Improving quality and domain-relevancy of paraphrase generation with graph-based retrieval augmented generation.

JAYAWARDENA, L. and YAPA, P.

2025

Improving Quality and Domain-Relevancy of Paraphrase Generation with Graph-Based Retrieval Augmented Generation

Lasal Jayawardena*
lasalcj@gmail.com
Robert Gordon University
Aberdeen, United Kingdom

Prasan Yapa
prasan.y@iit.ac.lk
Informatics Institute of Technology
Colombo, Sri Lanka

ABSTRACT

Paraphrase generation is a fundamental area of research in Natural Language Processing (NLP) and Natural Language Generation (NLG), due to its sequence-to-sequence (Seq2Seq) nature. Paraphrasing, spanning across various domains, poses challenges for simpler model architectures due to the extensive knowledge required to generate paraphrases. The added constraint of generating diverse paraphrases further complicates the task for models trained on existing datasets. We present a methodology that leverages Graph-Based Retrieval Augmented Generation (G-RAG), capable of utilizing both entity and phrasal knowledge to address this issue. We demonstrate through experiments that this approach enables both complex models like Large Language models (LLMs) and smaller Seq2Seq models to generate more diverse paraphrases without compromising semantic similarity. Furthermore, this approach's capacity to integrate domain-specific knowledge makes it particularly effective across different domains, enhancing its applicability in varied contexts. The results are further corroborated by human evaluation and extensive quantitative analysis focusing on semantic similarity, lexical diversity, syntactic diversity, and grammatical correctness to gauge high-quality paraphrases.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; **Information extraction**; • **Information systems → Retrieval models and ranking**; **Similarity measures**; *Language models*; **Graph-based database models**.

KEYWORDS

Paraphrase Generation, Natural Language Processing, Large Language Models, Graph-based Knowledge, Sequence-to-Sequence Models

ACM Reference Format:
Lasal Jayawardena and Prasan Yapa. 2024. Improving Quality and Domain-Relevancy of Paraphrase Generation with Graph-Based Retrieval Augmented Generation. In *2024 10th International Conference on Computing and Artificial Intelligence (ICCAI 2024), April 26–29, 2024, Bali Island, Indonesia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3669754.3669784

1 INTRODUCTION

Paraphrase Generation (PG) entails the rephrasing of text while preserving its original meaning, serving as a vital tool in Natural Language Processing (NLP). This has been a longstanding endeavor within the realm of NLP for decades. Such a generation of varied linguistic expressions that maintain the same informational core serves as a powerful instrument for data augmentation. This augmentation is instrumental in elevating the performance across a spectrum of NLP tasks. The ability to create a plethora of semantically similar expressions enhances the diversity of training datasets, which in turn, bolsters the robustness and the ability of NLP models to generalize across different contexts [19, 29].

The application of PG is vast and spans many domains, enhancing the capabilities of systems in understanding and generating human-like dialogue. In the domain of question-answering systems, PG is not just an asset but a necessity for generating responses that are not only precise but also display a rich variety in phrasing. This variety mirrors the nuanced ways in which humans converse, leading to more natural interaction [9, 42, 48]. Beyond dialogue systems, the influence of PG extends to the field of information retrieval. Here, PG plays a pivotal role by expanding the range of expressions that can be recognized and matched to user inputs during their search queries. This expansion significantly enhances the likelihood of retrieving relevant and comprehensive information, thereby improving user experience and the efficiency of information access [1]. Moreover, PG is indispensable in the process of paraphrase identification, which involves the critical evaluation of textual material to determine if disparate segments deliver equivalent meanings. This process is foundational in ensuring the integrity and reliability of communicative exchanges across various platforms [2, 40, 44]. In the rapidly evolving landscape of chatbot technology, PG contributes to the enrichment of conversational dynamics, enabling these automated agents to engage users with a greater degree of authenticity and variety [41].

Despite the broad applications and their vital contributions, persisting challenges remain, particularly in handling unknown words, entities, and phrases that are often domain-specific. The emergence of Large Language Models (LLMs) has propelled the field forward [31], offering substantial improvements in language understanding and generation. Yet, even these advanced models encounter difficulties when dealing with out-of-vocabulary terms or unique jargon that are characteristic of specialized fields [53]. This limitation is particularly evident when these models, trained on expansive but generic datasets, are applied to niche domains. The models' performance can be compromised as they struggle to adapt to the

ICCAI 2024, April 26–29, 2024, Bali Island, Indonesia

ACM ISBN 979-8-4007-1705-5/24/04
https://doi.org/10.1145/3669754.3669784

terminologies and contextual nuances specific to each domain, leading to a decrease in the quality and relevance of the generated paraphrases.

In response, our paper introduces an innovative methodology designed to enhance the domain adaptability of paraphrase generation models. By leveraging the power of Graph-Based Retrieval Augmented Generation (G-RAG), our approach facilitates the dynamic integration of domain-specific knowledge into the paraphrase generation process. This not only improves the diversity and contextual relevance of the paraphrases but also ensures their applicability across diverse domains. The efficacy of our models is rigorously tested through extensive qualitative and quantitative evaluations, showcasing their ability to address the existing challenges in PG effectively.

2 RELATED WORK

2.1 Paraphrase Generation

The field of PG has evolved significantly over the years, transitioning from traditional methods to more advanced, contemporary approaches. Early methods in PG primarily relied on rule-based systems [22], thesaurus-based techniques [17], and Statistical Machine Translation (SMT) based approaches [46]. The advent of neural networks introduced neural-based approaches [35], which marked a substantial leap forward. Another significant development was the use of back translation [14, 45], providing new dimensions in paraphrase generation.

Recent advancements have seen the introduction of multi-round generation [24], which iteratively refines paraphrases, and reinforcement learning-based methods [25], which optimize paraphrasing through adaptive learning algorithms. Prompt-tuning [5] has emerged as a technique to fine-tune generation prompts, enhancing output quality. In addition to these, there has been a focus on increasing syntactic diversity through methods like sampling from latent spaces [4] and controlling word order [12]. However, these techniques often prioritize one aspect of diversity and may compromise on lexical or syntactic richness.

2.2 Knowledge Integration in Paraphrase Generation

The contribution by [30] laid the groundwork for knowledge integration within paraphrase generation for natural language question-answering systems. Their study focused on leveraging lexical knowledge and features, highlighting the importance of lexical variety for supporting diverse queries. However, the emphasis on lexical features might underplay the pragmatic nuances of paraphrasing, such as idiomatic usage and sentence flow, which are integral for natural-sounding language generation. Building on this foundational work, [50] established a baseline for integrating prior knowledge into Neural Machine Translation (NMT), employing posterior regularization to embed prior knowledge. This work set a precedent for knowledge integration in paraphrase generation, emphasizing the need for background knowledge to enhance semantic consistency in translations—a principle directly applicable to paraphrasing tasks.

Expanding on this, [28] introduced the concept of integrating topic knowledge as prior information in paraphrase generation. Their study illuminated the advantages of contextual awareness, showing that including topic information leads to paraphrases that are lexically diverse and topically coherent. However, they acknowledged the risk of topical bias, where paraphrases might overly adhere to the given topic, potentially overlooking other contextual details. Further exploring the role of knowledge integration, [47] focused on utilizing entity knowledge within an attention mechanism for enhancing the generation of paraphrased questions. While their method shows promise in producing context-rich questions, it is limited by its sole focus on entity knowledge, potentially missing out on a wider spectrum of semantic information.

In a different area, [26] investigated unsupervised paraphrasing guided by syntactic knowledge. They employed unsupervised learning techniques to adhere to syntactic constraints in paraphrasing. While this approach showcases the strength of unsupervised methods, it may sometimes yield syntactically accurate but semantically incongruent results due to the lack of explicit semantic direction. Lastly, [39] explored diversified paraphrase generation using commonsense graphs. Their approach, integrating structured world knowledge into paraphrasing, aims at generating paraphrases that are not only diverse but also embedded in commonsense reasoning. The primary challenge here lies in effectively translating complex commonsense relations into varied linguistic expressions without compromising accuracy or falling into generality.

While the above research work has integrated knowledge in various forms, our study focuses on integrating both phrasal and entity knowledge. We also leverage contextual embeddings with a ranking mechanism to leverage the knowledge effectively, which is vital when dealing with limited context windows.

3 METHODOLOGY

The present study introduces a novel method to conduct G-RAG in paraphrase generation through knowledge integration. The knowledge integration component is a sophisticated process designed for processing and integrating knowledge from graph data sources which are essentially knowledge bases tailored for a specific domain. The methodology also discusses the training process of the Seq2Seq models, which are used to test the effectiveness of the G-RAG. The inference flow of the LLM and the Seq2Seq models will also be detailed.

3.1 Knowledge Integration Component

The knowledge integration component is the implementation of G-RAG. The high-level architecture of the knowledge integration component can be seen in Fig. 1. The process commences with the input of a *Source Sentence*. This sentence is the foundation from which entities and keyphrases are to be identified and extracted, to later generate the model prompt. The next stages will be discussed subsequently.

3.1.1 Entity Recognition and Keyphrase Extraction: In this stage, the system identifies and recognizes distinct entities within the source sentence. Entities typically represent real-world objects, concepts, or names. Simultaneously, key phrases are extracted from the source sentence. These are phrases that capture the main topics or concepts discussed in the sentence.

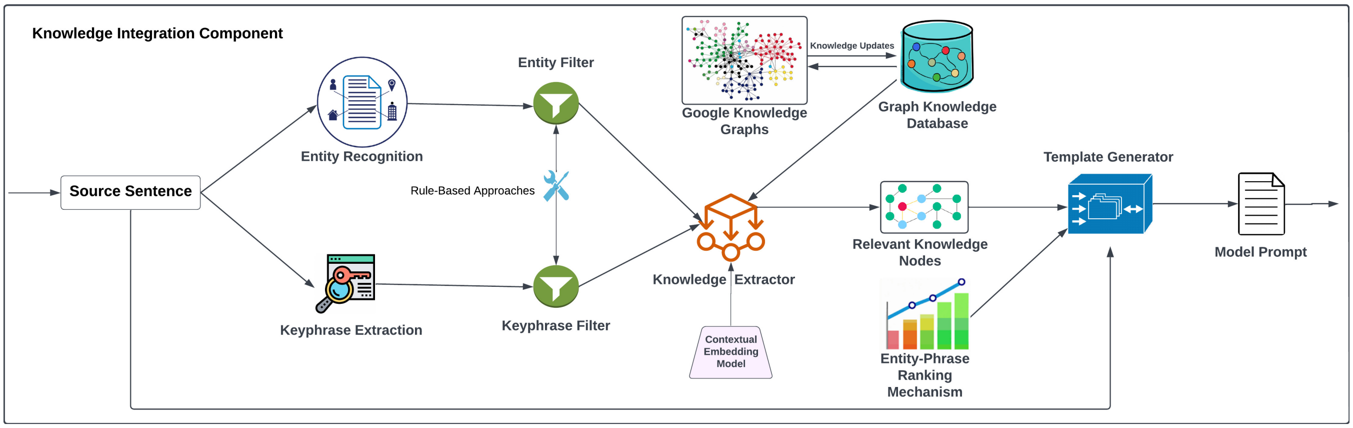


Figure 1: High-Level Architecture of the Knowledge Integration Component for G-RAG

For keyphrase and entity recognition, a combination of Azure language services [1] and the spaCy library [2] was used.

3.1.2 Entity and Keyphrase Filtering: The entities and keyphrases recognized in the previous step are filtered through a defined set of criteria to ensure the selection of only the most pertinent phrases and most important entities. We employ a rule-based approach for this. In this approach, we exclude entities that are numerical values, phone numbers, email addresses, and other types that typically do not contribute meaningful information. Furthermore, we apply a similar rule-based method to the keyphrases. Our objective here is to eliminate common phrases that do not provide significant insights. To identify these phrases and entities, we utilize regular expressions and analyze the types of entities returned from the entity recognition phase. This aids in highlighting unique and informative keyphrases and entities, enhancing the quality and relevance of our data for subsequent analysis.

3.1.3 Knowledge Extraction: The filtered entities and keyphrases are then passed to the Knowledge Extractor, which interacts with the graph knowledge database. This is a specialized knowledge base structured in a graph format, which contains both entity and phrasal knowledge. The initial knowledge base was built on entities and phrases built on the Quora Question Pair Dataset [15] and the PAWS Dataset [52], where the knowledge itself was obtained from Google Knowledge Graphs[3]. The knowledge base is stored as a Neo4J[4] graph database, but this can easily be interchanged with other graph-like storage methods like ontologies. Full visualization of the knowledge base and node attributes can be seen in Fig. 2 which was built using NeoDash[5]. The knowledge base initially has close to a million knowledge nodes which can be used in knowledge extraction. In cases where new entities to phrases are encountered new knowledge is updated using the source graphs. The extraction is done by first using an exact match on an entity or phrase and later based on a similarity basis of the node *text* attribute by leveraging the Cypher Query Language. The most relevant nodes are then found by utilizing contextual embeddings. By default, we use OpenAI's text-embedding-ada-002 model [32] due to its state-of-the-art performance, but this could be easily interchanged.

3.1.4 Template Generation: The culmination of the knowledge integration process is the *Template Generator*. This component takes the relevant knowledge nodes and the ranks entities and phrases to construct a model prompt. This prompt encapsulates the integrated knowledge in a format that is primed for paraphrase generation by using the most important entity and phrasal knowledge for the context size of the model.

When selecting the most appropriate entity and phrases, two separate methods are followed. Given a set of entities E, each entity $e \in E$ is represented by a tuple containing its category, subcategory, confidence score, and maximum similarity score to a knowledge node (cosine score). The ranking function R assigns a priority to each entity based on its category, its confidence score, and its maximum similarity score.

First, we define a priority function P for the categories as follows:

$$P(\text{category}) = \begin{cases} 1 & \text{if category is 'Person'}, \\ 2 & \text{if category is 'Location'}, \\ 3 & \text{if category is 'Organization'}, \\ 4 & \text{if category is 'Event'}, \\ 5 & \text{if category is 'PersonType'}, \\ 6 & \text{if category is 'Product'}, \\ 7 & \text{if category is 'Skill'}, \\ 8 & \text{otherwise}. \end{cases} \tag{1}$$

The ranking of entities is then performed by a ranking function $R(E)$, which sorts entities based on their category priority, confidence score, and maximum similarity score:

$$R(E) = \text{sort}\,(E, \text{key} = \left(P(e_{\text{category}}), -e_{\text{confidence}}, e_{\text{max\_similarity\_score}}\right)) \tag{2}$$

[1] https://learn.microsoft.com/en-us/azure/ai-services/language-service/
[2] https://github.com/explosion/spaCy
[3] https://developers.google.com/knowledge-graph
[4] https://neo4j.com/
[5] https://neo4j.com/labs/neodash/

Here, e_{category}, $e_{\text{confidence}}$, and $e_{\text{max\_similarity\_score}}$ represent the category, confidence score, and maximum similarity score of the entity e, respectively. The minus sign before the confidence score indicates that a higher confidence score results in a higher rank.

After reaching a certain minimum threshold in the similarity scores for entities, keyphrases are included in the ranking. These key phrases are considered only if their similarity scores are above the same threshold. This ensures that both entities and keyphrases included in the final ranking are of high relevance and similarity to the knowledge nodes.

The final product of the Knowledge Integration Component is the *Model Prompt*, which is a structured textual representation of the integrated knowledge, ready to be utilized. The prompt is generated with the help of a custom template generator, that creates the prompt with respect to the model context size and the relevant knowledge nodes of the sentence. This prompt is built in the most optimized manner so that it doesn't have redundant tokens to maximize what can fit into the model. This final prompt represents the result of applying G-RAG.

3.2 Seq2Seq Model Training

In the study, we also utilized small Seq2Seq models to integrate with G-RAG. Two distinct sets of models and datasets were created for this purpose and sequence-level knowledge distillation[18] was applied. This was done to demonstrate the effectiveness of the G-RAG component.

3.2.1 Dataset Creation: Initially, we generated a dataset using the ChatGPT (gpt-3.5-turbo) LLM, primarily leveraging the Quora Dataset and PAWSWiki [52], while consciously avoiding PAWSQQP due to its overlap with Quora. A similar second dataset was then created using ChatGPT in conjunction with G-RAG. The paraphrase pair generation for both datasets involved filtering out offensive content using OpenAI's Moderation Endpoint. This process yielded nearly 2 million unique sentence pairs for each dataset. The purpose of this was to create one set of models that do not use G-RAG and the other that can integrate it.

3.2.2 Training Models: For the training phase, we distilled two sets of three models each: T5-small [37], Flant5-small [6], and BART-base [21]. The first set was distilled with the dataset created without G-RAG, while the second set was distilled in conjunction with G-RAG. The training was conducted using the parameter efficient Low-Rank Adaptation of Large Language Models (LoRA) method [13], with hyperparameters optimized for both efficiency and G-RAG compatibility. The LoRA technique was applied through the use of a library specifically designed for Parameter-Efficient Fine-Tuning (PEFT), developed by Hugging Face[6].

Before choosing LoRa, we ran preliminary experiments where we tried using vanilla fine-tuning of the models, applying Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning (AdaLoRA) [51], and also running Efficient Finetuning of Quantized LLMs (QLoRA) [7]. The vanilla fine-tuning process would have been ideal given a large allocation of resources for tuning, especially GPUs, but since we were working with a large corpus and limited GPUs, the resources were not sufficient for this approach. Thus making a parameter efficient method suitable. Out of the parameter efficient techniques, LoRA was the most stable in model training compared to AdaLoRA and QLoRA. The preliminary experiments converged faster for LoRA and sample inference showed the model had adapted better compared to the other two. This was the reason we chose LoRA as the main training methodology for the models.

Each model was trained on the corresponding datasets with techniques like lower casing and setting the maximum sequence length to 512. We employed the Adam optimizer, an epsilon value of 1e-08, a learning rate of 0.0003, and a dropout rate of 0.1. The LoRA configuration included a rank of 8 and an alpha value of 32. Training times varied, with T5-small requiring 24 hours, Flant5-small 30 hours, and BART-base 50 hours, using RTX A100 40GB GPUs. The main difference between models is that one set is accustomed to the G-RAG prompt structure that incorporates the knowledge whereas the other follows the convention prefix-based generation method where in all models, "paraphrase" was used as the prefix for the input sentences during training.

3.3 Model Inference

Model inference involves two distinct approaches: standard model inference and inference using G-RAG. In standard model inference, each model (ChatGPT, BART, T5, and Flan T5) was configured with specific hyperparameters to optimize performance. These parameters included settings for maximum token generation, early stopping, sampling methods, temperature for randomness control, and n-gram repetition restrictions. For the inference, the LoRA weights will be loaded to the model architecture and inference will be conducted on top of the tuned architecture. Post-processing steps were taken to enhance the output's readability and accuracy. This involved capitalization of the start character, performing entity case correction by leveraging the SpaCy NER model, and applying de-duplication to maintain output diversity.

Inference with G-RAG involves an additional step of generating prompts using the knowledge integration component before passing them to the models. This is the main difference between G-RAG compared to the standard inference. Before the Inference when the source sentence is given for the paraphrase, the knowledge integration is done as shown in Fig.1. This process includes configuring the context window size in the knowledge integrator for each model, including ChatGPT. The context window size is important because it determines the amount of knowledge that can be passed in for model inference. For smaller models limited knowledge is passed, whereas in models with larger context windows, more knowledge can be integrated making the quality of inference higher. The order of knowledge injected, by the template generator, for inference is determined by the ranking mechanism discussed earlier in Section 3.1.4.

Once the prompt is generated from the template generator, this is used as the input for the LoRA tuned model. The model weights especially tuned for the G-RAG models will be loaded with LoRA. These weights are different from the standard inference models. The inference parameters for each model were adjusted accordingly to accommodate the additional context provided by G-RAG. The post-processing steps for G-RAG inference are similar to those of standard model inference, focusing on enhancing readability

[6] https://github.com/huggingface/peft

and ensuring the accuracy and uniqueness of the generated paraphrases. By employing both standard and G-RAG augmented inference methods, we aimed to explore the differences in output quality and relevance.

4 EVALUATION

In assessing the effectiveness of the G-RAG approach, we employed a comprehensive evaluation strategy, encompassing both quantitative and qualitative methods. Our quantitative evaluation was grounded in the analysis of a diverse set of just more than 100,000 paraphrase pairs, carefully curated from the Wiki Answer[10], MRPC [8], Twitter URL [20], and MSCOCO [23] datasets. This approach allowed us to gauge the models' performance with G-RAG and without it for both the LLM and Seq2Seq models. Complementing this, our qualitative evaluation involved a dual approach: first, engaging human annotators to provide insights on the paraphrases, and second, utilizing an innovative evaluation methodology leveraging LLMs. This section details the results and insights gleaned from this multifaceted evaluation process, which follows a gold standard evaluation strategy that is not commonly seen in most paraphrase generation research [54].

4.1 Quantitative Analysis

The focus of our quantitative analysis lies in determining the quality and variety of the paraphrases. This evaluation will concentrate on three critical aspects: similarity in semantics, diversity in syntax, and lexical diversity.

4.1.1 Semantic Similarity: In this study, our approach to evaluating semantic similarity involves generating sentence embeddings from both the original texts and their paraphrases using a variety of models. The similarity between these embeddings is quantified using the cosine similarity method. This involves computing the scores by subtracting the cosine similarity metric from one, and comparing the base text with its paraphrased counterpart.

For this purpose, several models have been utilized. The Ada Score is calculated with the assistance of the text-embedding-ada-002 model developed by OpenAI [32]. Meanwhile, the SimCSE Score is obtained using the sup-simcse-roberta-large model created by SimCSE [11], and the PromCSE Score leverages the sup-promcse-roberta-large model from PromCSE [16]. Additionally, models from the sentence-transformers library [38] are employed in this analysis. The Mpnet Score is extracted by using the all-mpnet-base-v1 model, and the Roberta Score is determined through the all-roberta-large-v1 model.

The results, as illustrated in Table 1, reveal that applying G-RAG has not comprised the semantic similarity, and in some cases, it has improved the model performance in this aspect. The variations in similarity scores are minor and likely stem from lexical differences and sentence structure differences.

4.1.2 Syntactic diversity. The concept of syntactic diversity in paraphrasing is quantified by comparing the complexity and range of sentence structures in a paraphrased text against the original. This aspect, a key indicator of paraphrase quality, signifies that the rephrased sentences exhibit not just variation but also depth in linguistic structure. To assess this, specific metrics that focus on the syntactic structures of sentences are employed.

For a detailed Tree Edit Distance analysis, the "Ted-F" metric is utilized. This involves forming constituency parse trees for both the paraphrased and original sentences using Stanza[36]. These trees are then converted into bracket notation using regex and the NLTK library [3]. The comprehensive Tree Edit Distance is subsequently computed using the APTED library [33].On the other hand, the "Ted-3" metric narrows its focus to the top three levels of the Tree Edit Distance. While it follows a similar procedure to "Ted-F", it specifically calculates the Tree Edit Distance for the tree's first three layers.

Employing the Kermit library [49], the "Kermit Score" is derived by first calculating the cosine similarity between the syntactic vectors of the original and paraphrased sentences. This similarity measure is then subtracted from one, with the syntactic embeddings being generated from the syntax trees of the sentences.

The diversity measured by the "Subtree K Score" pertains to the Subtree Kernel. It starts by creating constituency parse trees for both sentences using Stanza, transforming them into NLTK Tree format, and then identifying all subtrees. The diversity score is determined by the proportion of unique common subtrees with the total unique subtrees, subtracted from one. Similarly, the "Node Pair K Score" calculates Subtree Node Pair Kernel diversity, differing from the "Subtree K Score" only in its use of node pairs rather than subtrees.

The findings, as presented in Table 2, demonstrate that utilizing G-RAG improves the syntactic diversity quite significantly. The improvement is seen in both the LLM and the Seq2Seq model. Given the improvement was achieved with no comprise in semantic similarity is a huge achievement.

4.1.3 Lexical diversity. Lexical diversity refers to the variety and range of words used in a text, reflecting the breadth of vocabulary and the use of synonyms. This concept is particularly important in paraphrasing, as it helps assess the extent of vocabulary variation. To measure lexical diversity, we employed a variety of metrics.

The "BOW Overlap Score" measures the commonality of tokens between the original and paraphrased texts. This is done by calculating the shared tokens' proportion against the total token count and then deducting this value from one.

The "Corpus BLEU Score" and "Corpus BLEU2 Score" are both derived using the SacreBLEU Library [34], with the latter employing a specific smoothing function known as "method1". Both scores are adjusted by subtracting them from one.

Similar to the Corpus BLEU score but on a sentence level, the "Sentence BLEU Score" is computed using the SacreBLEU Library and then reduced by one.

The "METEOR Score", calculated with the NLTK library, and the "ROUGE Scores" (ROUGE 1, ROUGE 2, and ROUGE L), computed using the Google Research library[7], are all adjusted by subtracting one from each.

The "Token ∩/∪ Score" closely resembles the BOW Overlap Score but differs slightly in its calculation. It's based on the proportion of shared tokens to the total unique tokens, with the final value being reduced by one.

[7] https://github.com/google-research/google-research

Table 1: Semantic Similarity Scores Comparison of Models with and without G-RAG.

Model	ADA Score (↑)	SimCSE Score (↑)	PromCSE Score (↑)	Roberta Score (↑)	Mpnet Score (↑)
ChatGPT					
w/o G-RAG	95.60%	91.25%	99.41%	88.23%	87.03%
w/ G-RAG	**96.24%**	**93.14%**	**99.60%**	**90.20%**	**90.91%**
T5 Small (Distilled)					
w/o G-RAG	**97.28%**	**94.59%**	99.67%	**92.77%**	**92.60%**
w/ G-RAG	96.92%	93.85%	**99.76%**	92.09%	91.88%
Flan T5 Small (Distilled)					
w/o G-RAG	**97.75%**	**95.42%**	**99.71%**	**93.71%**	**93.69%**
w/ G-RAG	96.74%	94.34%	99.66%	92.61%	92.64%
BART Base (Distilled)					
w/o G-RAG	**98.07%**	**95.77%**	**99.72%**	**94.04%**	**93.77%**
w/ G-RAG	96.80%	95.61%	99.76%	92.98%	93.37%

Table 2: Syntactic Diversity Scores Comparison of Models with and without G-RAG.

Model	Ted-F Score (↑)	Ted-3 Score (↑)	Kermit Score (↑)	Subtree K Score (↑)	Node Pair K Score (↑)
ChatGPT					
w/o G-RAG	21.24	4.53	66.94%	92.77%	79.20%
w/ G-RAG	**21.29**	**4.48**	**70.88%**	**94.96%**	**84.54%**
T5 Small (Distilled)					
w/o G-RAG	17.29	3.99	54.89%	83.36%	66.58%
w/ G-RAG	**18.45**	**4.23**	**62.97%**	**87.33%**	**73.25%**
Flan T5 Small (Distilled)					
w/o G-RAG	18.38	4.40	54.96%	83.23%	65.45%
w/ G-RAG	**18.68**	**4.42**	**64.30%**	**86.55%**	**71.68%**
BART Base (Distilled)					
w/o G-RAG	23.45	5.06	61.98%	88.97%	72.30%
w/ G-RAG	**23.79**	**5.05**	**65.86%**	**89.63%**	**73.31%**

The "Google BLEU Score" is determined using Huggingface's Evaluate library[8], with the score being decreased by one. Similarly, the "TER Score" (Translation Error Rate), "WER Score" (Word Error Rate), and "CharacTER Score" (Character Error Rate) are all calculated using the same library.

As shown in Table 3 and Table 4, the models that integrated G-RAG show an improvement in lexical diversity. Manual inspection further showed that the G-RAG prompt encourages the model to generate longer paraphrases. This is especially beneficial when dealing with data augmentation in certain domain-specific tasks.

4.2 Qualitative Analysis

To gain comprehensive insights into our model's performance, we undertook a qualitative analysis through two primary approaches: evaluations by human reviewers and assessments using an LLM.

Human Evaluation: For the human evaluation, we engaged five independent reviewers proficient in English. The task involved examining 3200 pairs of paraphrases sourced from four diverse datasets: MRPC, MSCOCO evaluation subset, Twitter URL, and Wiki Answer. Each dataset source and model pair contributed 100 pairs, ensuring a balanced mix for each model set. The selection was meticulously done to represent the sentence length variation accurately across sources. The paraphrases given for human evaluation are a representation of the benchmark datasets used for quantitative analysis.

Our evaluation criteria were based on a 5-point Likert scale, as delineated in [43], focusing on Semantic Similarity, Lexical Diversity, Syntactic Diversity, and Grammatical Correctness. The criteria were defined as follows:

Semantic Similarity was rated on a scale where a score of 5 indicated a near-perfect alignment in meaning with the source text, encompassing similar ideas, conclusions, or arguments. A score of 1, on the other hand, indicated a complete divergence in meaning, reflecting different ideas or arguments.

Lexical Diversity was assessed with a score of 5 representing a wide and rich vocabulary, marked by the use of diverse words, synonyms, and phrases, and minimal repetition. A score of 1 indicated

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

Table 3: Lexical Diversity Scores Comparison of Models with and without G-RAG (Part 1).

Model	Lexical BOW (↑)	Corpus BLEU (↑)	Sentence BLEU (↑)	METEOR (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)
ChatGPT						
w/o G-RAG	50.55%	99.54%	84.01%	43.20%	42.56%	70.15%
w/ G-RAG	**53.97%**	**99.60%**	**86.90%**	**45.06%**	**46.46%**	**73.79%**
T5 Small (Distilled)						
w/o G-RAG	35.42%	99.46%	65.79%	29.50%	28.19%	50.78%
w/ G-RAG	**38.88%**	**99.48%**	**68.69%**	**30.66%**	**30.30%**	**52.03%**
Flan T5 Small (Distilled)						
w/o G-RAG	34.06%	99.48%	63.93%	27.80%	26.23%	47.05%
w/ G-RAG	**39.76%**	**99.52%**	**71.28%**	**32.73%**	**33.19%**	**50.12%**
BART Base (Distilled)						
w/o G-RAG	39.49%	99.50%	75.71%	36.27%	33.51%	59.79%
w/ G-RAG	**42.48%**	**99.53%**	**76.02%**	**36.70%**	**33.96%**	**60.82%**

Table 4: Lexical Diversity Scores Comparison of Models with and without G-RAG (Part 2).

Model	ROUGE-L (↑)	Token Intersection (↑)	Translation Edit Rate (↑)	Word Edit Rate (↑)	CharacTER (↑)
ChatGPT					
w/o G-RAG	54.17%	62.93%	63.05	77.44	77.90
w/ G-RAG	**56.85%**	**66.09%**	**66.11**	**78.06**	**85.95**
T5 Small (Distilled)					
w/o G-RAG	42.17%	43.52%	49.05	66.83	54.80
w/ G-RAG	**43.14%**	**48.60%**	**50.13**	**67.15**	**57.50**
Flan T5 Small (Distilled)					
w/o G-RAG	41.98%	40.94%	48.77	69.97	56.02
w/ G-RAG	**43.76%**	**51.48%**	**51.84**	**68.84**	**63.31**
BART Base (Distilled)					
w/o G-RAG	52.11%	49.68%	59.23	82.47	68.00
w/ G-RAG	**52.89%**	**54.17%**	**59.44**	**73.65**	**68.95**

a narrow range of vocabulary, characterized by repetitive usage of limited words or phrases.

Syntactic Diversity was measured where a score of 5 denoted significant variation in sentence structures, including a mix of sentence types, lengths, and constructions. A score of 1 suggested little to no variation in sentence structure, with reliance on limited sentence forms and frequent repetition.

Grammatical Correctness was evaluated with a score of 5 representing flawless grammar, including accurate punctuation, spelling, and syntax. Conversely, a score of 1 indicated substantial grammatical errors that impacted understanding, such as frequent mistakes in spelling, punctuation, or syntax.

The guidelines provided to the human reviewers are detailed in Fig. 4. It's important to note that while grammatical correctness is often assessed in similar research, it's particularly vital in evaluating the efficacy of paraphrasing. The results from this human evaluation are compiled in Table 5. The results further prove that using G-RAG has indeed enhanced the model performance to generate more high-quality paraphrases that are diverse.

4.2.1 LLM Evaluation: The recent surge in the application of LLMs for NLP evaluations is noteworthy. As highlighted in [27], LLMs have shown superior capabilities over traditional reference-free metrics, leading to their increased adoption in NLP research. Our study aligns with this trend, incorporating an LLM-based evaluation method, specifically utilizing the gpt-4 model from OpenAI, which was at the forefront of LLM technology during our research period [31].

For our LLM evaluation, the dataset used mirrored that of the human evaluators. To ensure consistency and a level playing field with the human evaluation, we carefully crafted the LLM prompts, replicating the guidelines provided to human annotators. These prompts are detailed in Fig. 3.

This methodological innovation in our research could significantly influence future studies in this domain. It underscores the importance and potential of LLMs in NLG evaluations, offering insights for future research and advancements in evaluation methods. The results derived from this LLM evaluation are presented in Table 6.

Table 5: Results of Human Evaluation on Models with and without G-RAG.

Model	Semantic Similarity (↑)	Lexical Diversity (↑)	Syntactic Diversity (↑)	Grammatical Correctness (↑)
ChatGPT				
w/o G-RAG	4.33	3.43	3.26	**4.97**
w/ G-RAG	**4.40**	**3.58**	**3.37**	4.96
T5 Small (Distilled)				
w/o G-RAG	4.21	3.23	2.95	4.83
w/ G-RAG	**4.22**	**3.35**	**3.01**	**4.88**
Flan T5 Small (Distilled)				
w/o G-RAG	4.05	3.19	2.83	4.74
w/ G-RAG	**4.20**	**3.32**	**3.04**	**4.86**
BART Base (Distilled)				
w/o G-RAG	**4.20**	3.25	3.16	**4.89**
w/ G-RAG	4.19	**3.30**	**3.20**	4.88

Table 6: Results of LLM Evaluation on Models with and without G-RAG.

Model	Semantic Similarity (↑)	Lexical Diversity (↑)	Syntactic Diversity (↑)	Grammatical Correctness (↑)
Original Dataset	3.40	2.83	2.88	4.37
ChatGPT				
w/o G-RAG	4.82	3.51	3.85	4.89
w/ G-RAG	**4.85**	**3.60**	**3.89**	**4.90**
T5 Small (Distilled)				
w/o G-RAG	**4.51**	2.48	3.20	4.68
w/ G-RAG	4.44	**2.91**	**3.36**	**4.82**
Flan T5 Small (Distilled)				
w/o G-RAG	4.26	2.20	2.85	4.41
w/ G-RAG	**4.33**	**3.01**	**3.41**	**4.78**
BART Base (Distilled)				
w/o G-RAG	**4.75**	2.63	3.41	4.69
w/ G-RAG	4.70	**2.89**	**3.63**	**4.71**

5 DISCUSSION

The introduction of the G-RAG approach in paraphrase generation has demonstrated notable improvements in various aspects of language model performance in paraphrase generation. The results from both quantitative and qualitative evaluations indicate that G-RAG enhances syntactic diversity, and lexical diversity in the generated paraphrases while retaining semantic similarity and in some cases enhancing that too. One of the key findings is the consistent improvement is that improvement is seen across varied model sizes where it was integrated into an LLM and also several Seq2Seq models which are in comparison small to the LLM. The ranking mechanism that utilizes the priority function, confidence scores, and contextual similarity helps rank the most appropriate knowledge to be integrated into the model. The flexibility of G-RAG is most effective when dealing with limited model context sizes, where only limited knowledge can be integrated. All of the models irrespective of size saw an improvement with the utilization of G-RAG.

In our analysis, a comprehensive set of evaluation metrics was utilized to thoroughly examine the different facets of the paraphrases produced. We adopted this multi-metric approach acknowledging that relying solely on a single metric might not adequately reflect the true effectiveness of a paraphrase. This is because single metrics often concentrate on just one aspect of the paraphrase. This consideration underscores a significant and ongoing research challenge. Relying on metrics that are either inappropriate or too narrow can result in a misleading interpretation of a model's capabilities. Future studies, therefore, should focus on creating an all-encompassing metric that can more comprehensively assess the quality of paraphrases. The development of such a metric would not only yield a more accurate evaluation of model performance but also significantly contribute to progressing the field of paraphrase generation.

Interestingly, the improvement in lexical diversity hints at G-RAG's potential to aid models in avoiding common pitfalls like repetitive or overly simplistic language use. Additionally, manual observations show that it can be used for improving domain relevance as well. The knowledge integration of G-RAG is optimally used with the help of the ranking mechanism and fitting as much information as possible with respect to the model context size. The flexibility of the knowledge extractor and ranking mechanism allows newer research to use other data sources like ontologies in the future.

Conducting a long-term study to assess the impact of G-RAG on model learning and adaptation over time would provide valuable insights into the sustainability and evolution of this approach would be valuable. Finally, extending G-RAG to support multiple languages and evaluating its effectiveness in cross-lingual contexts could greatly enhance its applicability in global settings.

6 CONCLUSION

This study has successfully demonstrated the efficacy of Graph-Based Retrieval Augmented Generation (G-RAG) in enhancing the quality and domain relevance of paraphrase generation. Through our innovative approach, both LLMs and smaller Seq2Seq models have shown marked improvements in generating diverse paraphrases while maintaining semantic accuracy. Particularly noteworthy is the ability of G-RAG to incorporate domain-specific knowledge, significantly broadening the applicability of our methodology across various fields. G-RAG is able to maximize the amount of knowledge integrated into the models, with respect to varying model context window sizes and it integrates the most appropriate information with the help of a unique ranking mechanism. Rigorous evaluation through a blend of human judgment and quantitative analysis has validated the improvement of our models in lexical richness, and syntactic variation. This research not only addresses existing challenges in paraphrase generation but also sets a new benchmark for future developments in this critical area of NLP and NLG.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the individuals who engaged in the human evaluation of our models. The contributions and constructive critiques provided by these research participants have been invaluable in evaluating the performance and impact of our study.

REFERENCES

[1] Koichi Akabe, Toshiki Takeuchi, Takashi Aoki, and Kunihiro Nishimura. 2021. Information retrieval on oncology knowledge base using recursive paraphrase lattice. *Journal of Biomedical Informatics* 116 (April 2021), 103705. https://doi.org/10.1016/j.jbi.2021.103705

[2] Mohammad AL-Smadi, Zain Jaradat, Mahmoud AL-Ayyoub, and Yaser Jararweh. 2017. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management* 53, 3 (May 2017), 640–652. https://doi.org/10.1016/j.ipm.2017.01.002

[3] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

[4] Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards Diverse Paraphrase Generation via Diversified Generative Adversarial Network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2411–2421. https://doi.org/10.18653/v1/2020.findings-emnlp.218

[5] Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty Controlled Paraphrase Generation with Retrieval Augmented Conditional Prompt Tuning. (2022). https://doi.org/10.48550/ARXIV.2202.00535

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. (2022). https://doi.org/10.48550/ARXIV.2210.11416

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. http://arxiv.org/abs/2305.14314 arXiv:2305.14314 [cs].

[8] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005).* https://aclanthology.org/I05-5002

[9] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to Paraphrase for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Copenhagen, Denmark, 875–886. https://doi.org/10.18653/v1/D17-1091

[10] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Sofia, Bulgaria, 1608–1618. https://aclanthology.org/P13-1158

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[12] Tanya Goyal and Greg Durrett. 2020. Neural Syntactic Preordering for Controlled Paraphrase Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 238–252. https://doi.org/10.18653/v1/2020.acl-main.22

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. (2021). https://doi.org/10.48550/ARXIV.2106.09685

[14] J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-Scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL).* Association for Computational Linguistics, Hong Kong, China, 44–54. https://doi.org/10.18653/v1/K19-1005

[15] Shankar Iyer, Nikhil Dandeka, and Kornél Csernai. 2017. First Quora Dataset Release: Question Pairs. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[16] Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022.* Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3021–3035. https://aclanthology.org/2022.findings-emnlp.220

[17] David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference.* Association for Computational Linguistics, New York City, USA, 455–462. https://aclanthology.org/N06-1058

[18] Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1317–1327. https://doi.org/10.18653/v1/D16-1139

[19] Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the second international workshop on Paraphrasing -*, Vol. 16. Association for Computational Linguistics, Sapporo, Japan, 1–8. https://doi.org/10.3115/1118984.1118985

[20] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A Continuously Growing Dataset of Sentential Paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Copenhagen, Denmark, 1224–1234. https://doi.org/10.18653/v1/D17-1126

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. (2019). https://doi.org/10.48550/ARXIV.1910.13461

[22] Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering* 7, 4 (Dec. 2001), 343–360. https://doi.org/10.1017/S1351324901002765

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common

Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Vol. 8693. Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 Series Title: Lecture Notes in Computer Science.

[24] Zhe Lin and Xiaojun Wan. 2021. Pushing Paraphrase Away from Original Sentence: A Multi-Round Paraphrase Generation Approach. (2021). https://doi.org/10.48550/ARXIV.2109.01862

[25] Mingtong Liu, Erguang Yang, Deyi Xiong, Yujie Zhang, Yao Meng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2020. A Learning-Exploring Method to Generate Diverse Paraphrases with Multi-Objective Deep Reinforcement Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 2310–2321. https://doi.org/10.18653/v1/2020.coling-main.209

[26] Tianyuan Liu, Yuqing Sun, Jiaqi Wu, Xi Xu, Yuchen Han, Cheng Li, and Bin Gong. 2023. Unsupervised Paraphrasing under Syntax Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 11 (June 2023), 13273–13281. https://doi.org/10.1609/aaai.v37i11.26558

[27] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. (2023). https://doi.org/10.48550/ARXIV.2303.16634 Publisher: arXiv Version Number: 3.

[28] Yuanxin Liu, Zheng Lin, Fenglin Liu, Qinyun Dai, and Weiping Wang. 2019. Generating Paraphrase with Topic as Prior Knowledge. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, Beijing China, 2381–2384. https://doi.org/10.1145/3357384.3358102

[29] Marie Meteer and Varda Shaked. 1988. Strategies for effective paraphrasing. In *Proceedings of the 12th conference on Computational linguistics -*, Vol. 2. Association for Computational Linguistics, Budapest, Hungry, 431–436. https://doi.org/10.3115/991719.991724

[30] Kyo-Joong Oh, Ho-Jin Choi, Gahgene Gweon, Jeong Heo, and Pum-Mo Ryu. 2015. Paraphrase generation based on lexical knowledge and features for a natural language question answering system. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*. IEEE, Jeju, South Korea, 35–38. https://doi.org/10.1109/35021BIGCOMP.2015.7072846

[31] OpenAI. 2023. GPT-4 Technical Report. (2023). https://doi.org/10.48550/ARXIV.2303.08774 Publisher: arXiv Version Number: 3.

[32] OpenAI. 2023. New and Improved Embedding Model. https://openai.com/blog/new-and-improved-embedding-model Publisher: OpenAI.

[33] Mateusz Pawlik and Nikolaus Augsten. 2015. Efficient Computation of the Tree Edit Distance. *ACM Transactions on Database Systems* 40, 1 (March 2015), 1–40. https://doi.org/10.1145/2699485

[34] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://doi.org/10.18653/v1/W18-6319

[35] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural Paraphrase Generation with Stacked Residual LSTM Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2923–2934. https://aclanthology.org/C16-1275

[36] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. (2019). https://doi.org/10.48550/ARXIV.1910.10683

[38] Nils Reimers and Iryna Gurevych. 2019. 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. http://arxiv.org/abs/1908.10084

[39] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the Evaluation Metrics for Paraphrase Generation. http://arxiv.org/abs/2202.08479 arXiv:2202.08479 [cs].

[40] Xinyao Shen, Jiangjie Chen, and Yanghua Xiao. 2021. Diversified Paraphrase Generation with Commonsense Knowledge Graph. In *Natural Language Processing and Chinese Computing*, Lu Wang, Yansong Feng, Yu Hong, and Ruifang He (Eds.). Vol. 13028. Springer International Publishing, Cham, 353–364. https://doi.org/10.1007/978-3-030-88480-2_28 Series Title: Lecture Notes in Computer Science.

[41] Darlyn Sirikasem and Siripen Pongpaichet. 2022. Thai Paraphrasing Tool for Chatbot Intent Recognition Training. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*. IEEE, Sakon Nakhon, Thailand, 111–116. https://doi.org/10.1109/ICSEC56337.2022.10049337

[42] Sarvesh Soni and Kirk Roberts. 2019. A Paraphrase Generation System for EHR Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 20–29. https://doi.org/10.18653/v1/W19-5003

[43] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 355–368. https://doi.org/10.18653/v1/W19-8643

[44] Tedo Vrbanec and Ana Meštrović. 2020. Corpus-Based Paraphrase Detection Experiments and Review. *Information* 11, 5 (April 2020), 241. https://doi.org/10.3390/info11050241

[45] John Wieting and Kevin Gimpel. 2018. 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)*. Association for Computational Linguistics, Melbourne, Australia, 451–462. https://doi.org/10.18653/v1/P18-1042

[46] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2010. Paraphrase Generation as Monolingual Translation: Data and Evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics. https://aclanthology.org/W10-4223

[47] Justin J. Xie and Ameeta Agrawal. 2023. Emotion and Sentiment Guided Paraphrasing. http://arxiv.org/abs/2306.05556 arXiv:2306.05556 [cs].

[48] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. (2018). https://doi.org/10.48550/ARXIV.1804.09541 Publisher: arXiv Version Number: 1.

[49] Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 256–267. https://www.aclweb.org/anthology/2020.emnlp-main.18

[50] Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior Knowledge Integration for Neural Machine Translation using Posterior Regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1514–1523. https://doi.org/10.18653/v1/P17-1139

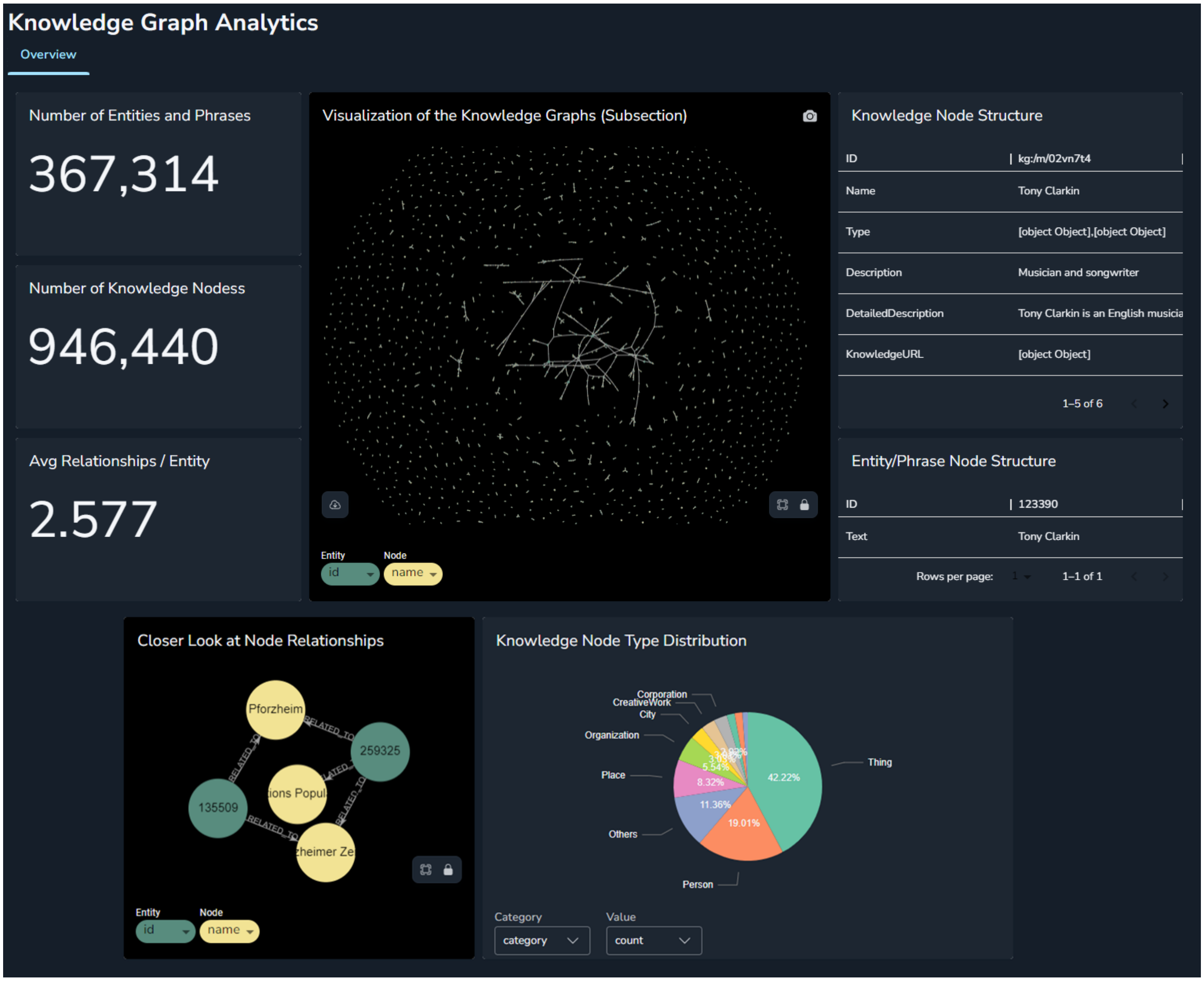
[51] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. http://arxiv.org/abs/2303.10512 arXiv:2303.10512 [cs].

[52] Yuan Zhang, Jason Baldridge, and Luheng He. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1298–1308. https://doi.org/10.18653/v1/N19-1131

[53] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. http://arxiv.org/abs/2303.18223 arXiv:2303.18223 [cs].

[54] Jianing Zhou and Suma Bhat. 2021. Paraphrase Generation: A Survey of the State of the Art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5075–5086. https://doi.org/10.18653/v1/2021.emnlp-main.414

A KNOWLEDGE BASE DASHBOARD



Knowledge Graph Analytics

Overview

Number of Entities and Phrases

367,314

Number of Knowledge Nodess

946,440

Avg Relationships / Entity

2.577

Visualization of the Knowledge Graphs (Subsection)

Entity: id

Node: name

Knowledge Node Structure

ID	kg:/m/02vn7t4
Name	Tony Clarkin
Type	[object Object],[object Object]
Description	Musician and songwriter
DetailedDescription	Tony Clarkin is an English musicia
KnowledgeURL	[object Object]

1–5 of 6

Entity/Phrase Node Structure

ID	123390
Text	Tony Clarkin

Rows per page: 1 1–1 of 1

Closer Look at Node Relationships

Entity: id

Node: name

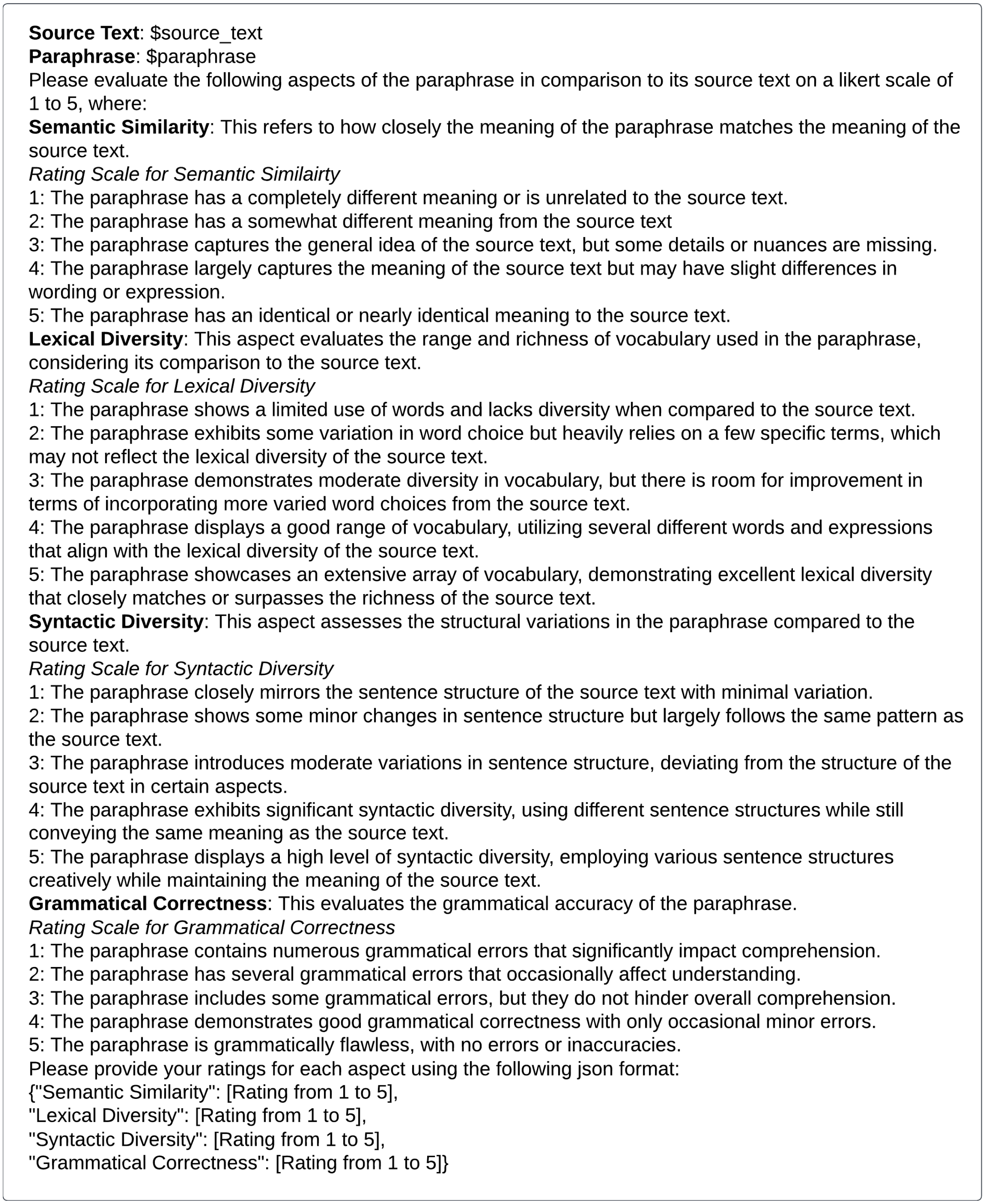
Knowledge Node Type Distribution

Category: category

Value: count

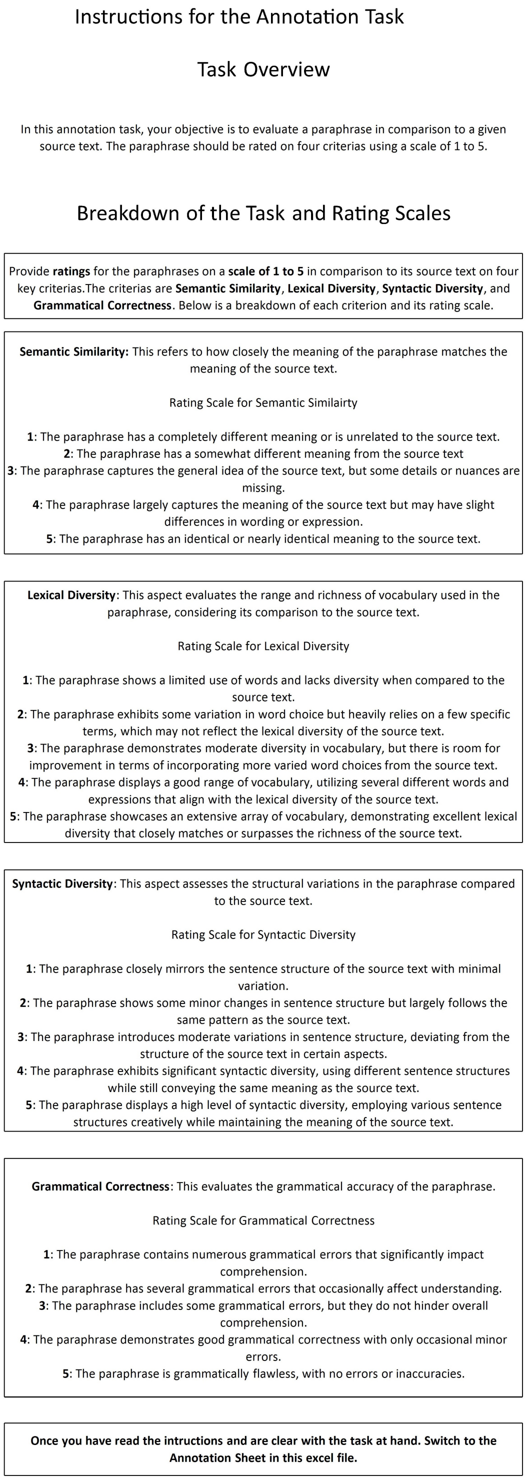
Figure 2: NeoDash Visualization of the Knowledge Base and Atrributes for G-RAG Integration

B QUALITATIVE EVALUATION INSTRUCTIONS



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]}

Figure 3: Prompt used during LLM Evaluation Process



Instructions for the Annotation Task

Task Overview

In this annotation task, your objective is to evaluate a paraphrase in comparison to a given source text. The paraphrase should be rated on four criterias using a scale of 1 to 5.

Breakdown of the Task and Rating Scales

Provide **ratings** for the paraphrases on a **scale of 1 to 5** in comparison to its source text on four key criterias.The criterias are **Semantic Similarity**, **Lexical Diversity**, **Syntactic Diversity**, and **Grammatical Correctness**. Below is a breakdown of each criterion and its rating scale.

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.

Once you have read the intructions and are clear with the task at hand. Switch to the Annotation Sheet in this excel file.

Figure 4: Instruction Sheet given to Human Evaluators