuring it does not bias the signal. This decorrelates quantization error, making it less perceptible.

2.2 Textual Data and LLMs

Drawing an analogy to the traditional use of noise in signal processing, applying noise addition to textual data for training LLMs can enhance the models' adaptability to various linguistic styles, contexts, and intricacies. This approach mimics the diversity and uncertainty of natural language, thereby improving model robustness in text generation. Given the rise of generative text models that rely on well-written prompts, there's a compelling need for a new benchmark to evaluate model robustness based on text variations in user's input. This study focuses on developing a robust evaluation metric that incorporates noise-added data to LLMs prompts. By introducing controlled noise into the questions posed to the model for answering this method challenges the models to maintain high performance even under less-than-ideal conditions, closely mirroring the unpredictable nature of human inputs.

3 Methodology

This section presents the methodology used to develop a robust evaluation metric for large language models (LLMs) under conditions where data is perturbed by noise. The objective is to systematically test the resilience of LLMs to both syntactical and semantic distortions.

Imagine we're trying to read a text message that has a few typos or some confusing word choices. We as humans can still understand it. But for a computer program trying to understand that same message, it's not so easy. This study focuses on evaluating how well large language models (LLMs) can handle messy, "noisy" data.

3.1 Experimental Setup

Model Selection: Pre-trained transformer models from the "transformers" library are chosen for their distinct architectural features and prevalence in NLP tasks.

Dataset: The “commonsense_qa” dataset from Hugging Face datasets is used, particularly its validation set of 1,221 samples. This dataset challenges models to integrate and interpret complex information, making it ideal for testing resilience to noise.

Types of Noise:

- Syntactical Noise: This involves making small mistakes in the text, like typos or misplaced punctuation. For example, turning “I love pizza!” into “I loev pizza!” or “I love pizza”.
- Semantic Noise: Entails replacing key words with synonyms, antonyms, or related terms that could subtly alter the meaning of sentences. For semantic adjustments, ChatGPT is used to generate variations of the same text or idea as if they were written by individuals with English language proficiency levels ranging from A1 to C2.

3.2 Evaluation Protocol

Baseline Performance: How well the models perform with clean, error-free data is measured to see their best-case scenario.

Performance Under Noise: Measure how model performance degrades as noise levels are incrementally introduced.

Proposed Metric: An aggregated metric, or “*ProposedMetric*”, is derived to provide a singular comprehensive view of model robustness across all tests, aiding in understanding each model’s overall resilience to noisy inputs.

Visualization and Analysis: Charts and graphs are used to show how the models’ performance changes as we add more noise. These visual aids help us see patterns and understand how well each model handles the messiness of real-world data.

A significant part of the methodology is the development of an aggregated metric that provides a comprehensive overview of model robustness across all noise tests. This metric aids in evaluating each model’s overall resilience to noisy inputs. Visualisation techniques are employed to graphically represent the relationship between noise intensity and model performance for each type of noise and model. Additionally, a proprietary metric is

calculated to quantitatively assess the robustness of each model in handling both syntactical and semantic noise.

3.3 Tools and Technologies

The study utilizes Python for data manipulation and analysis, with libraries such as Pandas for data handling, Matplotlib and Seaborn for visualization, and Scikit-learn for applying machine learning metrics. The 'transformers' library is used to manage model implementations.

3.4 Expected Outcomes

The expected outcomes of this methodology include a norm (*ProposedMetric*) of robustness that reflects the real-world capabilities of LLMs to process distorted inputs effectively. Insights into each model's ability to handle syntactic and semantic noise will guide future improvements in model training and development. Another important outcome is the showcase of the idea that noise can have benefits; we can utilize it instead of solely focusing on its removal. Also experiments with two types of noise and two text generative models are done to show the usage of metric.

4 Literature review

This section presents a structured literature review. First, papers related to the topic were searched in Web of Science and Scopus databases. Second, a subset was selected based on the criteria discussed below. Third, selected papers are analysed and discussed.

4.1 Papers search

The search strategy is designed to encompass a broad range of studies focusing on the integration of noise into textual data within the context of generative text models, ensuring a comprehensive review of methodologies and outcomes across this niche. By incorporating variations and synonyms related to noise addition and text-based generative models, including advanced architectures like GPT and BERT, the query aims to capture the depth and breadth of current research in the field.

The date when the search was conducted: 01.04.2024.

Four databases were used:

1. Scopus: Known for its broad coverage, Scopus indexes peer-reviewed journals, conference papers, and patents, making it a valuable resource for interdisciplinary research that spans AI, ML, and other fields.
2. Web of science: Web of Science offers a comprehensive index of high-quality, peer-reviewed research in AI and ML. Its citation tracking tools are vital for identifying seminal works and trends.
3. arXiv: A preprint server that provides access to the latest research in physics, mathematics, computer science, quantitative biology, quantitative finance, and statistics, including AI and ML. Usefull for finding cutting-edge research before it's formally published.
4. Papers with code: Papers with Code links academic papers to their source code and provides benchmarks, facilitating the validation and replication of machine learning research. It's made things far easier to read research papers, giving a quick glance at their summary, code implementation, result, and dataset.

4.1.1 Scopus database papers search

First trends in the research domain of NLP and noise injection are explored.

Search query: ("adding noise" OR "noise injection" OR "perturbation" OR "noise augmentation") AND ("text" OR "textual data" OR "text-based") AND ("generative text

models" OR "text generation models" OR "natural language generation" OR "GPT" OR "Transformer models" OR "seq2seq")

41 documents were identified.

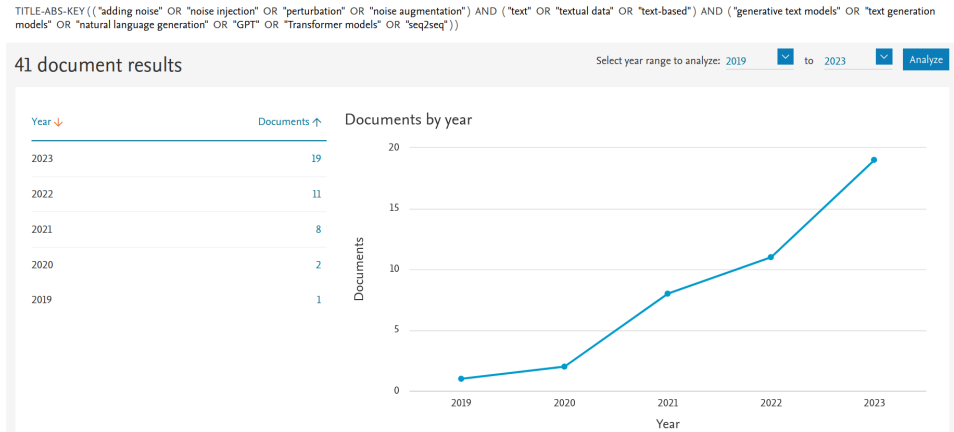


Figure 2. Scopus search results. 41 documents found.

Scopus search results show rising research interest in noise addition for text generation, with publications increasing from 1 in 2019 to 19 in 2023.

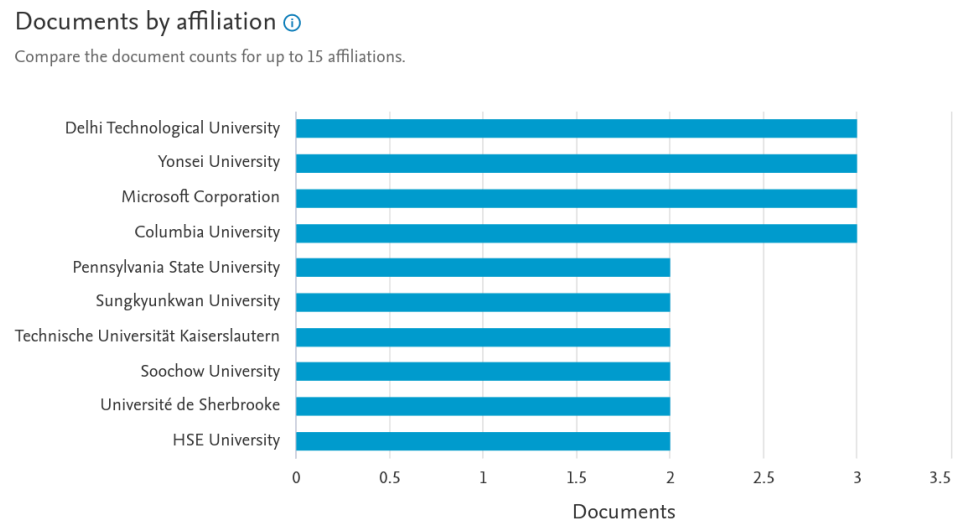


Figure 3. Scopus search documents by affiliation.

Delhi Technological University, Yonsei University, and Microsoft each have three documents, showing that research on noise addition in text data and generative text models is widely distributed, with interest from both academia and major tech companies like

Microsoft. Major number of articles are from USA universities. However, this is different for the papers found in Web of Science.

4.1.2 Web of Science database papers search

Search query: same as in 4.1.1 for Scopus.

10 documents [were found](#).

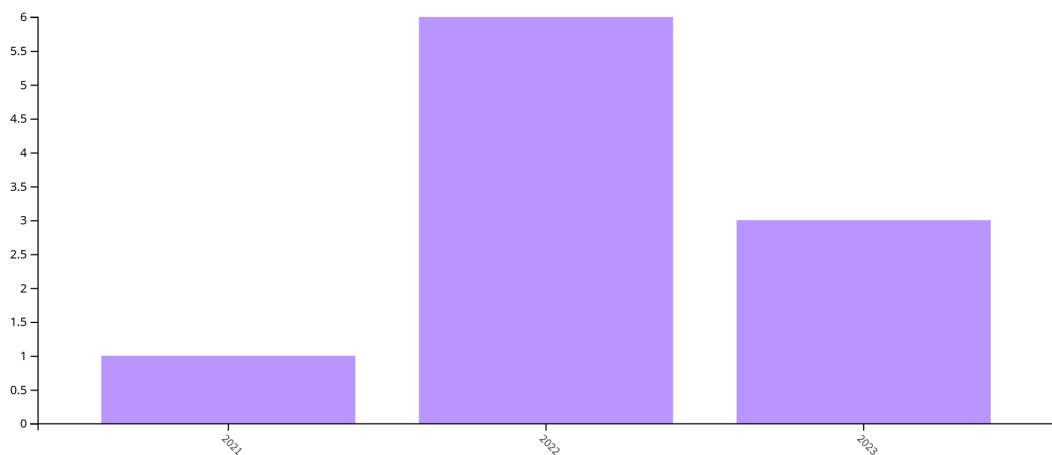


Figure 4. Web of search results. 10 documents found.

Document numbers peaked at six in 2022, dropped to three in 2023, and began with one in 2021, reflecting a growing focus on enhancing NLP model robustness.

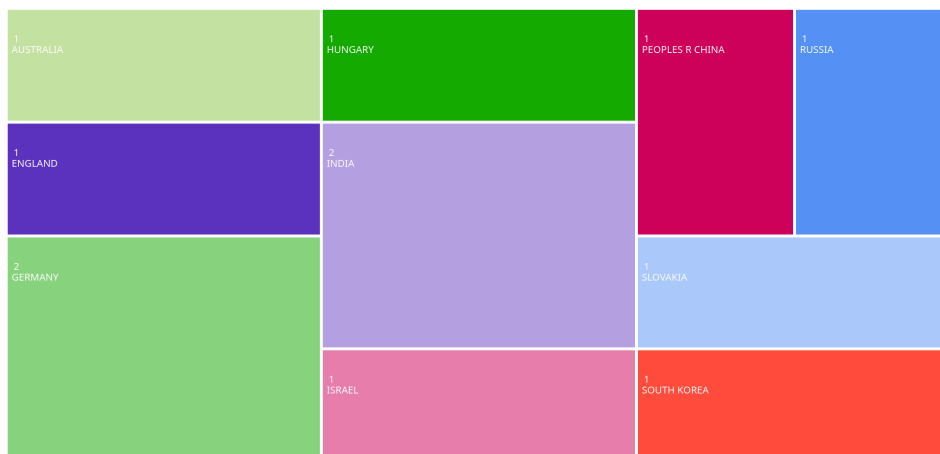


Figure 5. Web of science search documents by affiliation.

Web of Science documents reveal global interest in noise addition for NLP, with India and Germany leading with two documents each. Other contributing countries include Australia, Hungary, China, Russia, England, Israel, Slovakia, and South Korea. Both Web of Science and Scopus show rising interest, with Scopus peaking at 19 documents in 2023.

4.1.3 Other databases papers search

arXiv database papers search

Search query: noise methods in textual data and generative text models

Search query was changed, because arXiv doesn't provide advanced search options like Scopus and Web of Science does.

16 documents were found.

Papers with code database papers search

Search query: noise methods in textual data and generative text models

5 documents were found.

4.2 Papers selection

In total 41 (Scopus) + 10 (Web of Science) + 16 (arXiv) + 4 (Papers with code) = 71 documents found (with duplicates). Afterward, all of them were skimmed, and the abstracts were read to check their correlation to the topic of the current work as well as duplicates were removed.

42 papers were excluded from further analysis because they primarily addressed methods for avoiding noise rather than utilising it, or they were duplicates of other papers with similar content but different titles. Additionally, some papers were deemed irrelevant to the current work's topic.

Most of the selected papers were sourced from the Scopus database. All but one of the papers found in the Web of Science were duplicates of those found in Scopus. Only one

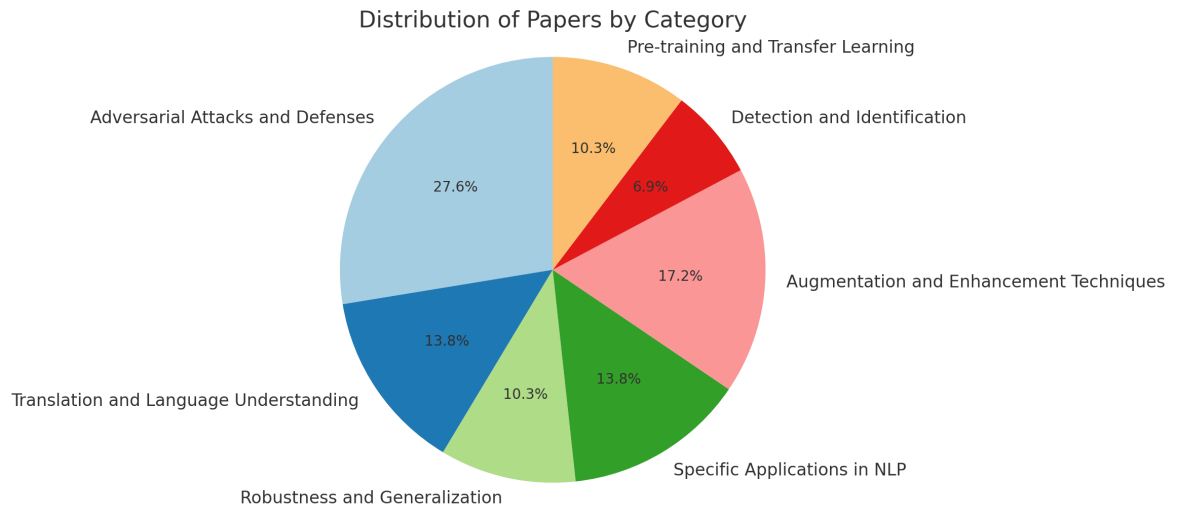


Figure 7. Distribution of 29 Research Papers by Category. This pie chart illustrates the proportional representation of topics covered in the selected research papers. The visualisation highlights the diversity of focus areas, with the largest segments representing "Adversarial Attacks and Defences" and "Augmentation and Enhancement Techniques".

Papers within their category are presented in Table 1. After that each category and related papers is reviewed.

Table 1. Distribution of 29 Research Papers by Category.

№	Category	Subcategory	Paper's title and citation reference
1	A. Adversarial Attacks and Defences	Adversarial Attacks on NLP Models	"A Differentiable Language Model Adversarial Attack on Text Classifiers" (Fursov et al., 2022)
2			"Block-Sparse Adversarial Attack to Fool Transformer-Based Text Classifiers" (Sahar Sadrizadeh, Ljiljana Dolamic and Frossard, 2022)
3			"Efficiently Generating Sentence-Level Textual Adversarial Examples with Seq2seq Stacked Auto-Encoder" (Li et al., 2023)
4		Adversarial Defense Techniques	"Adversarial Machine Learning in Text Processing: A Literature Survey" (Alsmadi et al., 2022)
5			"KTGAT: Improving the Robustness of Knowledge-enhanced Text Generation via Adversarial Training" (Zhu, Song and Liu, 2023)
6			"Adversarial Text Perturbation Generation and Analysis" (Guerrero, Liang and Izzat Alsmadi, 2023)
7			"Iterative Adversarial Attack on Image-guided Story Ending Generation" (Wang, Hu and Hong, 2024)
8			"HOMOCHAR: A Novel Adversarial Attack Framework for

			Exposing the Vulnerability of Text-based Neural Sentiment Classifiers" (Bajaj and Dinesh Kumar Vishwakarma, 2023)
9	B. Translation and Language Understanding	Machine Translation	"Dissecting In-Context Learning of Translations in GPTs" (Vikas Raunak, Menezes and Awadalla, 2023)
10		Semantic Analysis and Perturbations	"Perturbing Inputs for Fragile Interpretations in Deep Natural Language" (Sinha et al., 2021)
11			"Tailor: Generating and Perturbing Text with Semantic Controls" (Ross et al., 2022)
12		Contextual Understanding in Models	"Attention Analysis and Calibration for Transformer in Natural Language Generation" (Lu et al., 2022)
13			"Perturbation CheckLists for Evaluating NLG Evaluation Metrics" (Sai et al., 2021)
14	C. Robustness and Generalization		"On Robustness of Finetuned Transformer-based NLP Models" (Pavan et al., 2023)
15			"Randout-KD: Finetuning Foundation Models for Text Classification via Random Noise and Knowledge Distillation" (Khan, Dengel and Ahmed, 2023)
16			"Robustness Analysis Uncovers Language Proficiency Bias in Emotion Recognition Systems" (Tran et al., 2023)
17	D. Specific Applications in NLP	Healthcare and Biomedical Applications	"A Novel Approach to Train Diverse Types of Language Models for Health Mention Classification of Tweets" (Pervaiz Iqbal Khan et al., 2022)
18		Security and Privacy	"Nomen est Omen - The Role of Signatures in Ascribing Email Author Identity with Transformer Neural Networks" (Srinivasan et al., 2021)
19		Database and Information Retrieval	"MIGA: A Unified Multi-task Generation Framework for Conversational Text-to-SQL" (Fu et al., 2023)
20		Sentiment Analysis	"Bypassing Deep Learning-based Sentiment Analysis from Business Reviews" (Bajaj and Dinesh Kumar Vishwakarma, 2023a)
21	E. Augmentation and Enhancement Techniques	Enhancing Text Generation	"Diversity Regularized Autoencoders for Text Generation" (Ko et al., 2020)
22			"Differentially-Private Text Generation via Text Preprocessing to Reduce Utility Loss" (Sasada et al., 2021)
23			"FRSUM: Towards Faithful Abstractive Summarization via Enhancing Factual Robustness" (Wu et al., 2022)
24		Visual and Textual Data Integration	"TextManiA: Enriching Visual Feature by Text-driven Manifold Augmentation" (Ye-Bin et al., 2023)
25			"TEXTFIELD3D: Towards Enhancing Open-Vocabulary 3D Generation with Noisy Text Fields" (Huang et al., 2023)
26	F. Detection and Identification		"DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature" (Mitchell et al., 2023)
27			"Learning Representations through Contrastive Strategies for a

			More Robust Stance Detection" (Rajendran and Trabelsi, 2023)
28	G. Pre-training and Transfer Learning	Pre-training Approaches	"ReCode: Robustness Evaluation of Code Generation Models" (Wang et al., 2023)
29		Language Model Analysis	"Morphosyntactic Probing of Multilingual BERT Models" (Judit Ács et al., 2023)

A. Adversarial Attacks and Defences

a. Adversarial Attacks on NLP Models

The study by Fursov et al. (2022) look at how text classification models can be tricked by adversarial attacks. It show that small, carefully designed changes to text can lead to wrong predictions by machine learning models. The main achievement is developing a method to change text inputs using gradient-based optimization, making attacks on language models more effective. They also created a metric to measure how robust models are against these attacks and random noise.

In another study, Sadrizadeh, Dolamic, and Frossard (2022) examine how transformer-based models can be attacked by exploiting their structure. They present a block-sparse attack method that alters blocks of data to cause misclassifications. This method takes advantage of how transformers process data in segments using attention mechanisms. The study find that although transformers are generally robust, they are particularly weak against structured adversarial inputs. This emphasizes the need for specific metrics to measure their weaknesses.

Lastly, Li et al. (2023) focus on creating sentence-level adversarial text examples using a Seq2seq model combined with a stacked auto-encoder. This method keep the original text's meaning while causing misclassifications, effectively fooling advanced NLP models. The research show the potential of using deep learning architectures to create subtle but effective adversarial examples, highlighting the need for evaluation methods that can detect and resist such attacks.

b. Adversarial Defense Techniques

The literature survey by Alsmadi et al. (2022) offers a comprehensive review of adversarial techniques developed to challenge text processing systems, covering a spectrum from simple character-level alterations to complex semantic changes. This paper catalogs these techniques and discusses their effectiveness across various NLP tasks like sentiment analysis, text classification, and machine translation, highlighting the necessity for ongoing development of robust defense mechanisms and standardized benchmarks for evaluating NLP model resilience.

Zhu, Song, and Liu (2023) explore the application of adversarial training in enhancing the robustness of text generation models through their Knowledge-enhanced Text Generation Adversarial Training (KTGAT). This method integrates external knowledge into the text generation process and subjects the model to adversarial examples during training, significantly enhancing the model's ability to produce coherent and contextually appropriate text under adversarial conditions. This is vital for ensuring the reliability of text generation models in varied and potentially adversarial environments.

Guerrero, Liang, and Izzat Alsmadi (2023) analyze the creation and impact of subtle text perturbations designed to deceive NLP models. Their research systematically evaluates how minor changes like synonym replacement, word reordering, or grammatical distractions can severely impact model performance, providing insights into the specific vulnerabilities of current systems and guiding the development of more robust NLP models.

Wang, Hu, and Hong (2024) extend adversarial attack methodologies to the multimodal domain, where both text and images are manipulated to mislead models into generating incorrect or irrelevant story endings. This study shows that even advanced multimodal models are vulnerable to precise, systematic input modifications, underscoring the need for improved security measures in systems used for creative or narrative AI applications.

Bajaj and Dinesh Kumar Vishwakarma (2023) introduce HOMOCHAR, an adversarial attack framework that tests the resilience of sentiment analysis models by substituting characters with homoglyphs—characters that appear similar but are recognized differently by computer systems. This method effectively exposes how minor textual changes can disrupt sentiment classifiers, which is crucial for applications that depend on accurate sentiment analysis to avoid significant misinterpretations of user feedback or market data.

B. Translation and Language Understanding

a. Machine Translation

The study by Vikas Raunak, Menezes, and Awadalla (2023) investigates the capabilities of Generative Pre-trained Transformers (GPTs) for language translation without specific training for this task, a method known as zero-shot learning. Unlike conventional machine translation systems that rely on large datasets of bilingual sentence pairs, this research examines whether GPTs can leverage their extensive language understanding from general training to perform translations across languages they have not been explicitly trained on.

The findings suggest that GPTs are indeed capable of executing translations by utilizing the context provided in the input text, with the quality of these translations varying based on the complexity of the text and the languages involved. This ability of GPTs to adapt to translation tasks without direct training illustrates their potential flexibility and utility in scenarios featuring diverse and unforeseen linguistic inputs.

b. Semantic Analysis and Perturbations

10. "Perturbing Inputs for Fragile Interpretations in Deep Natural Language" (Sinha et al., 2021)

The study investigates how slight modifications to the input of deep natural language processing models can lead to significant misinterpretations or misclassifications. By introducing subtle perturbations such as synonym replacements, rephrasing, or inserting

small grammatical errors, researchers observed a disproportionate impact on the outputs of these models. These findings highlight the models' sensitivity to changes that might be minor or undetectable to human readers.

Experiments revealed that robust natural language models often fail to maintain accuracy when faced with these minor input perturbations. The tests included a variety of models and showcased a broad range of NLP tasks, demonstrating the widespread nature of this vulnerability across different types of neural networks and applications. This sensitivity to input perturbations presents a critical challenge for deploying NLP models in real-world scenarios, where input variability is common.

11. "Tailor: Generating and Perturbing Text with Semantic Controls" (Ross et al., 2022)

Focuses on a model that can alter text based on specified semantic parameters while maintaining the original context and message. The model manipulates aspects like sentiment or topic within the text, ensuring the modifications preserve the original intent. Practical applications demonstrated include customized content creation for automated journalism and targeted marketing messages, where maintaining semantic consistency is crucial. The model's effectiveness in adhering to predefined semantic parameters is validated through various controlled experiments showcasing its capacity to generate content that meets specific semantic requirements.

c. Contextual Understanding in Models

The study by Lu et al. (2022) focuses on the critical role of attention mechanisms within transformer models used in text generation. The research investigates how the distribution of attention affects the coherence and relevance of the text produced. The researchers propose a method for the calibration of attention weights to enhance the model's ability to concentrate on more pertinent parts of the input data, thereby improving the overall quality of the generated text. This calibration not only refines the precision of language generation but also helps minimize the production of irrelevant or repetitive content. The methodology employed involves meticulous tracking and adjustment of attention weights

across different layers of the transformer, grounded in an empirical analysis of the model's performance in various text generation tasks.

Sai et al. (2021) introduce a structured approach to evaluate the robustness of natural language generation (NLG) evaluation metrics using perturbation checklists. These checklists are crafted to methodically introduce a range of textual perturbations, simulating typical errors and variations encountered in natural language processing. The objective is to determine how effectively different NLG evaluation metrics can detect and adapt to these changes. The study categorizes perturbations into syntactic variations, semantic shifts, and pragmatic alterations, each designed to challenge the NLG metrics differently. This examination reveals the performance variability among popular NLG metrics, identifying specific strengths and weaknesses in handling different text modifications. This method provides a detailed framework for assessing the reliability and responsiveness of NLG evaluation tools in realistically challenging scenarios.

C. Robustness and Generalization

In their study, Pavan et al. (2023) look at how different fine-tuning methods impact the effectiveness of transformer models in handling noise and adversarial examples in NLP tasks. They find that while these models perform well on standard tests, their ability to cope with unfamiliar, noisy data decreases. By experimenting with different training setups, they identify techniques that keep the models robust without hurting their accuracy on regular tasks. This suggests that careful fine-tuning is crucial for preparing models to handle both expected and unexpected input scenarios effectively.

Khan, Dengel, and Ahmed (2023) investigate a new method called Randout-KD to improve text classification models by injecting random noise and using knowledge from a more extensive, pre-trained model. This technique enhances the models' ability to generalize and remain robust, showing notable improvements over traditional training methods. This method proves especially useful in preparing models to perform well in varied and unpredictable environments.

Tran et al. (2023) explore how emotion recognition systems perform when analyzing inputs from non-native speakers. Their findings indicate a significant bias, as these systems, trained mostly with data from native speakers, fail to accurately detect emotions from non-native speakers. This study highlights the importance of including diverse linguistic data in training to reduce bias and improve the fairness and accuracy of emotion recognition systems across different language proficiency levels.

D. Specific Applications in NLP

a. Healthcare and Biomedical Applications

The study by Pervaiz Iqbal Khan et al. (2022) introduces a training method for language models to identify health-related information in tweets. This method combines supervised learning, where models learn from labeled data, with unsupervised learning, which doesn't use specific labels, allowing the model to adapt to informal and varied language on social media. In their experiments, models were trained using a large dataset of tweets annotated with health-related terms such as symptoms, diseases, and medications. This hybrid training approach enhanced the models' ability to detect and classify health mentions accurately, improving both precision and recall. This capability supports monitoring public health trends and concerns on platforms like Twitter.

b. Security and Privacy

The study by Srinivasan et al. (2021) investigates how email signatures affect the ability of transformer-based models to correctly identify the authors of emails. Using the Enron email dataset, the researchers found that changes to or removal of signatures from emails significantly decrease the accuracy of these models, particularly with shorter emails. This outcome suggests that signatures are a crucial factor these models use to determine authorship. Additionally, the impact of removing signatures was more pronounced in shorter emails than in longer ones, indicating that signatures are particularly important for author identification in emails with less text. These results highlight the need for training

improvements in models to better manage variations in email input data, such as signatures.

c. Database and Information Retrieval

The study by Fu et al. (2023) focuses on improving conversational text-to-SQL conversion through a two-stage framework that incorporates multi-task pre-training followed by fine-tuning. This approach uses generative pre-trained models such as T5 and introduces specific SQL perturbations during the fine-tuning phase to minimize errors carried over from previous SQL outputs.

During the pre-training stage, the MIGA framework breaks down the primary task into several sub-tasks, including SQL Generation, Related Schema Prediction, Turn Switch Prediction, and Final Utterance Prediction. These sub-tasks are integrated within a sequence-to-sequence framework, each accompanied by task-specific prompts. The framework is designed to enhance the accuracy of SQL query generation in interactive settings by simultaneously addressing various elements of the SQL generation process, focusing on both the structural integrity and the continuity of the content in SQL queries.

d. Sentiment Analysis

The study by Bajaj and Dinesh Kumar Vishwakarma (2023a) examines the susceptibility of sentiment analysis models to adversarial attacks. The research utilizes deep learning models such as BERT, LSTM, and Word-CNN, which were trained on the Yelp polarity dataset, to demonstrate how strategic textual manipulations can deceive these models. The researchers calculated Attack Success Rates (ASR) to measure the effectiveness of various adversarial techniques on these models. Results showed significant variability in the models' resistance to these attacks.

The study found that the effectiveness of adversarial techniques varies among different models, with some being more vulnerable to specific types of attacks. Additionally, the

research pointed out how certain adversarial methods can alter the predictions of these sentiment analysis models without changing the text's meaning as perceived by human readers. This highlights critical vulnerabilities in current sentiment analysis models, emphasizing the need for more robust defenses against such adversarial attacks.

E. Augmentation and Enhancement Techniques

a. Enhancing Text Generation

Ko et al. (2020) developed a technique to enhance the diversity of text generated by autoencoders. They introduced a diversity-promoting regularization term during the training process. This term encourages the generation of a wider variety of textual outputs, helping to avoid repetitive or overly similar text. The effectiveness of these diversity regularized autoencoders was tested against standard autoencoders lacking such regularization. The results indicated that the regularized models produced text with greater diversity while maintaining relevance and coherence compared to the input, validating the regularization approach across various datasets.

Sasada et al. (2021) explored improving privacy in text generation processes by incorporating differential privacy techniques during the text preprocessing phase. This method aims to reduce the utility loss typically associated with implementing differential privacy, which can degrade the quality and usability of generated text. The methodology involves preprocessing input data with noise injection to ensure data privacy before it is used to train text generation models. Experimental results showed that this preprocessing technique diminishes the impact on text quality compared to applying differential privacy directly in the training phase. This approach enables the production of text that retains higher usefulness while maintaining data privacy, providing a framework for balancing data privacy and text generation quality effectively.

Wu et al. (2022) introduced the FRSUM model to improve the accuracy and trustworthiness of abstractive text summarization. This model focuses on enhancing factual robustness to ensure that summaries closely align with the content of the original text

without introducing common inaccuracies or distortions. The FRSUM model includes a fact-checking module that cross-verifies facts in the generated summary against the source document, prioritizing factual consistency in the summarization process. This approach helps ensure the reliability of abstractive summarization methods by maintaining factual integrity.

b. Visual and Textual Data Integration

Ye-Bin et al. (2023) introduce "TextManiA," an innovative approach to enhance the integration of textual information with visual data for improving machine learning models. This method uses text-driven manifold augmentation to enrich visual features with contextual data derived from text, creating a more detailed feature space. The process involves building a manifold where textual descriptions shape and refine the visual features extracted from images, aiming to reflect the natural association between text and visual perceptions. This enhancement is particularly beneficial for tasks requiring an in-depth understanding of both textual and visual information, such as image captioning or visual question answering. The results indicate that models using TextManiA show improved performance in tasks involving complex text-image interactions, highlighting the method's effectiveness in fostering a more sophisticated multimodal system.

Huang et al. (2023) present a method in "TEXTFIELD3D" for generating 3D models from textual descriptions that may include noise or incomplete information. The innovation lies in using noisy text fields as inputs, prompting the model to interpret and rectify inaccuracies during the model generation process. This method capitalizes on advancements in natural language processing and computer vision to handle diverse textual inputs and generate accurate 3D models that closely match the descriptions provided. The approach involves parsing and cleansing noisy text inputs before they guide the 3D modeling process, ensuring that the final models are true to the original specifications. This technique demonstrates the potential to enhance open-vocabulary 3D generation, making it possible to create detailed and accurate models from a wide array of textual descriptions.

F. Detection and Identification

Mitchell et al. (2023) developed "DetectGPT," a new method for identifying machine-generated text without using labeled data. This technique analyzes the probability distribution of text produced by large language models such as GPT to pinpoint distinctive patterns that indicate machine generation. DetectGPT focuses on the probability curvature of the text, which shows abrupt changes or anomalies in the probability distribution, allowing the model to differentiate between human-written and machine-generated content. This zero-shot method does not require labeled data for training, making it highly useful in situations where annotated datasets are limited or not available.

Rajendran and Trabelsi (2023) propose a stance detection method that utilizes contrastive learning strategies to enhance the robustness of representation learning. Stance detection aims to determine the attitude or viewpoint conveyed in a text towards a specific subject. This method improves the model's ability to handle input variations and noise by training it to distinguish between similar and dissimilar text pairs, each representing different stances on a topic. By training the model to recognize these subtle differences, it can more effectively capture nuanced variations in stance expression, thus improving its robustness against inconsistencies in language usage and context.

G. Pre-training and Transfer Learning

a. Pre-training Approaches

Wang et al. (2023) discuss "ReCode," a framework created to assess the robustness of code generation models. ReCode is essential for determining the reliability and performance stability of models that generate code based on natural language descriptions or other inputs. The framework aims to develop detailed evaluation metrics and methods that cover various dimensions of model robustness.

ReCode evaluates several critical aspects of code generation models. These include the sensitivity of the models to changes in input, the effects of noise on the quality of the

generated code, and the consistency of the model outputs over different attempts or data sets. By thoroughly examining these factors, ReCode offers a comprehensive overview of a code generation model's robustness, helping researchers and developers pinpoint potential weaknesses and suggesting areas for enhancement.

b. Language Model Analysis

Judit Ács et al. (2023) explore the capabilities of multilingual BERT models using morphosyntactic probing, which assesses how effectively these models understand syntactic and morphological details across different languages. Morphosyntactic probing involves creating diagnostic tasks to test the model's grasp of linguistic phenomena like word order, agreement, and morphological inflections.

The study involves experiments in various languages to measure how multilingual BERT models perform on these probing tasks. Through detailed analysis of the models' predictions and their performance metrics on these tasks, the researchers can determine the linguistic strengths and weaknesses of these models, pinpointing the specific areas where the models perform well or where they need improvement.

4.4 Summary of reviewed insights

In the papers review section of the thesis, a thorough analysis of 29 significant studies from the field of NLP is presented. The research explored in these papers primarily concentrates on adversarial attacks, where researchers demonstrate how minor, intentional changes to text inputs can deceive models into making errors. Additionally, several papers investigate the resilience of models against these attacks, aiming to strengthen them against such vulnerabilities.

A notable observation from the review is that the majority of research focuses on the negative impacts of noise in texts, such as how noise can degrade the performance of language models in tasks like text classification and sentiment analysis. The exploration of

beneficial aspects of noise - how it might be used to enhance model robustness or performance - is less common and typically not the main focus of these studies.

The potential benefits of noise in training more robust models are recognized as a promising area of research. There remains a research gap in research of benefits that noise in text can have. Existing studies often focus on overcoming noise rather than leveraging it to test and enhance model resilience. This thesis addresses this gap by proposing a new metric for evaluating model robustness that specifically incorporates noise in a controlled manner.

This insight is important as it aligns with the thesis's objective to develop a new metric for evaluating model robustness by incorporating noise in a controlled manner to test and enhance model resilience. This approach is novel, as the existing literature often treats noise as a challenge to overcome rather than as a tool that could potentially improve model performance when used appropriately. Thus, this thesis aims to shift the perspective on noise from a purely negative factor to a beneficial one in the context of building stronger, more reliable models.

5 Proposed approach

In the presented research, the idea of using noise or test perturbations in a beneficial way is presented. As an illustration the metric (referred as *ProposedMetric*) is suggested to assess the robustness of large language models (LLMs) by incorporating both orthographic and semantic noise into the testing process. The core of this approach is the systematic injection of noise into a standard dataset used for model evaluation. The CommonsenseQA dataset, known for its reliance on contextual reasoning, will be employed. It consists of questions that require understanding of everyday scenarios, making it an ideal candidate for testing model performance under distorted input conditions.

5.1 Dataset

For evaluation of models robustness questions from a dataset called CommonsenseQA is used, which is designed to test how well someone can use common sense to answer questions. Dataset is proposed in (Talmor et al., 2019). Unlike other datasets that focus on using specific information from texts, CommonsenseQA requires a broader understanding of general knowledge. The dataset is unique because it uses ideas from ConceptNet, a resource that links different concepts together. It includes multiple-choice questions created by people known as crowd-workers. These questions involve a main idea and several related ideas. The questions are designed to be challenging and require understanding beyond simple facts.

The creation of the dataset leverages ConceptNet (Speer, Chin and Havasi, 2018), a knowledge graph that links various concepts through common relations. From ConceptNet, several related concepts are selected around a central source concept. Crowdworkers then use these concepts to craft questions that not only mention the source concept but require differentiation among the related concepts. This design aims to test the respondent's deeper understanding and reasoning, going beyond straightforward question-answer formats.

CommonsenseQA contains 12,247 multiple-choice questions, each intricately designed to demand more than simple recall of facts split into train, test and validation (Figure 8). For the *ProposedMetric* 1221 questions from validation dataset were used. Because it's reasonable amount of questions for metric computation.

```

train: Dataset({
  features: ['id', 'question',
            'question_concept', 'choices', 'answerKey'],
  num_rows: 9741
})
validation: Dataset({
  features: ['id', 'question',
            'question_concept', 'choices', 'answerKey'],
  num_rows: 1221
})
test: Dataset({
  features: ['id', 'question',
            'question_concept', 'choices'],
  num_rows: 1140
})

```

Figure 8. CommonsenseQA dataset questions number and fields.

Dataset consists of multiple options questions.

Task: The task involves selecting the most appropriate answer among multiple choices given a question about a common scenario.

Dataset metric: Evaluation is typically done using accuracy, measuring the percentage of questions for which the model provides the correct answer.

Examples of questions:

1. What do people aim to do at work?

Options: “complete job”, “learn from each other”, “kill animals”, “wear hats”, “talk to each other”

Correct option: complete job

2. Where are you likely to find a hamburger?

Options: “fast food restaurant”, “pizza”, “ground up dead cows”, “mouth”, “cow carcass”

Correct option: fast food restaurant

3. Crabs live in what sort of environment?

Options: “maritime”, “bodies of water”, “saltwater”, “galapagos”, “fish market”

Correct options: saltwater

5.2 Types noise

There are several types of noise or text perturbations:

1. Orthographic Noise. Typos or variations in word spelling. This can help the model better handle errors in the source data and be more resilient to such errors when generating text.
2. Semantic Noise: Adding words or phrases that alter or expand the semantic content of the text. This may include synonyms, antonyms, or even contextually appropriate nonsense that prompts the model to pay closer attention to the meaning of words and phrases.

3. Syntax Noise: Minor syntax errors or changes that can mimic different styles and sentence structures. For example, intentionally altering word order or using alternative sentence constructions.
4. Data Variability Noise: Changing styles, dialects, or even languages in text data. This adds additional complexity for the model, training it to handle the variability of language use in the real world.
5. Latent space Noise: Controlled randomness in latent space (encoded representation) can lead to a new type of information retrieved at the end by Decoder.

For the current work two types of text perturbations were used - Orthographic Noise and Semantic Noise. Two types is enough to show the idea of noise usage. In future work more types of noise can be explored.

5.2.1 Orthographic Noise

Based on the random variable each letter in the question has the possibility to change to another random character. Randomness is determined by the variable that refers to the alteration chance – alpha. Example of text perturbation for the different alpha is presented on Figure 9.

```
alpha = 0, sentence = two roads diverged in a yellow wood
alpha = 0.005, sentence = two roads diverged vn a yellow wood
alpha = 0.01, sentence = two roads diverged in a yellow wood
alpha = 0.03, sentence = two roads diverged in a yelzow wood
alpha = 0.05, sentence = two roads diverged in a yellox wood
alpha = 0.08, sentence = two roads divdrged in a dellowVwood
alpha = 0.1, sentence = two rogds diverged in a yellow wood
alpha = 0.2, sentence = two rmads diveugedGxn aHyfleowFwood
alpha = 0.3, sentence = twe moadd tmoerwwdVcn nJeellcv wopf
alpha = 0.4, sentence = xtaMroads diveoqeb ix a qvclolCwwod
alpha = 0.5, sentence = tqoQriadgHkiuegrrd jsIa wellww wolo
alpha = 1, sentence = dylIkdyxk0odxidvavFahGbJuovfiyJogfy
```

Figure 9. Text perturbation for the different alpha.

5.2.2 Semantic Noise

Semantic noise is designed by usage of chatGPT (ChatGPT, 2024). The GPT-4 model is prompt to rephrase the same questions as presented by individuals with varying English proficiency levels, ranging from A1 to C2. It is important that the model is asked to keep the meaning of initial question.

Example of text perturbations using Semantic Noise:

Original question: "A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?"

A1: "A turning door is good for going both ways. It also keeps a place safe. Where?"

A2: "A revolving door is useful for going in and out. It also helps keep a place secure. At what kind of place?"

B1: "A revolving door is convenient for entering and exiting in both directions, and it also acts as a safety feature. In what kind of place?"

B2: "A revolving door allows easy access and exit while serving as a security feature. Where is it typically used?"

C1: "A revolving door facilitates bidirectional entry and exit, simultaneously serving as a security mechanism. In what context is it employed?"

C2: "A revolving door offers the dual benefits of streamlined ingress and egress, alongside functioning as a safeguard. Within what establishment is it customarily implemented?"

5.3 ProposedMetric calculation

To calculate metric, simple linear combination proposed as a formula. It's fast to calculate as well as it captures the model's differences effectively.

$$ProposedMetric = \lambda \sum_{i=1}^n \frac{acc_i}{n},$$

where

- n is the number of data points,
- acc_i is model's accuracy,
- λ is scaling coefficient, equal to 1000 in presented experiments in Section 6.

6 Experiments

For testing Proposed metric on 1221 questions from CommonsenseQA dataset with text perturbations two generative models from Python transformers library were used: facebook/bart-large-mnli and , nli-deberta-v3-small. Let's take a closer look at them.

1. facebook/bart-large-mnli (2019)

The BART-large model is explored, specifically the version that was fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset and hosted on the Hugging Face Model Hub under the identifier "facebook/bart-large-mnli." This model is built on the BART architecture, which stands for Bidirectional and Auto-Regressive Transformers. BART is notable for its design, which incorporates both a bidirectional encoder (similar to the one used in BERT) and an autoregressive decoder (similar to the one used in traditional sequence-to-sequence models like those found in machine translation).

The BART-large model with MNLI fine-tuning is designed to perform tasks that require understanding the relationships between sentences, which is fundamental to natural language inference (NLI). NLI tasks involve determining whether a premise sentence can imply, contradict, or remain neutral with respect to a hypothesis sentence. The MNLI dataset, which is used to fine-tune this model, includes a wide range of genres of spoken and written text, and is designed to help the model generalise across different forms of language and various contexts. (Yin, Hay and Roth, 2019)

2. nli-deberta-v3-small (2020)

We also investigate the use of the cross-encoder/nli-deberta-v3-small model, available on the Hugging Face Model Hub. This model is built on the DeBERTa (Decoding-enhanced BERT with disentangled attention) architecture.

The DeBERTa model introduces several innovative techniques, including disentangled attention mechanisms that separately model the content and position of each word in a sentence, and an enhanced mask decoder that improves the model's understanding of context and word relationships. These features allow DeBERTa to outperform other models on benchmarks that require deep understanding of text semantics, such as natural language inference (NLI).

The specific model we use, cross-encoder/nli-deberta-v3-small, is fine-tuned on multiple NLI datasets. In NLI tasks, the model evaluates the relationship between two sentences - typically categorized into entailment, contradiction, or neutrality. The cross-encoder setup involves processing both the premise and hypothesis together in a single forward pass, which allows for more comprehensive interaction between the input pair compared to bi-encoder models where the inputs are encoded separately.

For each type of noise both models are tested and *ProposedMetric* is calculated after that (Fig. 10 and 11). Each dot represents accuracy on 1221 questions from the CommonsenseQA dataset. Initially text perturbations was performed to each question and dataset for models testing is created ¹. The Orthographic noise alpha step is smaller, around 0, to capture small variations and changes up to 0.2, as beyond that point, the question becomes too noisy and the meaning is lost. For Semantic noise, all levels of English language are represented, from A1 to C2.

¹ Dataset and notebooks for experiments could be found in <https://github.com/seriozh1/metric-robustness-noise>

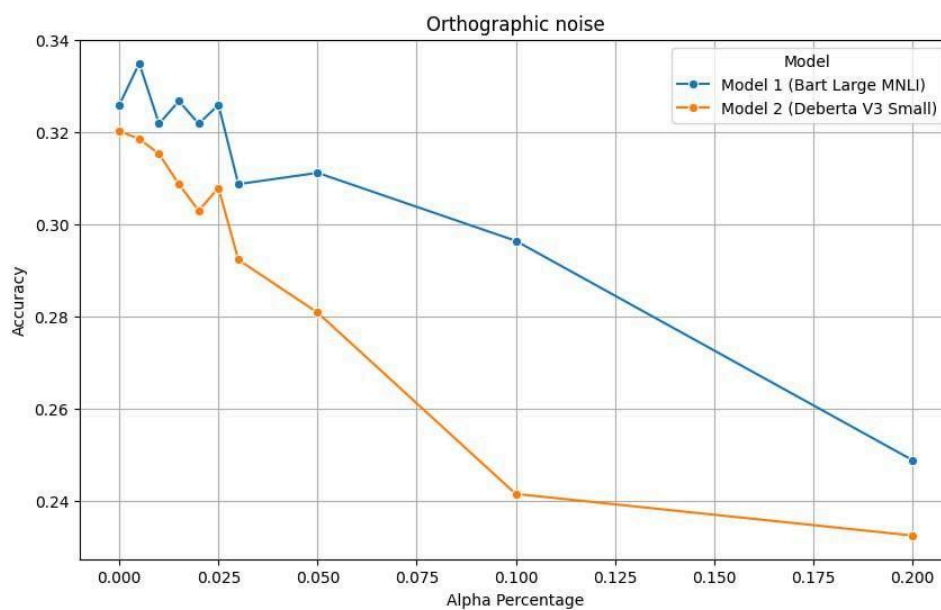


Figure 10. Accuracy calculation for the Orthographic type of noise for two models.

The second model, DeBERTa, was proposed and developed one year after BART, achieving better results as a result of its improvements. BART as presented by model's developers in their paper (Lewis et al., 2019) archives 84.3 accuracy on MNLi dataset, while DeVERTa archives 91.9 (HuggingFace, 2020). However, our experiments show that from the robustness point of view BART is slightly better. Accuracy degradation is smaller for the BART model (Fig. 10) and expected in real world scenarios to respond better on human inputs.

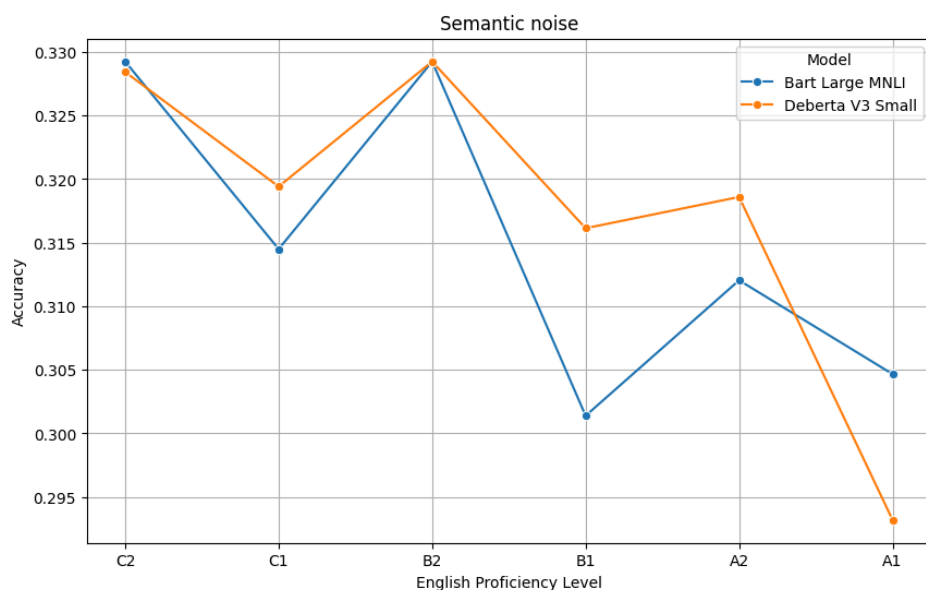


Figure 11. Accuracy calculation for the Semantic type of noise for two models. For the chatGPT or Semantic variations both models show similar behaviour.

Using the formula for *ProposedMetric* described in Section 5.4 we got:

1. facebook/bart-large-mnli (2019)

ProposedMetric = 313

2. nli-deberta-v3-small (2020)

ProposedMetric = 301

Thus, the section bears the main result of the study, it's proposed and demonstrated how *ProposedMetric* can be applied. Using the approach other researchers can also try to evaluate other models to assess robustness.

7 Limitations

ProposedMetric utilises different types of text perturbations to assess generative text models robustness. Two types of noise and two types of generative models were tested. This brings us to limitations in the number of noise and generative models used. The presented results suffice to illustrate the proposed approach of utilising text perturbations. However, additional types of noise, as described in Section 5.3, could be integrated into the metric. Moreover, for the existing types of text perturbations, alternative approaches could be explored. For instance, different alpha values could be employed for Orthographic noise, and new proficiency levels of English could be introduced for Semantic noise. Instead of using only ChatGPT for text variations, other text generative models such as Llama3 (Meta, 2024) or Claude (Anthropic, 2024) could be employed. Additionally more advanced formulas for metric calculation could be developed, for example using nonlinear functions, integrals instead of addition.

8 Future work

The proposed approach could be enhanced by incorporating additional types of text perturbations, improving those already used, or developing new formulations for calculating the *ProposedMetric* as described in the Limitations section.

Also text perturbation can be used in the training process of generative text models. In several experiments, it is observed that small variations in the data led to an increase in the prediction power of models. However, it remains a question whether this was a random occurrence or a trend.

9 Discussion

Nowadays models are improved by scaling - adding more training data, more epochs in training, more training time. However, during the use of generative models internet, as a

source for new human created data, will be polluted with artificial data created by models like chatGPT. At some point, we will reach a stage where we have gathered information from all available sources and the newly created information will no longer be of the same quality as before. Research shows that models break down and degrade when fine-tuned or trained on syntactic data. (Bertrand et al., 2023). But with the proposed method, we may be able to create new data that is only slightly different from the original data, but still provides value for image extensions of image models. Therefore, focusing on the training process seems to be a promising avenue for future work.

10 Conclusion

This thesis developed a new method to measure how well large language models handle messy data. The measurement tool, called *ProposedMetric*, uses both simple spelling mistakes and meaning changes in text to test the models. This helps us understand how these models might perform in real-world situations where data isn't perfect.

The experiments, which involved adding different kinds of noise to a dataset known as CommonsenseQA, showed that this method could effectively measure how different models cope with errors in the data. The metric was applied to two specific models, and the results were clear and helpful.

The *ProposedMetric* is valuable not only for evaluating model performance but also for guiding future improvements. By using this metric, researchers can see where a model struggles with noise and make adjustments to improve it. This could lead to models that are better at understanding and processing real-world, imperfect data.

This method could be highly beneficial for the research community. It offers a way to test and improve models more thoroughly. Also, using the ProposedMetric during the training phase of model development might help make models naturally better at handling noisy data.

In conclusion, the development of this new metric could significantly help in making language models more reliable for all kinds of applications, from automated chatbots to systems that help doctors with medical records.

References

Alsmadi, I., Aljaafari, N., Nazzal, M., Alhamed, S., Sawalmeh, A.H., Vizcarra, C.P., Khreishah, A., Anan, M., Algozaibi, A., Al-Naeem, M.A., Aldalbahi, A. and Al-Humam, A. (2022). Adversarial Machine Learning in Text Processing: A Literature Survey. *IEEE Access*, 10, pp.17043–17077. doi:<https://doi.org/10.1109/access.2022.3146405>.

Anthropic (2024). *Introducing the next generation of Claude*. [online] www.anthropic.com. Available at: <https://www.anthropic.com/news/claude-3-family>.

Bajaj, A. and Dinesh Kumar Vishwakarma (2023a). Bypassing Deep Learning based Sentiment Analysis from Business Reviews. *ViTECoN 2023 - 2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, Proceedings*, (2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, ViTECoN 2023). doi:<https://doi.org/10.1109/vitecon58111.2023.10157098>.

Bajaj, A. and Dinesh Kumar Vishwakarma (2023b). HOMOCHAR: A novel adversarial attack framework for exposing the vulnerability of text based neural sentiment classifiers. *Engineering applications of artificial intelligence*, 126, pp.106815–106815. doi:<https://doi.org/10.1016/j.engappai.2023.106815>.

Bender, E., McMillan-Major, A., Shmitchell, S. and Gebru, T. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. [online] doi:<https://doi.org/10.1145/3442188.3445922>.

Bertrand, Q., Bose, A.J., Duplessis, A., Jiralerspong, M. and Gidel, G. (2023). *On the Stability of Iterative Retraining of Generative Models on their own Data*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.2310.00429>.

Braiek, H.B. and Khomh, F. (2024). *Machine Learning Robustness: A Primer*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.2404.00897>.

Brown, O., Curtis, A. and Goodwin, J. (2021). *Principles for Evaluation of AI/ML Model Performance and Robustness*. [online] arXiv.org.

doi:<https://doi.org/10.48550/arXiv.2107.02868>.

ChatGPT (2024). *ChatGPT - Latest News and Chat About AI*. [online] chatgpt.com.

Available at: <https://chatgpt.com/>.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1810.04805>.

Fu, Y., Ou, W., Yu, Z. and Lin, Y. (2023). MIGA: A Unified Multi-Task Generation Framework for Conversational Text-to-SQL. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, 37(11), pp.12790–12798.

doi:<https://doi.org/10.1609/aaai.v37i11.26504>.

Fursov, I., Zaytsev, A., Pavel Burnyshev, Dmitrieva, E.F., Klyuchnikov, N., Kravchenko, A., Ekaterina Artemova and Evgeny Burnaev (2022). A Differentiable Language Model Adversarial Attack on Text Classifiers. *IEEE Access*, 10, pp.17966–17976.

doi:<https://doi.org/10.1109/access.2022.3148413>.

Guerrero, J., Liang, G. and Izzat Alsmadi (2023). Adversarial Text Perturbation Generation and Analysis. *ICSC*. doi:<https://doi.org/10.1109/icsc60084.2023.10349981>.

Huang, T., Zeng, Y., Dong, B., Xu, H., Xu, S., Lau, and Zuo, W. (2023). TextField3D: Towards Enhancing Open-Vocabulary 3D Generation with Noisy Text Fields. *arXiv (Cornell University)*. doi:<https://doi.org/10.48550/arxiv.2309.17175>.

HuggingFace (2020). *microsoft/deberta-large · Hugging Face*. [online] huggingface.co. Available at: <https://huggingface.co/microsoft/deberta-large> [Accessed 22 May 2024].

Jia, R. and Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. *arXiv:1707.07328 [cs]*. [online] Available at: <https://arxiv.org/abs/1707.07328>.

Judit Ács, Endre Hamerlik, Schwartz, R., Smith, N.A. and András Kornai (2023). Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, pp.1–40. doi:<https://doi.org/10.1017/s1351324923000190>.

- Khan, P., Dengel, A. and Ahmed, S. (2023). Randout-KD: Finetuning Foundation Models for Text Classification via Random Noise and Knowledge Distillation. *IEEE Open Access*, Volume 3, Pages 457 - 465. doi:<https://doi.org/10.5220/0011687800003393>.
- Ko, H., Lee, J., Kim, J., Lee, J. and Shim, H. (2020). Diversity regularized autoencoders for text generation. *Proceedings of the ACM Symposium on Applied Computing*, Pages 883 - 891(35th Annual ACM Symposium on Applied Computing, SAC 2020). doi:<https://doi.org/10.1145/3341105.3373998>.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*. [online] Available at: <https://arxiv.org/abs/1910.13461>.
- Li, A., Zhang, F., Li, S., Chen, T., Pan S and Wang, H. (2023). Efficiently generating sentence-level textual adversarial examples with Seq2seq Stacked Auto-Encoder. *Expert Systems with Applications*, 213, pp.119170–119170. doi:<https://doi.org/10.1016/j.eswa.2022.119170>.
- Lu, Y., Zhang, J., Zeng, J., Wu, S. and Zong, C. (2022). Attention Analysis and Calibration for Transformer in Natural Language Generation. *IEEE/ACM transactions on audio, speech, and language processing*, 30, pp.1927–1938. doi:<https://doi.org/10.1109/taslp.2022.3180678>.
- Meta (2024). *Introducing Meta Llama 3: The most capable openly available LLM to date*. [online] ai.meta.com. Available at: <https://ai.meta.com/blog/meta-llama-3/>.
- Mitchell, E., Yoon-Ho Alex Lee, Khazatsky, A., Manning, C.D. and Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *arXiv (Cornell University)*. doi:<https://doi.org/10.48550/arxiv.2301.11305>.
- Pavan, Subba Oota, Mounika Marreddy, Venkateswara Kagita and Gupta, M. (2023). On Robustness of Finetuned Transformer-based NLP Models. *ACL*. doi:<https://doi.org/10.18653/v1/2023.findings-emnlp.477>.

Pervaiz Iqbal Khan, Razzak, I., Dengel, A. and Ahmed, S. (2022). A Novel Approach to Train Diverse Types of Language Models for Health Mention Classification of Tweets. *Lecture notes in computer science*, Volume 13530 LNCS, Pages 136 - 147(31st International Conference on Artificial Neural Networks, ICANN 2022), pp.136–147. doi:https://doi.org/10.1007/978-3-031-15931-2_12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2018). *Language Models are Unsupervised Multitask Learners*. [online] Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rajendran, U.K. and Trabelsi, A. (2023). *Exploration of Contrastive Learning Strategies toward more Robust Stance Detection*. [online] ACLWeb. doi:<https://doi.org/10.18653/v1/2023.wassa-1.37>.

Ross, A., Wu, T., Peng, H., Peters, M.E. and Gardner, M. (2022). Tailor: Generating and Perturbing Text with Semantic Controls. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. doi:<https://doi.org/10.18653/v1/2022.acl-long.228>.

Sahar Sadrizadeh, Ljiljana Dolamic and Frossard, P. (2022). Block-Sparse Adversarial Attack to Fool Transformer-Based Text Classifiers. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:<https://doi.org/10.1109/icassp43922.2022.9747475>.

Sai, A.B., Dixit, T., Dev Yashpal Sheth, Mohan, S. and Khapra, M.M. (2021). Perturbation CheckLists for Evaluating NLG Evaluation Metrics. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. doi:<https://doi.org/10.18653/v1/2021.emnlp-main.575>.

Sasada, T., Kawai, M., Yuzo Taenaka, Fall, D. and Youki Kadobayashi (2021). Differentially-Private Text Generation via Text Preprocessing to Reduce Utility Loss. *3rd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2021*, Pages 42 - 47. doi:<https://doi.org/10.1109/icaaic51459.2021.9415242>.

See, A., Pappu, A., Saxena, R., Yerukola, A. and Manning, C.D. (2019). Do Massively Pretrained Language Models Make Better Storytellers? *arXiv:1909.10705 [cs]*. [online] Available at: <https://arxiv.org/abs/1909.10705>.

Sinha, S., Chen, H., Sekhon, A., Ji, Y. and Qi, Y. (2021). *Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.2108.04990>.

Speer, R., Chin, J. and Havasi, C. (2018). *ConceptNet 5.5: An Open Multilingual Graph of General Knowledge*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.1612.03975>.

Srinivasan, S., Edmon Begoli, Mahbub, M. and Knight, K. (2021). Nomen est Omen - The Role of Signatures in Ascribing Email Author Identity with Transformer Neural Networks. *Proceedings - 2021 IEEE Symposium on Security and Privacy Workshops, SPW 2021*, Pages 291 - 297(2021 IEEE Symposium on Security and Privacy Workshops, SPW 2021). doi:<https://doi.org/10.1109/spw53761.2021.00049>.

Stuart, R. and Craven, P. (2019). The Gentle Art of Dithering. *Journal of the Audio Engineering Society*, 67(5), pp.278–299. doi:<https://doi.org/10.17743/jaes.2019.0011>.

Talmor, A., Herzig, J., Lourie, N. and Berant, J. (2019). *CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.1811.00937>.

Tran, Q., Krystsina Shpileuskaya, Zaunseder, E., Salg, J., Putzar, L. and Blankenburg, S. (2023). Robustness Analysis uncovers Language Proficiency Bias in Emotion Recognition Systems. *ACII*. doi:<https://doi.org/10.1109/acii59096.2023.10388123>.

Vikas Raunak, Menezes, A. and Awadalla, H. (2023). Dissecting In-Context Learning of Translations in GPT-3. *EMNLP 2023*, Pages 866 - 872. doi:<https://doi.org/10.18653/v1/2023.findings-emnlp.61>.

Wang, S., Li, Z., Qian, H., Yang, C., Wang, Z., Shang, M., Kumar, V., Tan, S., Ray, B., Bhatia, P., Ramesh Nallapati, Ramanathan, M., Roth, D. and Xiang, B. (2023). ReCode: Robustness Evaluation of Code Generation Models. *Proceedings of the Annual Meeting of*

the Association for Computational Linguistics, Volume 1, Pages 13818 - 13843.

doi:<https://doi.org/10.18653/v1/2023.acl-long.773>.

Wang, Y., Hu, W. and Hong, R. (2024). Iterative Adversarial Attack on Image-guided Story Ending Generation. *IEEE transactions on multimedia*, pp.1–14.

doi:<https://doi.org/10.1109/tmm.2023.3345167>.

Wu, W., Li, W., Liu, J., Xiao, X., Cao, Z., Li, S. and Wu, H. (2022). FRSUM: Towards Faithful Abstractive Summarization via Enhancing Factual Robustness. *Findings of the Association for Computational Linguistics: EMNLP 2022*, Pages 3640 - 3654.

doi:<https://doi.org/10.18653/v1/2022.findings-emnlp.267>.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Ruslan Salakhutdinov and Chaudhuri, K. (2020). A Closer Look at Accuracy vs. Robustness. *arXiv (Cornell University)*.

Ye-Bin, M., Kim, J., Kim, H., Son, K. and Oh, T.-H. (2023). TextManiA: Enriching Visual Feature by Text-driven Manifold Augmentation. *Proceedings of the IEEE International Conference on Computer Vision*, Pages 2526 - 2537.

doi:<https://doi.org/10.1109/iccv51070.2023.00239>.

Yin, W., Hay, J. and Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *arXiv:1909.00161 [cs]*. [online] Available at: <https://arxiv.org/abs/1909.00161>.

Zames, G. and Francis, B. (1983). Feedback, minimax sensitivity, and optimal robustness. *IEEE Transactions on Automatic Control*, 28(5), pp.585–601.

doi:<https://doi.org/10.1109/tac.1983.1103275>.

Zhu, H., Song, Y. and Liu, B. (2023). KTGAT: Improving the Robustness of Knowledge-enhanced Text Generation via Adversarial Training. *ICCEA*.

doi:<https://doi.org/10.1109/iccea58433.2023.10135313>.