ParaFusion: A Large-Scale LLM-Driven English Paraphrase Dataset Infused with High-Quality Lexical and Syntactic Diversity

Lasal Jayawardena and Prasan Yapa

School of Computing, Informatics Institute of Technology,
Colombo 00600, Sri Lanka

Abstract. Paraphrase generation is a pivotal task in natural language processing (NLP). Existing datasets in the domain lack syntactic and lexical diversity, resulting in paraphrases that closely resemble the source sentences. Moreover, these datasets often contain hate speech and noise, and may unintentionally include non-English language sentences. This research introduces ParaFusion, a large-scale, high-quality English paraphrase dataset developed using Large Language Models (LLM) to address these challenges. ParaFusion augments existing datasets with high-quality data, significantly enhancing both lexical and syntactic diversity while maintaining close semantic similarity. It also mitigates the presence of hate speech and reduces noise, ensuring a cleaner and more focused English dataset. Results show that ParaFusion offers at least a 25% improvement in both syntactic and lexical diversity, measured across several metrics for each data source. The paper also aims to set a gold standard for paraphrase evaluation as it contains one of the most comprehensive evaluation strategies to date. The results underscore the potential of ParaFusion as a valuable resource for improving NLP applications.

Keywords: Paraphrase Generation, Natural Language Generation, Deep Learning, Large Language Models, Data Centric AI.

1 Introduction

Paraphrase Generation, also known as Question Paraphrase Generation, is a fundamental task and a significant area of focus in NLP. This field has been the subject of research for several decades. Paraphrase Generation plays a critical role in data augmentation, a process that is vital for enhancing the performance of numerous NLP tasks. By generating diverse expressions of identical information, it significantly enriches the training data, thereby improving the robustness and generalization capabilities of NLP models [1][2][3][4].

In recent years, neural-based approaches, such as sequence-to-sequence models, have been increasingly used for paraphrase generation due to their ability to learn complex patterns and generate fluent text. However, they require large amounts of high-quality annotated data for training, which can be difficult and costly to obtain. Data quality determines the capability of the model to generate diverse

David C. Wyld et al. (Eds): AIBD, MLSC, ACSTY, NATP, CCCIoT, SVC, SOFE, ITCSS -2024
pp. 199-208, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140418

paraphrases. Existing models often struggle with maintaining the semantic equivalence between the original text and the generated paraphrase, especially for longer and more complex sentences. These limitations stem from the issue that the current datasets that are available do not have high-quality paraphrases. A paraphrase to be considered a high-quality paraphrase needs to be lexically diverse, syntactically diverse, grammatically correct, and semantically similar. Section 4 outlines an analysis of existing data sources that highlight this issue. Apart from dataset quality, there is no proper evaluation strategy employed by researchers that assesses the quality of diverse paraphrases. Most existing work primarily use lexical metrics to determine model quality whereas the other three components of high-quality paraphrases are often ignored [5].

In light of these challenges, ParaFusion is introduced as a more precise English Dataset to address these issues. LLMs have gained a lot of attraction in recent years, significantly outperforming several state-of-the-art models (SOTA) in several domains [6]. In this paper, we employ existing datasets and augment the data to generate high-quality paraphrases using the ChatGPT (gpt-3.5-turbo) LLM to create ParaFusion. This paper provides a comprehensive analysis of ParaFusion, investigating it using a range of evaluation metrics to explore various facets of the dataset's quality, as shown in Section 4, to demonstrate that the paraphrases generated by ParaFusion are more diverse in terms of syntax and lexical compared to existing datasets, while simultaneously maintaining strong semantic similarity between paraphrased sentences. We utilize a rigorous framework for dataset evaluation in hopes of setting a gold standard for future paraphrase evaluation research. Such a comprehensive strategy is needed to precisely evaluate paraphrases which is drastically different from other Sequence-to-Sequence NLP Taks. Moreover, our human evaluation results Section 4.3 corroborate that ParaFusion indeed offers higher-quality paraphrases compared to previous datasets. The results suggest a notable potential for realizing enhancements in paraphrase generation tasks, underlining ParaFusion's ability to shepherd future advancements in NLP.

2 Related Work

The landscape of paraphrase generation datasets is critical to understanding the research context and challenges in this domain. This section presents a review of noteworthy paraphrase datasets, highlighting their strengths and limitations.

Paraphrase Database (PPDB) PPDB is a comprehensive resource that houses over 220 million paraphrase pairs [7]. The PPDB is compiled through a technique known as bilingual pivoting. The rationale behind this approach is that if two English phrases are translated into the same foreign language phrase, they can be inferred to have identical meanings. Each pair within the PPDB is accompanied by a range of scores, such as paraphrase probabilities and monolingual distributional similarity scores. However, despite its extensive content and detailed scoring

system, the PPDB's utility has been questioned recently due to its exclusive focus on phrasal and lexical paraphrases, neglecting sentence paraphrases.

Twitter URL The Twitter URL dataset [8] is a comprehensive collection of large-scale sentential paraphrases sourced from Twitter and connected through shared URLs. This dataset is bifurcated into two subsets, each encompassing both paraphrases and non-paraphrases. The labeling of one subset is performed by human annotators, while the other subset is labeled automatically. It should be noted that the annotation does contain some noise due to the automatic labeling of sentence pairs. Due to the noisiness of the labels, this dataset is not widely used.

Wiki Answer The Wiki Answer dataset [9] encompasses an estimated 18 million pairs of questions that are paraphrased. The dataset was constructed by mapping open-domain questions to queries over a database of web extractions. The dataset also includes word alignments that connect synonyms within the paraphrased sentences. The dataset is limited in scope as all the sentences provided are in the form of questions, thereby confining the paraphrases to question format only. The dataset is also noisy such that paraphrases do not have the needed semantic similarity of a high-quality dataset.

MSCOCO The MSCOCO dataset [10] was primarily characterized as a comprehensive object detection dataset. It comprises over 120,000 images, each of which is accompanied by five distinct captions, contributed by five separate annotators. Typically, the annotators focus on detailing the most conspicuous object or action within an image, rendering this dataset particularly useful for tasks related to paraphrasing.

Microsoft Research Paraphrase Corpus The Microsoft Research Paraphrase Corpus (MRPC) Dataset [11] comprises 5800 sentence pairs derived from online news sources. It also includes human annotations that denote whether each pair represents a paraphrase or semantic equivalence relationship. This was one of the oldest datasets which is still being used for model evaluation but its only downside is that there are very few sentences in the corpora. The sentence length is quite longer compared to other phrase heavy datasets making it a valuable addition.

Quora The Quora Dataset or Quora Question Pair Dataset [12] which is predominantly used for training and evaluation, contains 150,000 question pairs that are annotated as paraphrases. These validated paraphrase question pairs were specifically employed for the training and testing phases of the paraphrase generation task. The dataset is similar to WikiAnswer in its limitation of having only questions.

ParaNMT The ParaNMT dataset [13] comprises over 50 million pairs of English sentential paraphrases. These pairs were autonomously generated through the application of back-translation to translate the non-English component of a substantial Czech-English parallel corpus. A Czech-English Neural Machine Translation (NMT) system was employed to translate Czech sentences from the training

data into English. These translations were then paired with the English references to form English-English paraphrase pairs. This is the first paraphrase dataset that utilized back-translation. Upon analysis, one downside is the inclusion of improperly formed paraphrases and non-English sentences.

ParaBank The ParaBank datasets [14, 15] were developed using a Czech-English Neural Machine Translation (NMT) system to generate new paraphrases of English reference sentences. The first version, ParaBank1, introduced lexical constraints to the NMT decoding process, allowing for the generation of multiple high-quality sentential paraphrases for each source sentence. This resulted in an English paraphrase resource that exhibits a higher degree of lexical diversity. Its successor, ParaBank2, addressed the issue of syntactic diversity by providing multiple diverse sentential paraphrases. These paraphrases were generated from a bilingual corpus using negative constraints, inference sampling, and clustering. Even with the improvements both datasets still lack syntactic diversity.

PAWS (Paraphrase Adversaries from Word Scrambling) The PAWS Dataset [16] contains sentences with high bag-of-words (BOW) overlap but having different word order. The PAWS dataset creation involved a two-step process. Initially, a language model was used to generate sentence pairs with high lexical overlap through word swapping, ensuring naturalness and well-formedness. Subsequently, back translation was employed to create paraphrases with high bag-of-words overlap but distinct word order. The PAWS dataset is further divided into two subsets: PAWSQQP and PAWSWiki. The PAWSQQP subset is derived from the Quora Question Pairs (QQP) corpus, while the PAWSWiki subset is derived from Wikipedia. Subsets of the dataset were then subjected to human review for sentence correction and paraphrase identification. PAWSWiki has significantly better quality paraphrases than all the other datasets, yet improvements can be made for syntactic diversity. There is a considerable portion of PAWS that is noisy and the other portion that is labeled exhibit high-quality.

3 ParaFusion

3.1 Data Sources

ParaFusion is a comprehensive tool constructed on the foundation of several datasets, each contributing unique elements to the overall structure. The first dataset we utilized was the MRPC Dataset [11], which provided a solid base for our work. Following this, we incorporated a subset of the Quora Dataset [12], specifically selecting sentences that were labeled as paraphrases to enrich our data pool. To further diversify our data, we included PAWSWiki, a component of the PAWS Dataset [16]. However, we consciously decided against using PAWSQQP as it contained source sentences identical to those in the Quora Dataset, which would have introduced

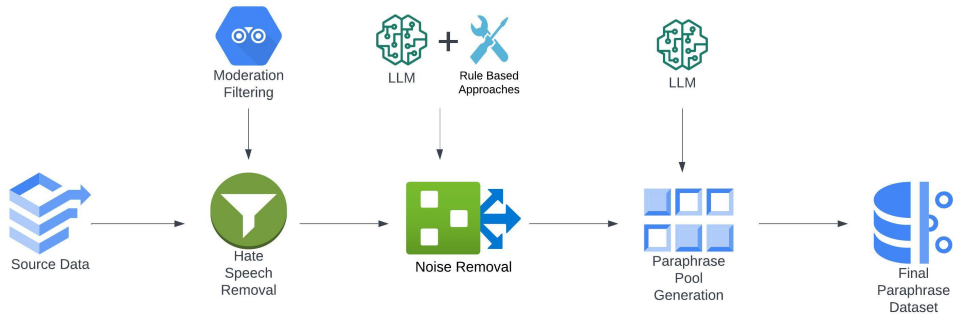

Fig. 1: High-level diagram outlining the dataset creation process.

unnecessary redundancy into our data. We also considered the use of three additional datasets: ParaNMT [13], Parabank1 [14], and Parabank2 [15]. However, due to financial constraints, we were unable to utilize these datasets in their entirety. Instead, we strategically selected all the common source sentences from these datasets and supplemented this with an additional 250,000 source sentences.

The decision to incorporate multiple datasets into our research methodology was a strategic one, aimed at enhancing the comprehensiveness and diversity of the dataset. Different datasets inherently possess varying sentence lengths and unique content, style, and context. By amalgamating several datasets, we ensured that ParaFusion captured a broad spectrum of sentence lengths, and provided an extensive range of topics, writing styles, and contexts for training. This approach also helped us to mitigate the risk of potential data bias, ensuring a more balanced and representative dataset.

It should be emphasized that the method we suggest is not confined to these particular datasets. Instead, it has the potential to be utilized with any general text to generate paraphrases.

3.2 Base Dataset Creation

In the construction of ParaFusion, we initially amalgamated the aforementioned data sources, selecting approximately 750,000 source sentences for paraphrase generation. An initial pass was conducted to filter out offensive content, utilizing OpenAI's Moderation Endpoint [17]. This process flagged any source sentence that fell under the categories of "sexual", "hate", "harassment", "self-harm", "sexual/minors", "hate/threatening", "violence/graphic", "self-harm/instructions", "self-harm/intent", "harassment/threatening", or "violence". This enabled us to filter out approximately 50,000 source sentences containing offensive content.

Subsequently, we employed the ChatGPT (gpt-3.5-turbo) LLM to augment the source sentences. The prompt used for each dataset varied slightly, and the reasons

for these variations are discussed in the Additional Processing Section 3.3. See Figure 2 for an example illustration of this.

```
Given a Source Sentence: "'$Source Sentence"', generate 5 diverse paraphrases. Try to generate
paraphrases that are both lexical and syntactically diverse from the Source Sentence. Give the
output as a numbered list.
```

Fig. 2: This figure illustrates a sample prompt fed to the gpt-3.5-turbo model for generating diverse paraphrases.

We adopted an iterative prompt engineering approach to construct an effective prompt. This prompt was then input into the gpt-3.5-turbo model, along with the source sentences. For paraphrase generation, we set the temperature parameter to 0. The model output was a string containing augmented, diverse sentence paraphrases in a numbered list format. This string was subsequently processed to generate a list of strings, which was then used to construct the dataset. By utilizing the source sentences, we successfully generated nearly 3.5 million sentence paraphrases.

3.3 Additional Processing

Our study revealed that datasets constructed using back-translation [13][14][15] often contained a significant amount of noise, including non-English source sentences. To tackle this, we iteratively developed a new prompt for the gpt-3.5-turbo model, instructing it to identify English sentences and generate paraphrases, or output "Error" for non-English sentences. This method, coupled with a rule-based approach to filter out certain responses from the model, proved efficient in significantly reducing noise, and eliminating approximately 10,000 non-English source sentences during generation.

In the final stage of dataset creation, we didn't merely use the source sentence and generated paraphrase pairs as they were. Instead, we treated the source sentence and generated paraphrase as a pool of sentences, from which we created unique paraphrase pairs. This approach diversified the paraphrase pairs in our dataset, ensuring a wider range of sentence structures and expressions.

This method offers several benefits. Firstly, it enhances the diversity of our dataset, which is crucial for training robust and generalizable models. Secondly, it helps to mitigate the risk of overfitting by exposing our models to a more representative sample of the data they will encounter in the real world. Lastly, it helps to reduce the impact of any noise in the source sentence. By creating unique paraphrase pairs, we can ensure that any noise in the source sentences or generated paraphrases are not consistently paired with the same sentences, thereby reducing the likelihood that our models will learn to associate this noise with specific inputs or outputs.

The final dataset comprises around 2 million unique paraphrase sentence pairs. These techniques significantly reduced noise in ParaFusion, contributing to the construction of a higher-quality dataset, comparable to human annotation or the next best alternative.

4 Evaluation

In this section, we evaluate ParaFusion, which was created using several source datasets: the Microsoft Research Paraphrase Corpus (MRPC), a subset of the Quora Dataset, PAWSWiki, ParaNMT, Parabank1, and Parabank2. For a fair evaluation, we only consider source sentences common to both ParaFusion and these source datasets.

We use "Para-Common Subset" to denote paraphrases in ParaFusion common to ParaNMT, Parabank1, and Parabank2. "MRPC Subset", "PAWS Subset", and "QQP Subset" refer to paraphrases in ParaFusion common to MRPC, Quora Dataset, and PAWSWiki Dataset, respectively. Each subset is analyzed separately for fairness.

4.1 Quantitative Analysis

We adopt a comprehensive quantitative evaluation methodology to assess the data sources and sentence pairs in ParaFusion. To guarantee a fair evaluation, the dataset is partitioned into four segments based on the source sentences. Our evaluation focuses on three crucial characteristics: semantic similarity, syntactic diversity, and lexical diversity.

Semantic Similarity Semantic similarity is a measure of the degree to which two pieces of text are related in terms of their meaning. In our research, we quantify this similarity by leveraging various models to obtain sentence embeddings of the source and the paraphrase, and then calculating the cosine similarity between these embeddings.

For instance, we use the "Ada Score" which is derived from OpenAI's text-embedding-ada-002 model [18]. Similarly, the "SimCSE Score" is calculated using SimCSE's sup-simcse-roberta-large model [19], and the "PromCSE Score" is based on PromCSE's sup-promcse-roberta-large model [20].

We also utilize several models from the sentence-transformers library [21]. The "Mpnet Score" is calculated using the all-mpnet-base-v2 model, while the "Mpnet-qa Similarity Score" is derived from the multi-qa-mpnet-base-dot-v1 model. The "Roberta Score" is based on the all-distilroberta-v1 model, and the "Mini Score" and "Mini Score2" are calculated using the all-MiniLM-L12-v2 and all-MiniLM-L6-v2 models, respectively.

The comprehensive evaluation of semantic similarity is presented in Table 1. It is evident that ParaFusion not only maintains semantic similarity but also, in some cases, surpasses the quality of the original data source. The perceived low similarity in less sophisticated models can be attributed to the lack of lexical or syntactic diversity in the original source sentences, which results in highly similar sentences. However, in ParaFusion, the complexity of the sentences allows for a more accurate measurement of semantic similarity using advanced models such as the text-embedding-ada-002 or SIMCSE.

Data Source	**Ada Score (↑)**	**SimCSE Score (↑)**	**PromCSE Score (↑)**	**Mpnet Score (↑)**	**Mpnet-qa Score (↑)**	**Roberta Score (↑)**	**Mini Score (↑)**	**Mini Score2 (↑)**
MSR Original	95.53%	86.50%	99.22%	84.27%	86.23%	82.80%	83.27%	82.47%
MSR Subset (Ours)	**96.59%**	**93.30%**	**99.56%**	**88.33%**	**90.80%**	**86.45%**	**86.82%**	**86.01%**
QQP Original	**95.52%**	89.61%	**99.65%**	**88.74%**	**91.30%**	**87.00%**	**88.75%**	**88.37%**
QQP Subset (Ours)	94.56%	**89.87%**	**99.65%**	86.86%	89.33%	84.29%	85.12%	83.85%
PAWS Original	**98.90%**	**97.42%**	**99.90%**	**96.96%**	**97.06%**	**97.20%**	**97.24%**	**97.13%**
PAWS Subset (Ours)	96.22%	92.20%	99.34%	91.74%	91.57%	91.72%	92.08%	91.33%
ParaNMT Original	94.83%	87.50%	99.88%	87.00%	90.67%	82.05%	87.69%	87.58%
ParaBank1 Original	**95.58%**	**89.85%**	**99.91%**	**89.00%**	**92.09%**	**85.31%**	**89.60%**	**89.51%**
ParaBank2 Original	94.49%	85.52%	98.87%	82.89%	87.60%	79.76%	83.63%	83.26%
Para-Common Subset (Ours)	95.04%	74.32%	98.60%	67.16%	74.11%	63.00%	67.19%	65.80%

Table 1: Comparison of Semantic Similarity of the original data sources and corresponding ParaFusion data subsets.

Syntactic Diversity Syntactic diversity refers to the variety and complexity of sentence structures of a paraphrase given a source sentence. High syntactic diversity indicates that the paraphrase sentences are diverse and linguistically rich. We assess this diversity using several metrics.

The "Ted-F Score" and "Ted-3 Score" are calculated by building constituency parse trees for the source and paraphrase sentences using Stanza [22], converting the trees to bracket notation using the NLTK library [23] and regex, and then using the APTED library [24] to calculate the Full Tree Edit Distance and the Tree Edit Distance of the first three layers, respectively.

The "Kermit Score" is calculated by obtaining the cosine similarity of the source and the paraphrase syntactic embeddings using the Kermit library [25], and then subtracting this similarity from one.

The "ST Kernel Score" and "NP Kernel Score" are calculated by first building constituency parse trees for the source and paraphrase sentences using Stanza, converting the trees to an NLTK Tree, and then finding all the subtrees or node

pairs, respectively. Then the kernel similarity is calculated using the number of unique common subtrees or node pairs over the number of unique total subtrees or node pairs, and then subtracting this similarity from one. "ST Kernel Score" is the score calculated related to the subtrees in the constituency parse trees whereas "NP Kernel Score" is the score calculated using the node pair similarity.

The full syntactic diversity evaluation is shown in Table 2. We can see that ParaFusion has a significant improvement over all the original data sources. This is because it was able to generate more syntactically rich paraphrases.

Data Source	**Ted-F Score (↑)**	**Ted-3 Score (↑)**	**Kermit Score (↑)**	**ST Kernel Score (↑)**	**NP Kernel Score (↑)**
MSR Original	18.65	3.76	55.23%	65.18%	82.87%
MSR Subset (Ours)	**29.09**	**4.92**	**64.37%**	**73.55%**	**89.34%**
QQP Original	9.30	2.02	57.59%	70.14%	88.29%
QQP Subset (Ours)	**16.91**	**3.35**	**71.62%**	**81.13%**	**93.64%**
PAWS Original	9.46	1.70	37.74%	47.44%	73.47%
PAWS Subset (Ours)	**27.02**	**4.67**	**61.06%**	**69.48%**	**87.35%**
ParaNMT Original	3.12	2.01	54.06%	73.99%	82.51%
ParaBank1 Original	2.06	1.48	45.85%	63.22%	70.37%
ParaBank2 Original	3.76	2.07	61.94%	85.01%	96.78%
Para-Common Subset (Ours)	**9.33**	**3.99**	**80.78%**	**91.73%**	**97.94%**

Table 2: Comparison of Syntactic Diversity of the original data sources and corresponding ParaFusion data subsets.

Lexical Diversity Lexical diversity refers to the range and variety of words used in a text. It is a measure of the breadth of vocabulary and the use of synonyms. In the context of paraphrasing, assessing lexical diversity is crucial to understand the extent of vocabulary variation. We used several metrics to assess lexical diversity.

The "BOW Overlap" is calculated by determining the intersection of tokens between the source and the paraphrase, divided by the total number of tokens. This value is then subtracted from one.

The "Corpus BLEU" and "Corpus BLEU2" scores are calculated using the SacreBLEU Library [26]. The Corpus BLEU2 score uses the "method1" smoothing function in the SacreBLEU library. Both scores are then subtracted from one.

The "Sentence BLEU" score is calculated in a similar manner to the Corpus BLEU score using the SacreBLEU Library but at the sentence level. This score is also subtracted from one.

The "METEOR" score is calculated using the NLTK library and then subtracted from one.

The "ROUGE 1", "ROUGE 2", and "ROUGE L" scores are calculated using the Google Research library [27] and then subtracted from one.

The "Token ∩/∪" score is similar to the BOW Overlap score, but with a small difference. It is calculated using the intersection of tokens between the source and the paraphrase, divided by the total number of unique tokens. This value is then subtracted from one.

The "Google BLEU" score is calculated using Huggingface's Evaluate library and then subtracted from one.[1]

The "TER (Translation Error Rate)", "WER (Word Error Rate)", and "CharacTER (Character Error Rate)" scores are calculated using Huggingface's Evaluate library.[1]

The full lexical diversity evaluation is shown in Table 3 and Table 4. We can see that ParaFusion has a significant improvement over all the original data sources.

Data Source	**1 - BOW Overlap (↑)**	**1 - Corpus BLEU (↑)**	**1 - Corpus BLEU2 (↑)**	**1 - Sentence BLEU (↑)**	**1 - METEOR (↑)**	**1 - ROUGE 1 (↑)**
MSR Original	35.37%	93.92%	99.62%	59.75%	31.43%	29.49%
MSR Subset (Ours)	**43.72%**	**96.53%**	**99.65%**	**75.93%**	**38.96%**	**36.81%**
QQP Original	36.69%	**95.44%**	99.15%	70.80%	35.33%	33.46%
QQP Subset (Ours)	**53.04%**	82.21%	**99.46%**	**84.33%**	**50.20%**	**51.86%**
PAWS Original	19.33%	94.21%	99.60%	34.96%	8.58%	5.96%
PAWS Subset (Ours)	**40.26%**	**98.15%**	**99.65%**	**72.39%**	**34.59%**	**30.22%**
ParaNMT Original	53.56%	62.19%	97.75%	58.26%	27.33%	18.37%
ParaBank1 Original	46.50%	62.47%	97.73%	51.62%	24.54%	16.08%
ParaBank2 Original	55.36%	63.06%	97.86%	66.85%	39.24%	30.43%
Para-Common Subset (Ours)	**74.31%**	**82.24%**	**98.72%**	**86.45%**	**69.03%**	**64.82%**

Table 3: Comparison of Lexical Diversity of the original data sources and corresponding ParaFusion Data Subsets.

4.2 Qualitative Evaluation

During the qualitative analysis, we uncovered intriguing information. A prevalent issue in existing datasets is that most paraphrases are merely sentences with substituted synonyms. Figure 3 illustrates an example of this, where "A" is replaced by "One", leaving the rest of the sentence identical to the source. This does not constitute an effective paraphrase.

Another problem we observed is that existing data sources, particularly those relying on back-translation, often contain paraphrases with altered meanings due to word choice. Figure 4 exemplifies a situation where the meaning of a paraphrase has deviated from the original source datasets. This occurs when inappropriate words

[1] https://github.com/huggingface/evaluate

Data Source	1 - ROUGE 2 (↑)	1 - ROUGE L (↑)	1 - Token ∩/∪ (↑)	TER (↑)	WER (↑)	CER (↑)	1 - Google BLEU (↑)
MSR Original	47.55%	33.97%	44.95%	70.34	76.07	42.98%	55.96%
MSR Subset (Ours)	**63.22%**	**51.86%**	**54.79%**	**85.35**	**98.47**	**64.95%**	**69.81%**
QQP Original	57.72%	36.85%	49.04%	56.56	62.77	44.74%	64.96%
QQP Subset (Ours)	**76.95%**	**57.89%**	**66.33%**	**69.92**	**75.34**	**76.94%**	**79.12%**
PAWS Original	21.44%	12.78%	15.74%	14.81	23.63	14.19%	31.80%
PAWS Subset (Ours)	**56.97%**	**47.10%**	**66.33%**	**56.10**	**85.66**	**62.38%**	**65.92%**
ParaNMT Original	33.20%	19.25%	61.39%	30.10	66.10	25.74%	44.99%
ParaBank1 Original	31.69%	16.90%	52.51%	23.71	50.47	21.61%	37.37%
ParaBank2 Original	55.24%	32.97%	67.69%	46.84	68.01	35.39%	66.87%
Para-Common Subset (Ours)	**89.12%**	**69.68%**	**82.23%**	**82.46**	**89.31**	**85.10%**	**81.30%**

Table 4: Comparison of Lexical Diversity of the original data sources and corresponding ParaFusion Data Subsets.

are used without considering the context. For instance, replacing "culverts" with "driers" is not suitable in this context.

In contrast, paraphrases in ParaFusion demonstrate superior lexical diversity and more syntactic changes, while preserving the original meaning.

Source Sentence: A poetic example of early modern philosophical thought can be found in the surprising works of the renowned intellectual Stoyan Mihaylovski.
Original Paraphrase: One poetic example of early modern philosophical thought can be found in the surprising works of the renowned intellectual Stoyan Mihaylovski.
ParaFusion Paraphrase: Stoyan Mihaylovski's works are a remarkable representation of early modern philosophical thought, expressed in a poetic manner.

Fig. 3: This figure illustrates an instance where the paraphrase in a source dataset has only word substitutions.

Furthermore, we observed that the original data sources contained a substantial number of paraphrase pairs where the source and the paraphrase were identical. This presents a significant problem because training a model on such datasets could encourage the model to simply reproduce the input, thereby defeating the purpose of paraphrasing. This qualitative analysis further underscores the effectiveness and reliability of ParaFusion in these scenarios.

Source Sentence: The water is moved by the gravity and is controlled by huge valves in the driers .
Original Paraphrase: The water is moved by gravity and is controlled by huge valves in the culverts .
ParaFusion Paraphrase: Huge valves in the driers regulate the movement of water, which is facilitated by the force of gravity.

Fig. 4: This figure illustrates an instance where the paraphrase in a previous dataset has a different meaning.

4.3 Human Evaluation

In our research, we conducted human evaluations using four annotators who assessed approximately 7000 paraphrase pairs across various datasets, including the source and ParaFusion. The source sentences were selected as follows: 200 from the MRPC Dataset, 250 from the Quora Dataset, 250 from the PAWSWiki, and 300 common source sentences from ParaNMT, Parabank1, and Parabank2. The sentences were sampled to ensure that the lengths of the data source sentences were properly represented. The corresponding paraphrases from the source datasets and ParaFusion were then selected, resulting in a total of 7000 paraphrase pairs for evaluation.

For the evaluation, we used a 5-point Likert scale [28] to assess key metrics, including Semantic Similarity, Lexical Diversity, Syntactic Diversity, and Grammatical Correctness. The full breakdown of the Likert scale can be seen in Figure 6. In this scale, 5 represents the highest level of similarity, diversity, or correctness, while 1 indicates the lowest.

Specifically, semantic similarity ratings ranged from 5 for identical meaning to the source text, to 1 for completely different or unrelated meaning. Lexical diversity was evaluated based on vocabulary range and richness, with 5 indicating excellent diversity and 1 signifying limited diversity. Syntactic diversity was assessed by structural variations, with 5 denoting high diversity and 1 signifying minimal variation. Lastly, grammatical correctness was evaluated, with 5 indicating flawless grammar and 1 representing significant errors impacting comprehension.

Table 5 provides a summary of the human evaluation, and a full breakdown can be seen in Table 8. The results clearly indicate that ParaFusion is more lexically and syntactically diverse than the original data sources.

	Original	**ParaFusion**
Semantic Similarity	4.36	4.46
Lexical Diversity	2.29	3.09
Syntactic Diversity	2.44	3.40
Grammatical Correctness	4.37	4.79

Table 5: Comparison of Semantic Similarity, Lexical Diversity, Syntactic Diversity, and Grammatical Correctness between two data sets in Human Evaluation.

4.4 LLM Evaluation

In the past year, the use of LLMs for evaluation in NLP has gained traction. The primary reason for this trend is the ability of LLMs to outperform existing reference-free metrics [29]. In light of this, we conducted an LLM evaluation using OpenAI's gpt-4 model on the same data provided to our human annotators. The gpt-4 model was selected due to its status as the SOTA LLM at the time of writing this paper [6]. We designed the prompt using the same instructions given to the human annotators, as illustrated in Figure 5.

The summary of the results is presented in Table 6, while a full breakdown is provided in Table 7. The observations clearly indicate that ParaFusion is more lexically and syntactically diverse than the original data sources, mirroring the results of the Human Evaluation.

	Original	**ParaFusion**
Semantic Similarity	4.49	4.94
Lexical Diversity	1.75	3.34
Syntactic Diversity	2.02	3.84
Grammatical Correctness	4.75	4.99

Table 6: Comparison of Semantic Similarity, Lexical Diversity, Syntactic Diversity, and Grammatical Correctness between two data sets in LLM Evaluation.

5 Conclusion

This research paper introduces ParaFusion, a large-scale, high-quality English paraphrase dataset developed using LLMs. The dataset is designed to address the limitations of the lack of syntactic and lexical diversity in existing datasets. Additionally, it shows potential as a very high-quality alternative to human-annotated paraphrase pairs which are costly to obtain. ParaFusion augmented existing datasets to generate high-quality data, significantly enhancing both lexical and syntactic diversity while maintaining semantic similarity which was seen in the evaluation

section. It also mitigates the presence of hate speech and reduces noise, ensuring a cleaner, more focused English dataset. The evaluation of ParaFusion demonstrated its potential as a valuable resource for improving NLP applications.

Limitations

While ParaFusion addresses several challenges in paraphrase generation, there are a few limitations to be considered. Firstly, the dataset is focused on English paraphrases, which may limit its applicability to other languages. Future research could explore the development of similar datasets for other languages to enhance the diversity and inclusivity of NLP applications. This could be accomplished using multi-lingual LLMs or language-specific LLMs.

Secondly, although efforts were made to ensure the quality and reliability of the dataset, there may still be instances of noise, inaccuracies, or in rare cases, offensive language. The use of LLMs for paraphrase generation introduces the possibility of generating incorrect paraphrases. Researchers and practitioners should exercise caution and conduct thorough evaluations when using the ParaFusion dataset.

Another point to be considered is the issue of error propagation. Our dataset was generated using gpt-3.5-turbo, thus inheriting all potential risks associated with it. This can also include phenomena like quality drift where the model output can change as the model adapts, if expansions were to be done. This could potentially introduce additional inaccuracies into the paraphrase generation process, and users should be aware of this when using the ParaFusion dataset and expanding it.

Lastly, the evaluation metrics used in this research provide valuable insights into the quality and diversity of the dataset. However, they may not capture all aspects of paraphrase generation. Future research could explore additional evaluation metrics or approaches to further assess the effectiveness and performance of paraphrase generation models using our ParaFusion dataset.

Ethics Statement

In conducting this research and developing the ParaFusion dataset, we took several ethical considerations into account. Firstly, we ensured that the dataset contained minimal hate speech or offensive language by implementing a moderation filtering process. This was done to ensure that the dataset is safe and suitable for using in NLP applications.

Secondly, we made efforts to reduce noise in the dataset, such as removing non-English sentences and filtering out responses that did not meet the criteria for high-quality paraphrases. This was done to ensure that the dataset is of high quality and reliable for training NLP models.

Additionally, we considered the potential impact of our research on the broader NLP community. By providing a large-scale, high-quality paraphrase dataset, we

aim to contribute to the advancement of NLP applications. This dataset can be used to improve the performance and robustness of NLP models, leading to more accurate and reliable NLP.

References

1. K. R. McKeown, "Paraphrasing using given and new information in a question-answer system," in *Proceedings of the 17th annual meeting on Association for Computational Linguistics* -. La Jolla, California: Association for Computational Linguistics, 1979, p. 67. [Online]. Available: http://portal.acm.org/citation.cfm?doid=982163.982182
2. M. Meteer and V. Shaked, "Strategies for effective paraphrasing," in *Proceedings of the 12th conference on Computational linguistics* -, vol. 2. Budapest, Hungry: Association for Computational Linguistics, 1988, pp. 431–436. [Online]. Available: http://portal.acm.org/citation.cfm?doid=991719.991724
3. R. Kozlowski, K. F. McCoy, and K. Vijay-Shanker, "Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources," in *Proceedings of the second international workshop on Paraphrasing* -, vol. 16. Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1118984.1118985
4. I. A. Bolshakov and A. Gelbukh, "Synonymous Paraphrasing Using WordNet and Internet," in *Natural Language Processing and Information Systems*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, F. Meziane, and E. Métais, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3136, pp. 312–323, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-540-27779-8_27
5. J. Zhou and S. Bhat, "Paraphrase Generation: A Survey of the State of the Art," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5075–5086. [Online]. Available: https://aclanthology.org/2021.emnlp-main.414
6. OpenAI, "GPT-4 Technical Report," 2023, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/abs/2303.08774
7. J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB: The Paraphrase Database," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 758–764. [Online]. Available: https://aclanthology.org/N13-1092
8. W. Lan, S. Qiu, H. He, and W. Xu, "A Continuously Growing Dataset of Sentential Paraphrases," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1224–1234. [Online]. Available: http://aclweb.org/anthology/D17-1126
9. A. Fader, L. Zettlemoyer, and O. Etzioni, "Paraphrase-Driven Learning for Open Question Answering," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1608–1618. [Online]. Available: https://aclanthology.org/P13-1158
10. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, vol. 8693, pp. 740–755, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-10602-1_48

11. W. B. Dolan and C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. [Online]. Available: https://aclanthology.org/I05-5002
12. S. Iyer, N. Dandeka, and K. Csernai, "First Quora Dataset Release: Question Pairs," 2017. [Online]. Available: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs
13. J. Wieting and K. Gimpel, "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 451–462. [Online]. Available: http://aclweb.org/anthology/P18-1042
14. J. E. Hu, R. Rudinger, M. Post, and B. Van Durme, "ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation," Jan. 2019, arXiv:1901.03644 [cs]. [Online]. Available: http://arxiv.org/abs/1901.03644
15. J. E. Hu, A. Singh, N. Holzenberger, M. Post, and B. Van Durme, "Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 44–54. [Online]. Available: https://www.aclweb.org/anthology/K19-1005
16. Y. Zhang, J. Baldridge, and L. He, "PAWS: Paraphrase Adversaries from Word Scrambling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1298–1308. [Online]. Available: https://aclanthology.org/N19-1131
17. OpenAI, "Moderation," 2023. [Online]. Available: https://platform.openai.com/docs/guides/moderation
18. OpenAI, "New and Improved Embedding Model," 2023, publisher: OpenAI. [Online]. Available: https://openai.com/blog/new-and-improved-embedding-model
19. T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.552
20. Y. Jiang, L. Zhang, and W. Wang, "Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3021–3035. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.220
21. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. [Online]. Available: http://arxiv.org/abs/1908.10084
22. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.
23. S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.
24. M. Pawlik and N. Augsten, "Efficient Computation of the Tree Edit Distance," *ACM Transactions on Database Systems*, vol. 40, no. 1, pp. 1–40, Mar. 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2699485
25. F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, and F. Fallucchi, "KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic

Interpretations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 256–267. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.18

26. M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319
27. Google, "Google Research," 2023. [Online]. Available: https://github.com/google-research/google-research/tree/master
28. C. Van Der Lee, A. Gatt, E. Van Miltenburg, S. Wubben, and E. Krahmer, "Best practices for the human evaluation of automatically generated text," in *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, 2019, pp. 355–368. [Online]. Available: https://www.aclweb.org/anthology/W19-8643
29. Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," 2023, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/abs/2303.16634

Authors

Lasal Jayawardena is a fourth-year undergraduate student at the Informatic Institute of Technology (IIT) Sri Lanka, affiliated with Robert Gordon University in Aberdeen, Scotland. He is currently pursuing a Bachelor of Science honors degree in Artificial Intelligence and Data Science. Lasal's research interests primarily revolve around the domain of Natural Language Processing (NLP), with a strong focus on Large Language Models and Deep Learning techniques.

Prasan Yapa received Master of Computer Science (Research Major) and Bachelor of Science in Information Technology and Management from University of Moratuwa, Sri Lanka. Currently, he is pursuing his PhD in Computer Science & Engineering from Kyoto University of Advanced Science in Japan. His research interests include Natural Language Processing, Deep Learning, Affective Computing, Health Informatics and Computer Vision.

A Human and LLM Evaluation

Data Source	**Semantic Similarity: Scale from 1 to 5**	**Lexical Diversity: Scale from 1 to 5**	**Syntactic Diversity: Scale from 1 to 5**	**Grammatical Correctness: Scale from 1 to 5**
MSR Original	4.61	2.61	2.98	4.99
MSR Subset (Ours)	**4.99**	**3.56**	**3.91**	**4.99**
QQP Original	4.35	2.14	2.63	4.96
QQP Subset (Ours)	**4.97**	**3.52**	**3.82**	**4.99**
PAWS Original	4.87	1.30	1.82	4.86
PAWS Subset (Ours)	**4.95**	**3.20**	**3.88**	**4.98**
ParaNMT Original	4.49	1.34	1.41	4.78
ParaBank1 Original	4.66	1.34	1.40	4.81
ParaBank2 Original	4.07	2.07	2.26	4.27
Para-Common Subset (Ours)	**4.84**	**3.08**	**3.78**	**5.00**

Table 7: Full Breakdown of the LLM Evaluation using gpt-4

Data Source	**Semantic Similarity: Scale from 1 to 5**	**Lexical Diversity: Scale from 1 to 5**	**Syntactic Diversity: Scale from 1 to 5**	**Grammatical Correctness: Scale from 1 to 5**
MSR Original	4.26	2.66	2.96	4.80
MSR Subset (Ours)	**4.53**	**3.10**	**3.46**	**4.88**
QQP Original	4.25	2.38	2.62	4.92
QQP Subset (Ours)	**4.49**	**3.15**	**3.38**	**4.96**
PAWS Original	**4.79**	2.07	2.25	4.13
PAWS Subset (Ours)	4.42	**3.02**	**3.39**	**4.69**
ParaNMT Original	4.51	2.00	2.10	4.36
ParaBank1 Original	**4.66**	1.97	2.04	4.24
ParaBank2 Original)	4.11	2.37	2.59	4.04
Para-Common Subset (Ours)	4.42	**3.02**	**3.45**	**4.76**

Table 8: Full Breakdown of the Human Evaluation

Source Text: $source_text
Paraphrase: $paraphrase
Please evaluate the following aspects of the paraphrase in comparison to its source text on a likert scale of 1 to 5, where:
Semantic Similarity: This refers to how closely the meaning of the paraphrase matches the meaning of the source text.
Rating Scale for Semantic Similairty
1: The paraphrase has a completely different meaning or is unrelated to the source text.
2: The paraphrase has a somewhat different meaning from the source text
3: The paraphrase captures the general idea of the source text, but some details or nuances are missing.
4: The paraphrase largely captures the meaning of the source text but may have slight differences in wording or expression.
5: The paraphrase has an identical or nearly identical meaning to the source text.
Lexical Diversity: This aspect evaluates the range and richness of vocabulary used in the paraphrase, considering its comparison to the source text.
Rating Scale for Lexical Diversity
1: The paraphrase shows a limited use of words and lacks diversity when compared to the source text.
2: The paraphrase exhibits some variation in word choice but heavily relies on a few specific terms, which may not reflect the lexical diversity of the source text.
3: The paraphrase demonstrates moderate diversity in vocabulary, but there is room for improvement in terms of incorporating more varied word choices from the source text.
4: The paraphrase displays a good range of vocabulary, utilizing several different words and expressions that align with the lexical diversity of the source text.
5: The paraphrase showcases an extensive array of vocabulary, demonstrating excellent lexical diversity that closely matches or surpasses the richness of the source text.
Syntactic Diversity: This aspect assesses the structural variations in the paraphrase compared to the source text.
Rating Scale for Syntactic Diversity
1: The paraphrase closely mirrors the sentence structure of the source text with minimal variation.
2: The paraphrase shows some minor changes in sentence structure but largely follows the same pattern as the source text.
3: The paraphrase introduces moderate variations in sentence structure, deviating from the structure of the source text in certain aspects.
4: The paraphrase exhibits significant syntactic diversity, using different sentence structures while still conveying the same meaning as the source text.
5: The paraphrase displays a high level of syntactic diversity, employing various sentence structures creatively while maintaining the meaning of the source text.
Grammatical Correctness: This evaluates the grammatical accuracy of the paraphrase.
Rating Scale for Grammatical Correctness
1: The paraphrase contains numerous grammatical errors that significantly impact comprehension.
2: The paraphrase has several grammatical errors that occasionally affect understanding.
3: The paraphrase includes some grammatical errors, but they do not hinder overall comprehension.
4: The paraphrase demonstrates good grammatical correctness with only occasional minor errors.
5: The paraphrase is grammatically flawless, with no errors or inaccuracies.
Please provide your ratings for each aspect using the following json format:
{"Semantic Similarity": [Rating from 1 to 5],
"Lexical Diversity": [Rating from 1 to 5],
"Syntactic Diversity": [Rating from 1 to 5],
"Grammatical Correctness": [Rating from 1 to 5]}

Fig. 5: This figure illustrates the prompt fed to the gpt-4 model for evaluation.

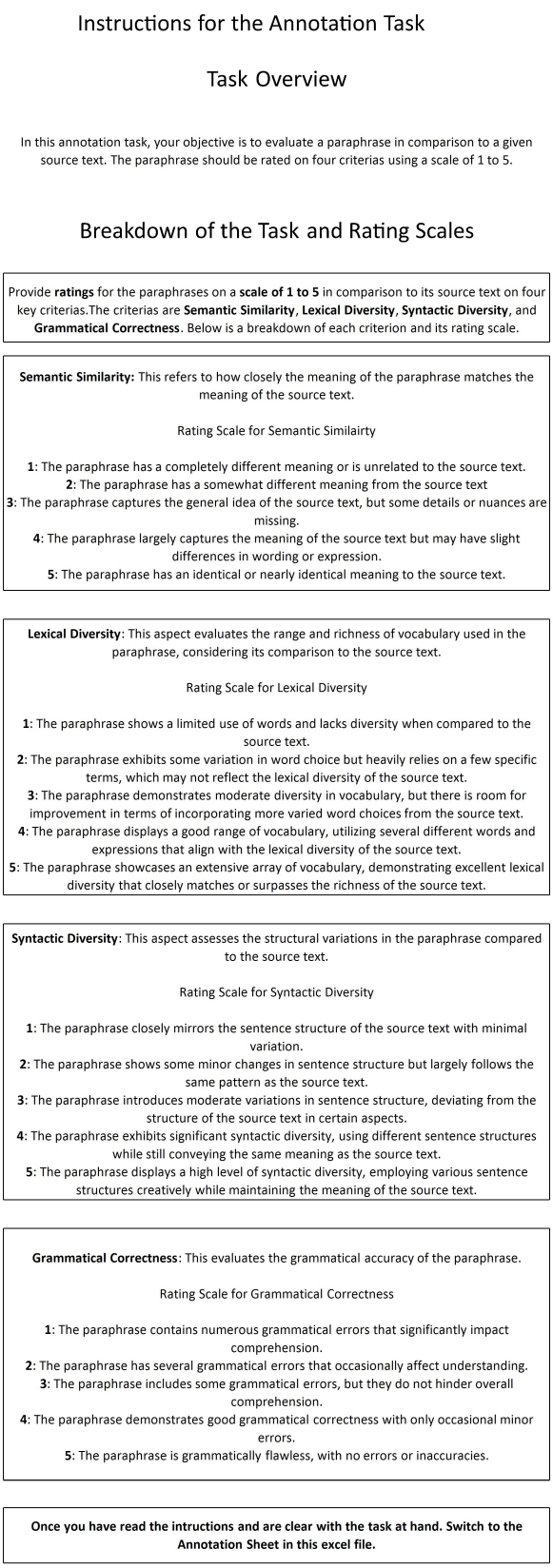

Fig. 6: Instructions given to Human Annotators and breakdown of the 5 point Likert Scale.