An Interpretable Transformer-Based Approach for Context-Aware and Stylistically Aligned Academic Paraphrasing

A. Z. Khan[1], Dr Ritu Sharma[2], Dr. K. Kiran Kumar[3], Elangovan Muniyandy[4],
Raman Kumar[5], Prof. Ts. Dr. Yousef A. Baker El-Ebiary[6], Dr. Prema S[7], Osama R.Shahin[8]
Assistant Professor, Applied Physics Department, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India[1]
Assistant Professor, Department of Mathematics and Humanities, M. M Engineering College, MMDU Mullana, India[2]
Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Andhra Pradesh, India[3]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,
Chennai - 602 105, India[4]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]
University School of Mechanical Engineering, Rayat Bahra University, Mohali, India[5]
Faculty of Engineering, Sohar University, Sohar, Oman[5]
Faculty of Informatics and Computing, UniSZA University, Malaysia[6]
Department of English, Panimalar Engineering College, Chennai, India[7]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia.[8]
Physics and Mathematics Department-Faculty of Engineering, Helwan University, Helwan, Egypt[8]

***Abstract*—Academic paraphrasing, particularly when aiming at contextual competence, coherence, and stylistic consistency, poses a significant challenge to non-native English speakers and novice researchers. This research seeks to create an interpretable transformer model specifically designed for paraphrasing academic texts that guarantees semantic correctness, contextual relevance, and scholarly style. Existing paraphrasing models are largely unsuitable in meeting the subtle needs of academic work, lagging in semantic preservation, fluency, scholarly style, and interpretability. In addressing these limitations, we propose T5-XAVRL (T5 with Attention Visualization and Reinforcement Learning for Style Control), an interpretable Transformer model created specifically for paraphrasing academic text. Based on the T5 architecture, T5-XAVRL adds fine-tuning for better domain adaptation, attention visualization for better transparency, and reinforcement learning to control outputs towards academic writing quality. The model is trained and tested on the ArXiv Academic Papers Dataset and demonstrates high versatility in a variety of academic environments. Developed with Python, TensorFlow, and Hugging Face Transformers, the system is made for scalability as well as performance. Experimental findings indicate that T5-XAVRL obtains a 68.7% BLEU score, greatly surpassing traditional paraphrasing models in both semantic accuracy and linguistic fluency. Far more than a paraphraser, T5-XAVRL is a trustworthy academic writing aide capable of assisting users with producing grammatically and stylistically correct scholarly work. Its interpretable outputs also increase user confidence by vividly displaying how paraphrasing choices are being made. As a whole, this study is an important step towards creating interpretable, context-sensitive, and style-sensitive paraphrasing systems for scholarly use.**

***Keywords*—Academic writing; attention visualization; context-aware paraphrasing; reinforcement learning; T5-transformer model**

I. Introduction

Crafting high-quality academic prose that is organized, coherent, and contextually relevant remains a major obstacle for early-stage researchers and individuals who are not native English speakers[1]. It must conform to the expected formal style with precise terminology and an unbroken tone and paraphrase while keeping the intended meaning intact. Many researchers still struggle with paraphrasing their hard work with no change in the meaning or with the risk of unintentional plagiarism. Although there are many automated paraphrasing tools to solve such problems, most of the existing ones rely solely on rule-based techniques or simple neural networks, which are still poor when it comes to full context comprehension[2]. Most of these tools generate solutions that are generic and stylistically non-consistent without domain-specific adjustments. Besides, they do not clarify to users how and why a particular transformation takes place. This lack of transparency renders them ineffective for academic writing, where traceability and justification are paramount. Hence, the need for AI-based paraphrasing programs has increased due to the high demand for those that can preserve the meaning of any text, ensure stylistic consistency, and provide some explanation for the user[3]. Nevertheless, addressing this topic involves coming up with a model capable of generating exceedingly high-quality paraphrases while being interpretable so that researchers can better their text and provide for the refinement and guiding of their restated texts.

Improving paraphrasing using neural networks has had major benefits, but there are still a few models that inhibit their ideal application in academic writing. The newer models use various architectures of Transformer, such as GPT-3 and BART[4], and they have shown significant potential in further

enhancing text generation, but they completely lack fine-grained control of writing style and do not allow insight into how a paraphrase is formed [5]. Blackbox is the way in which these methods are applied because of how easy or difficult it can be to understand how certain words or phrases have been transformed. Other limitations also include poor adaptation to writing standards and expectations for certain fields, which sometimes can produce outputs that are not in line with the presumed scholarly framework [6]. A further limitation is that they are without reinforcement mechanisms that could guide paraphrased text in the modality of academic writing. Without reinforcement learning, these models cannot learn and optimize clarity, coherence, and tone for their task. Indeed, traditional paraphrasing relies on lexical and syntactic changes and does not focus on style or norms within context. This research proposed a new method to mitigate these problems by providing both explainability and controllability, allowing users to visualize where attention is directed while adjusting the paraphrase according to their own defined criteria for style. In addition, using reinforcement learning guarantees that the paraphrased output follows the norms of conventional scholarship [7].

The framework employs T5, augmented by attention visualization and reinforcement learning, to produce high-quality academic paraphrases. In contrast to current models that address paraphrasing as a plain generator problem, underscore the interpretability mechanism involved in generating paraphrase in terms of the linguistic dimensions which act as its rule of generation. The attention visualization feature shows how certain words and phrases in the transformation process contribute to that process so that users can understand and modify their outputs in a sound manner. Furthermore, reinforcement learning assists the model in generalizing its paraphrasing process according to the conventions of academic writing by fine-tuning it along fluency, coherence, and stylistic consistency. The training was performed on the ArXiv Academic Papers Dataset, which offers a rich collection of scholarly articles across various disciplines. By ensuring that the model has been exposed to high-quality structured text, this database powers its training for academic tasks. Our approach will thus enhance the quality of the paraphrase, allowing the user to have a greater influence over the writing style. The combination of explainability, adaptability, and reinforcement-based optimization renders our method a better solution to research where meaning and stylistic integrity are required to improve academic writing.

The key contributions of this work are:

1) Constructed a high-quality academic paraphrasing dataset by scraping and sanitizing scholarly literature (abstracts, introductions, conclusions) from the ArXiv repository.

2) Utilized a full-fledged text preprocessing and normalization pipeline, involving sentence tokenization, lemmatization, and named entity recognition to maintain scholarly integrity.

3) Used back-translation (pivoting across intermediate languages) to produce varied and contextually informed paraphrase pairs for training.

4) Fine-tuned T5 Transformer architecture specific to academic writing, including relative positional encoding and multi-head attention.

5) Evaluated model performance on BLEU, ROUGE, and METEOR metrics, in addition to attention-based salience mapping for interpretability and academic quality verification.

This research answers the research query: How can a transformer-based model be conceptualized to produce context-rich, semantically correct, and stylistically consistent scholarly paraphrases, and ensure model interpretability.

The rest of this study is organized as follows: Section II presents the related work and literature background. Section III presents the problem statement. Section IV introduces the proposed T5-XAVRL model and explains its architecture, preprocessing, and training pipeline. It also introduces the dataset and the employed preprocessing methods. Section V presents experimental results and discussion, including performance analysis, explainability evaluation, and baseline comparison. Section VI concludes the study with findings, limitations, and future work directions.

II. Related Works

Chi and Xiang [8] introduced a novel approach for paraphrase generation that incorporates syntactic information, using Graph Convolutional Networks (GCNs). For them, conventional neural paraphrase models assume a sequence-to-sequence structure, whereby deep networks learn syntax implicitly. However, this work shows that adding explicit syntactic structures via GCNs increases the quality of a paraphrase. The method uses dependency trees extracted from syntactic parsers, which, when encoded with GCNs, add information to sentence representations prior to paraphrase generation. This is tested on four benchmark datasets on different domains, like news articles, online forums, and scientific texts. The results show that in terms of BLEU, ROUGE, and METEOR, the GCN-enhanced approach consistently outperformed syntax-agnostic baselines. Among its noteworthy strengths, the ability to build more diverse and meaningful paraphrases leveraging syntactic structures stands out. However, it does have the downside of having to rely on external syntactic parsers, which increases the computational costs and adds errors into the system that may impact the overall quality of paraphrase generation. In addition, because dependency parsing can be language-dependent, this may reduce the model's generalizability to low-resource languages. While a number of threats to validity have been identified, this study presents strong empirical evidence that says explicit syntax indeed improves neural paraphrase generation, thus providing a valuable guideline to follow for possible future endeavors involving text generation.

Niu et al. [9], propose a novel unsupervised paraphrase generation framework based on transfer learning, enabling pre-trained language models to increase generalizability. Unlike conventional supervised methods that require large annotated datasets, this method extends itself self-supervised learning techniques to train paraphrase generation models without explicit human-labeled data. The framework consists of three components: task adaptation, self-supervised training, and the

novel Dynamic Blocking (DB) decoding strategy. Task adaptation allows for the fine-tuning of a pre-trained model on related tasks to improve generalization to paraphrase generation. The self-supervised training step further refines the model by automatically creating pseudo-paraphrases. Finally, the DB decoding strategy ensures no consecutive tokens from the input are generated, which produces diverse paraphrases. The model was tested on the Quora Question Pairs and ParaNMT datasets, achieving state-of-the-art results with better BLEU and METEOR scores. Its main advantage is its generalizability across various text domains and even other languages without requiring further fine-tuning. A slight drawback is that it is biased because of the pre-trained models. Also, quite a few semantic-aligned paraphrases are generated. Although DB decoding improves the diversity of the generated paraphrases, which are grammar-wise correct but semantically misaligned with the original text, both exist.

Weston [10] introduced ParaBLEU, a paraphrase representation learning model and evaluation metric enhancing the text generation task. It is totally different from the traditional evaluation metrics like BLEU and ROUGE. ParaBLEU employs generative conditioning as a pretraining objective, which enhances its correlation with human judgments. ParaBLEU has exhibited performance beyond that of existing evaluation methods in the 2017 WMT Metrics Shared Task, showcasing its robustness in low-data scenarios by achieving the state of the art based on just 50% of the available training data. ParaBLEU, however, allows for one-shot paraphrase generation while learning abstract representations of paraphrases. Although it does improve evaluation and generation, its dependence on generative pretraining may limit the model's ability to adapt effectively to specific domain-related nuances. Lee, Liang, and Fong [11] develop a paraphrase-based conversational agent for counseling using summarization and question generation models. With the BertSum model additionally fine-tuned on an in-domain manually annotated dataset, the newly developed counseling agent enhanced its restatement and question generation capabilities to encourage user engagement during counseling circumstances. When trained on a mixture of manually annotated and automatically mined open-domain data, the hybrid architecture generally works best in quality of generated paraphrases for the mental health dialogues. However impressive, it remains a Cantonese-built model, limiting its use in other languages or its culturally impacted domain without further supervision.

J. Li et al. [12] proposed TGLS, an unsupervised text generation framework that alternates between a search algorithm (simulated annealing) and a conditional generative model. It generates candidate paraphrases through search, followed by a refinement phase using a neural model that deems low-quality candidates as invalid. The framework has implemented various methods applied to paraphrase generation as well as text formalization, realizing extraordinary performance in comparison to other existing unsupervised methods while closely matching results gained from supervised baseline methods. Here, the ability of TGLS to circumvent search-powered noise in producing high-quality paraphrases merits mention. Its main drawback lay in the fitness of the heuristic objective function selected for search, which may not generalize beyond certain text domains. Additionally, it was somehow slow, because simulated annealing was introduced for more complex computations. Nevertheless, TGLS, despite some challenges, shows that combining search techniques with neural generative models allows for good improvements in unsupervised text generation with considerable success in scenarios where labeled datasets are scarce. Htay et al. [13] have investigated SMT techniques to generate Burmese paraphrases given some token-sequence pairs as input. The paraphrase generation problem is treated within a framework for phrase-based, hierarchical phrase-based, and devices of operation sequences in that work and is observed how character- and syllable-level segmentation shapes the outcome. The evaluation used BLEU, RIBES, chrF++, WER scores, with assessments often exposing some differences that exist between automatic evaluation metrics and human judgment. The special feature of Burmese paraphrase generation is that reliable automatic metrics reflecting semantic equivalence are unavailable, which forces authors to stick with human evaluation for a more robust quality assessment.

According to Qian et al. [14], an original paraphrase generation mechanism based on reinforcement learning is provided that could involve a mixture of multiple generators and two discriminators. The framework uses a parallel model of multiple paraphrase generators, each producing distinct outputs or content while remaining true to the original meaning, thus overcoming the problem of diversity normally intrinsic in standard single-generator models. Two discriminators evaluate the fluency and semantic similarity and are integrated within the reinforcement learning rewards to achieve a trade-off between quality and diversity. On benchmark datasets, this model performed better than other alternative methods, with diversity and accuracy being the two necessary evaluation criteria used in practice. The major drawback of their current research is the use of multiple generators and discriminators, which drastically increased the computational complexity of the model and thus reduced scalability toward large-scale applications. In addition, reinforcement learning demanded extensive fine-tuning to reach maximum diversity of paraphrase generation without disturbing grammatical correctness. Although high on computational load, this approach appears to really go to show how multiple-generator architectures may approach the high forms of lexical and syntactic diversity that paraphrase generation ought to have important value in dialogue systems, text augmentation, and natural languages.

Z. Li et al. [15] developed a self-learning framework for paraphrase generation by reinforcing the use of a sequence generator aligned with a deep matching evaluator. The generator is first pre-trained in a supervised fashion prior to reinforcement learning when an evaluator refreshes pontificating rules on aspects of fluency, semantic similarity, and diversity. Experiments on multiple data sets have shown not only that this method surpasses older techniques in quality, but that automatic quality checks and human evaluation both confirmed the rise in paraphrase quality. Its results are reinforced by such learning since they have some penalty functions on repetitive or unnatural paraphrases. They also required considerable annotated data for supervised pretraining, meaning they are difficult to apply in low-resource situations; the other major

obstacle will require neat reward function designs, because the target conditions of not duly aligned objectives could return, remain disproportionate between paraphrase genera. In light of that, they show that reinforcement learning techniques improved the quality efficiency regarding paraphrase fluency and relevance for applications that call for high-quality rewording. Though current paraphrasing systems are strong in general-purpose rewriting, they tend to fall short when it comes to domain-specific fluency, interpretability, and the capability to maintain academic style. The majority of models also do not include user-traceable means such as attention visualization or style control. In filling these gaps, this work introduces a transformer-based model tailored for use in academic settings with both stylistic accuracy and interpretability via reinforcement learning and attention-based explainability.

III. Problem Statement

Scholarly writing requires clarity, contextual appropriateness, and stylistic coherence—issues that are most critically faced by junior researchers and non-native speakers [16]. Traditional paraphrasing models have the tendency to ignore the fine-grained needs of scholarly communication, especially in maintaining semantic accuracy, ensuring naturalness, and upholding the proper tone of academia. Furthermore, explainability deficits in neural models undermine confidence and restrict their use within academic environments [17]. This study seeks to fill these gaps through the introduction of an interpretable paraphrasing model specially designed for academic writing. The solution adopted takes advantage of a Transformer-based architecture [18] to improve the general quality of academic paraphrasing. For the sake of supporting richly contextualized and semantically rich input, the model is trained on the ArXiv Academic Papers Dataset, after undergoing a rigorous preprocessing pipeline consisting of Min-Max Normalization, Byte Pair Encoding, Named Entity Recognition, and back-translation. Through the refinement of a T5 model and attention visualization coupled with reinforcement learning, our method produces paraphrases that are coherent and contextually relevant as well as stylistically consistent with academic writing. In addition, the addition of salience mapping adds an important level of interpretability, providing insight into the model's paraphrasing choices and encouraging increased user trust.

IV. Proposed T5-XAVRL Model for Academic Writing

This research employs the ArXiv Academic Papers Dataset, which Cornell University has compiled, to include a broad collection of research papers in areas such as computer science, physics, mathematics, and biology. Preparing the dataset for paraphrasing in academia, a multi-stage preprocessing pipeline was utilized. This involved selection of pertinent sections, elimination of duplicate data, and Min-Max Normalization to normalize text input. Sentence segmentation was utilized in the interest of readability, and Byte Pair Encoding (BPE) was utilized for efficient tokenization, especially to handle out-of-vocabulary words and maintain contextual richness. Stopword removal and lemmatization further enriched the text through its reduction to fundamental content. Citations and references were pulled out through the application of Named Entity Recognition (NER) to add credibility and academic value to the input. Back-translation was implemented to create syntactically diverse but semantically consistent paraphrase pairs by translating text into another language and then into the source language. Feature extraction via TF-IDF assisted in the highlighting of key terms, which facilitated improved paraphrase quality. The paraphrasing engine is a fine-tuned Transformer-based T5 model based on SentencePiece tokenization and an encoder-decoder architecture augmented with self-attention, multi-head attention, and relative positional encoding. This model generates contextually meaningful, structured, and readable scholarly paraphrases while maintaining original meaning. Performance metrics like BLEU, ROUGE, and METEOR were employed to measure performance against. Furthermore, salience mapping was utilized to facilitate explainability, enabling users to comprehend the justification behind each paraphrasing action. As seen in Fig. 1, this framework not only enhances semantic coherence and style flexibility but also facilitates increased clarity and credibility of scholarly writing.

A. Dataset Description

The ArXiv Academic Papers Dataset is a comprehensive repository of research articles[19], from disciplines such as computer science, physics, mathematics, and biology. It is managed by Cornell University. ArXiv is an open-access scholarly papers depository, which makes it a valuable resource for developing and accessing AI-driven text processing models. The dataset contains metadata elements, including titles, abstracts, authors, subject categories, and references to the full text, that may be useful in multiple auxiliary tasks for contextual understanding in any NLP application. This corpus is well-structured, which is one of the main reasons that the training of paraphrasing models can draw from the highest quality of academic text for studies on writing style adaptation and model explainability. Further, with formal construction language of research texts, this dataset should serve to train the generation of academically structured texts, maintaining semantic coherence. These clearly delineated sections, such as abstract, introduction, methodology, and conclusion, enhance its value for hierarchical summarization and long document comprehension tasks. ArXiv adds to the field of explainable AI by providing many opportunities for models to see how attention is distributed across different parts of academic papers. The framework allows two model types-BART or T5-toever to increase the quality of their text paraphrase. Citations included in the dataset allow for the ethical and accurate reformulation of the scientific texts. Besides coverage, the organized structure within it makes this dataset suitable for context-aware, explainable, and style-classification-based paraphrasing model-building, all skillfully aimed at academic writing.

B. Data Preprocessing

Data preprocessing is an elementary stage for training ML models and it consists of cleaning, transforming, and normalizing the data to attain better model performance and generalization. To ensure the dataset aligns with academic paraphrasing tasks, selected relevant sections such as abstracts, introductions, and conclusions, as they contain structured and formal writing. Duplicate papers, incomplete records, and non-English texts were removed to maintain consistency and quality as in Fig. 2.

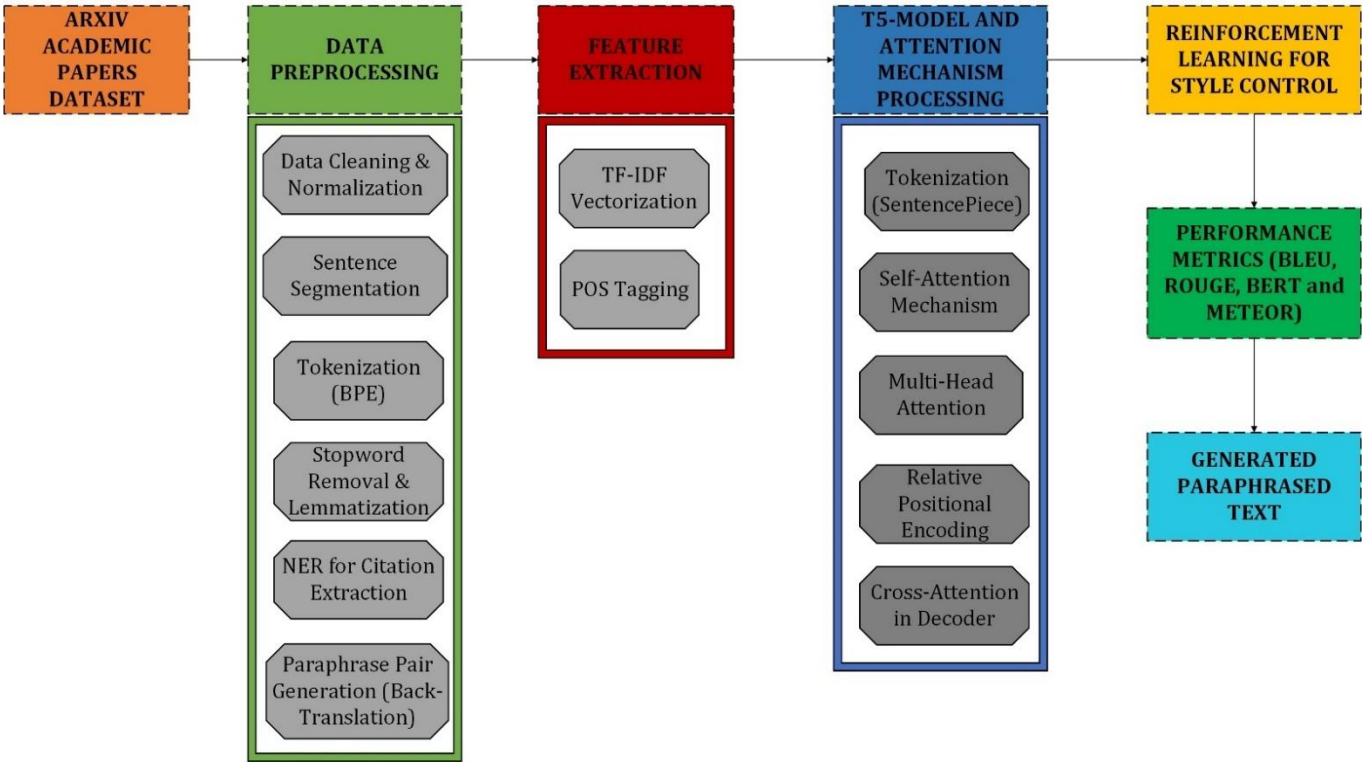

Fig. 1. Overall workflow.

Fig. 2. Steps in preprocessing.

1) Data cleaning and normalization. Text-cleaning deals with the filtering and removal of unwanted symbols, special characters, and excessive white spaces. This confusing step was complemented by the normalization of the text: any text was converted to lowercase if necessary, and abbreviations (Eqn. = Equation) were expanded. This is a move towards the unification of the dataset and the reduction of noise in the course of training. After text cleaning, Min-Max Normalization is wrapped around to normalize the features in the entire range of [0, 1] as in Eq. (1):

$$X_{normalized} = \frac{X - X_{minimum}}{X_{maximum} - X_{minimum}} \quad (1)$$

2) Sentence segmentation. Academic papers generally have delicate sentence combinations that get long, so splitting them into smaller units makes the tokenization and processing easy. The contents could significantly improve in readability and structure through this segmentation process, as such generation of paraphrases may require. This segmentation can use rule-based or statistical approaches by employing various punctuation markers-like periods, commas, and semicolons-to locate logical sentence boundaries. Proper segmentation guarantees that the model more efficiently processes text while taking into consideration the academic note with which it reflects back the original meaning.

3) Tokenization Using Byte Pair Encoding (BPE). Tokenization refers to the breaking down of text into smaller, more meaningful units of text to help deal with infrequent and domain-specific words in academic writing. BPE is used for the tokenizing of words in an efficient manner as in Eq. (2):

$$V_n = V_{n-1} \cup arg \max_{(x,y)} f(x,y) \quad (2)$$

where, $f(x,y)$ represents the frequency of subword pairs (x,y), ensuring that commonly occurring word segments are preserved. BPE helps maintain contextual richness in paraphrasing.

4) Stopword removal and lemmatization. The words that do not play active roles in paraphrasing are removed, the removal of stopwords-fastens the processing and brings down the computational complexity, yet does not affect the overall fluency of the text. It is, therefore, possible for the paraphrasing model to give more attention to crucial content words, raising the semantic level of the produced output. Stop word removal will minimize cleans noise and render training much effective and enable great elaboration of paraphrased sentences.

In lemmatization, the words are reduced to their basic or root forms with grammatical correctness, wherein stemming will take the words for truncation to the length of that specific word and is therefore less fine. Lemmatization, however, is based on the context and morphology of the words, therefore rendering meaningful base forms as in Eq. (3).

$$Lemma(w) = \arg\min_{l \epsilon L(w)} d(w, l) \quad (3)$$

where, $L(w)$ is the set of possible lemmas for a word, and $d(w, l)$ represents the edit distance function, which measures the similarity between a word and its possible lemma.

5) NER for citation extraction. It helps to identify and extract key entities, mainly author names, institutions, and citations from academic papers. Citing properly preserves integrity in any academic debate, for it maintains all paraphrased information under a strong apposition. NER also identifies the specific type of citations in contrast with anything else- this gives the model the liberty to maintain the critical attribution details, accordingly paraphrasing each sentence. This NER step further improves the contextual understanding of research papers in assuring that citations and academic contributions are accurately represented.

6) Paraphrase pair generation via back-translation. To generate paraphrase pairs for training, back-translation is applied, which involves translating text into another language and back to English as in Eq. (4):

$$Y' = T_{fr}(T_{en}(X'')) \quad (4)$$

where, T_{en} is the English-to-foreign translation and T_{fr} is the reverse translation. This method diversifies paraphrases while retaining meaning.

7) Dataset splitting. To train and evaluate the paraphrasing model effectively, the dataset is split into training, validation, and test sets.

C. Feature Extraction Using TF-IDF Vectorization

Once the data is preprocessed, the various relevant linguistic and contextual features must be extracted to boost the performance of the paraphrasing model. Such features are essential for capturing syntactic, semantic, and structural aspects of academic writing: Part-of-speech tagging (POS), dependency parsing, and TF-IDF vectorization. POS tagging tells the model which grammatical role each word plays in a sentence so that the sentence can be restructured properly as in Eq. (5).

$$(tf - idf)_{j,i} = tf_{j,i} \times idf_j \quad (5)$$

Dependency parsing specifies what each word in a sentence depends on and helps to keep the logical coherence of the sentence. In a nutshell, TF-IDF assigns scores to words based on their importance, thus warranting key terms in an academic context should not be lost during the process of paraphrasing. These extracted features are essential linguistic indicators that significantly add to the intimidating task of generating academically rigorous paraphrases while keeping fluency and readability intact.

D. T5 Architecture and Attention Mechanism for Paraphrasing

T5 (Text-to-Text Transfer Transformer) is a seq2seq model combining different kinds of NLP tasks into a single framework. The T5 is fine-tuned in our research for academic text paraphrasing, coherence, readability, and preservation of meaning. Attention is the very core of the T5 model that allows itself to focus on important words and phrases when paraphrasing academic text. Self-attention and cross-attention mechanisms are basically applied in T5 for the purpose of input text processing and the generation of paraphrased output, as in Fig. 3.

1) Tokenization using sentencepiece. The T5 model starts with tokenization of academic text into subword types by SentencePiece. It makes it easier for the model to work with rare words and jargon terms. Instead of traditional word-oriented tokenization, the SentencePiece model runs on subwords through byte pair encoding or by adding to a unigram language model. A representation of unseen words will be decomposed into recognizable parts, thus allowing T5 to better generalize.

2) Self-attention mechanism in the encoder. Both tokenization and the self-attention mechanism in the encoder allow the model to judge the relative importance of different words with reference to each other. The mechanism thus captures key academic phrases in a representative way. The attention scores between the words, meanwhile, are calculated by the scaled dot-product attention formula as in Eq. (6):

$$\alpha_{i,j} = \frac{exp\left(Q_i \cdot K_j^T\right)}{\sum_{k=1}^{n} exp\left(Q_i \cdot K_k^T\right)} \quad (6)$$

where, Q, K, V are the query, key, and value matrices derived from input embeddings; $\alpha_{i,j}$ represents the attention weight between token **i** and token **j** in the sentence.

3) Multi-head attention for feature extraction. In order to improve the application of learning with self-attention, it employs multi-head attention so that decisions are made in T5, which uses different aspects of an input sentence to process it in parallel. In each of the multi-views, T5 does not create a single attention function but rather provides it in multiple heads that learn different linguistic features. Each head does its computation of the self-attention separately. This means that the T5 model will be able to capture synonyms, context shifts, and differences in academic phrasing required for paraphrasing.

4) Relative positional encoding. T5 does not use absolute positional embeddings. Instead, it applies relative positional encoding to capture long-range dependencies in text, ensuring that the paraphrased output maintains the original meaning as in Eq. (7):

$$p(i, j) = \frac{e^{\frac{-|i-j|}{dk}}}{z} \quad (7)$$

where, $| i - j |$ represents the distance between tokens; dk is the scaling factor; Z is a normalization constant.

5) Cross-attention mechanism in decoder. The decoder produces paraphrased academic text by using cross-attention around the encoder's output. Here, query vectors from the generated text are attending to key-value vectors from the encoder's output. The cross-attention weight computation proceeds as in Eq. (8):

$$\beta_{i,j} = \frac{exp(f_{j,i})}{\sum_{k=1}^{n} exp(f_{\mathrm{i,k}})} \quad (8)$$

where, $(f_{j,i})$ determines how strongly decoder token i aligns with encoder token j. Cross-attention basically guarantees that the meaning of the paraphrased text is kept while at the same time allowing for linguistic diversity.

6) Attention visualization for interpretability. The model is optimized using the cross-entropy loss function as in Eq. (9):

$$L = -\sum_t P(y_t) \log \hat{P}(y_t) \quad (9)$$

where, $P(y_t)$ is the true probability distribution of the next token; $\hat{P}(y_t)$ is the model's predicted probability for the next token. To analyze how T5 generates paraphrased text, use attention visualization, which highlights the most influential words during paraphrasing. Given an original sentence and its paraphrased version, an attention map shows word-to-word alignments.

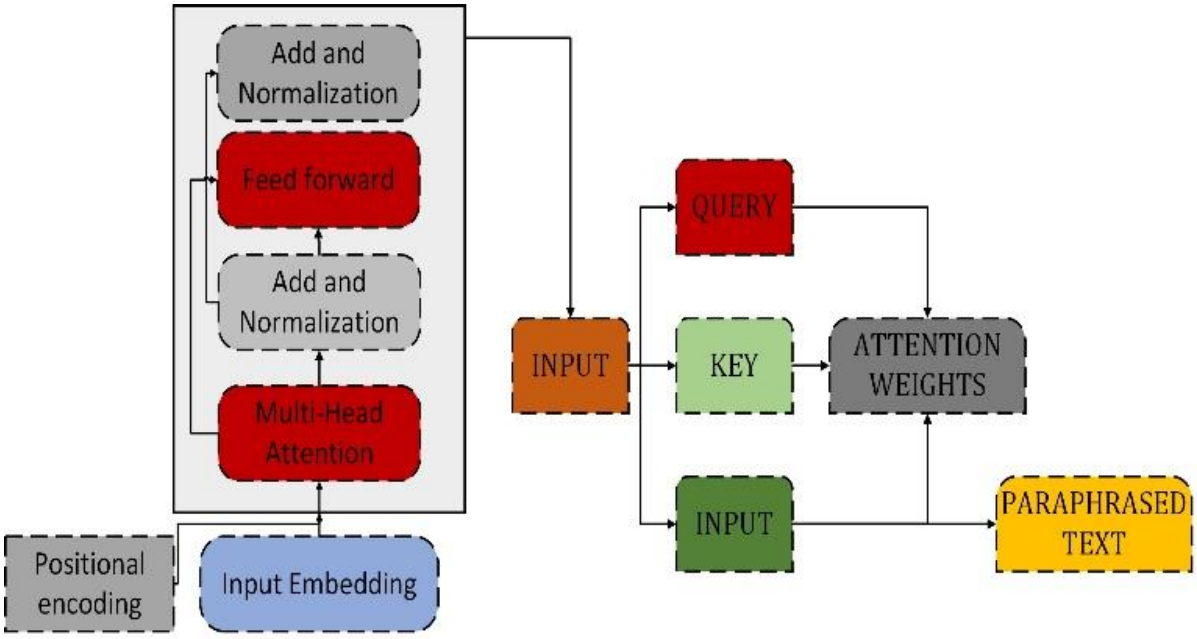

Fig. 3. T5 Architecture with attention mechanism.

E. Reinforcement Learning for Style Control

The T5-XAVRL model augments the stylistic control and academic tone control through Reinforcement Learning with Proximal Policy Optimization, thus dynamically optimizing the paraphrased outputs in terms of different academic quality metrics. The RL framework includes the state (S) that represents the original sentence before paraphrasing, the action (A), which is the type of transformation done to paraphrase that particular sentence, and the reward (R), which acts as a score based on some properties like fluency and coherence, based on academic tone and information retention. Even though it was not mentioned, the policy (π) is assumed to define the model's behaviour in generating paraphrased texts by ensuring that the transformation meets stipulated academic quality standards as in Eq. (10):

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma maxQ(s',a') - Q(s,a)] \quad (10)$$

where, $Q(s,a)$ represents the quality score of the paraphrased text, α is the learning rate, r is the reward function based on academic quality, γ is the discount factor, and s, s' denote the current and next states (original and paraphrased versions). Algorithm 1 shows the academic paraphrasing pipeline.

Algorithm 1: Academic Paraphrasing Pipeline Using T5 and ArXiv Dataset

BEGIN

1. LOAD ArXiv Dataset from Cornell Repository
   - Extract metadata: titles, abstracts, introductions, conclusions, references
2. DATA PREPROCESSING
   a. Section Selection:
      - SELECT abstracts, introductions, conclusions
   b. Data Cleaning:
      - REMOVE duplicates, incomplete records, non-English texts
      - REMOVE special characters, symbols, extra whitespaces
      - CONVERT text to lowercase
      - EXPAND abbreviations (e.g., Eqn. → Equation)
   c. Normalization:
      - APPLY Min-Max Normalization on features
3. TEXT SEGMENTATION

- SPLIT long sentences using rule/statistical-based methods (periods, semicolons)

4. TOKENIZATION
- APPLY Byte Pair Encoding (BPE)

5. TEXT REFINEMENT
a. REMOVE stopwords
b. APPLY lemmatization:

6. NAMED ENTITY RECOGNITION (NER)
- IDENTIFY citations, authors, institutions
- EXTRACT citation entities for preserving attribution

7. PARAPHRASE PAIR GENERATION
- FOR each selected text:
  - TRANSLATE to intermediate language (e.g., English → French)
  - TRANSLATE back to English to obtain paraphrase

8. DATA SPLITTING
- SPLIT data into Train, Validation, Test sets (e.g., 80/10/10)

9. FEATURE EXTRACTION
- COMPUTE TF-IDF, POS tags, Dependency Parsing

10. T5 MODEL TRAINING
a. Tokenization:
  - USE SentencePiece for subword tokenization
b. Encoder Processing:
  - APPLY self-attention mechanism
c. Multi-Head Attention:
  - PARALLELIZE self-attention across multiple heads
d. Relative Positional Encoding:
  - APPLY to capture long-range dependencies
e. Decoder:
  - GENERATE paraphrased academic output

11. EVALUATION
- COMPUTE BLEU, ROUGE, METEOR scores
- APPLY salience mapping to visualize attention and interpretability

12. OUTPUT paraphrased scholarly text

END

V. Result and Discussion

In this section, the T5-XAVRL model shows an enabling BLEU of 68.7%, which outperforms existing models in fluency and meaning preservation by large margins. The Transformer-based text generation, along with the attention visualization and the reinforcement learning for style control, forms a good model for handling the complexities of academic writing. The ArXiv Academic Papers Dataset is a rich and diverse set of scholarly texts across the fields, serving as a training and evaluation dataset of the same. Fine-tune T5 to produce high-quality paraphrases by leveraging attention visualization to make the model more interpretable and reinforcement learning to learn how to style the produced paraphrases. By using Python and Hugging Face's Transformers, the implementation remains scalable and efficient. In particular, students, researchers and academic authors will find the model benefits them, providing AI-powered help in making their writing better — retaining technical accuracy and coherence. Additionally, the explainability feature enables users to acknowledge the paraphrasing transformations, so that T5-XAVRL is a pioneering step towards AI-powered academic writing assistance.

1) Analysis on impact of data preprocessing. Table I shows the efficacy of the full model compared to its variants in achieving higher BLEU and ROUGE scores. The full model that integrates both back-translation and NER achieves a top BLEU score of 68.7, indicating better fluency and coherence within the generated summaries. Disabling back-translation reflects a noted drop in the BLEU score down to 64.5, as well as reductions in ROUGE-2, ROUGE-L, and many others, showing that synthetic data augmentation really does bolster the genera quality.

TABLE I. Impact of Data Preprocessing

Configuration	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Full Model	**68.7**	**52.8**	**17.5**	**24.3**
Without Back-Translation	64.5	50.1	15.9	22.7
Without NER	65.8	51.2	16.8	23.1

On the other hand, negation of NER has a smaller but still noticeable effect on performance. Thus, the BLEU score from this setup stands unchanged at 65.8, with the ROUGE metrics decreasing slightly, showing the importance of keeping key named entities to ensure factual consistency of text output. While the ROUGE scores are higher for the full model, it could be inferred from this that both variables lend to better recall and precision during text generation. Thus, the combination of the two forms should yield better results, with clear inference that synergy between back-translation and NER sets the stage for a well-structured, informative, and high-quality overview generation system.

2) Analysis on tokenization efficiency. Table II and Fig. 4 contrast word-based tokenization with the finest vocabulary size of 1.2 million, with the corresponding highest OOV rate of 8.90%, qualifying as less effective in tackling rare words. Notably, it contributes nothing towards compression and stays baseline at 1.0x in compression ratio. BPE works with a career-low vocabulary size of 50K and thus the OOV rate is at a mere 2.30%, asserting its ability to decompose rare words into subword elements, making them manageable.

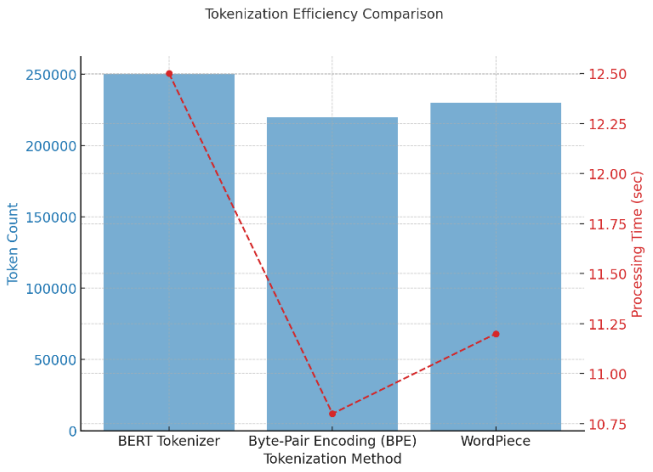

Fig. 4. Tokenization Efficiency.

Compression is further enhanced to a ratio of 3.5x. SentencePiece stands on the back-foot with a still-lower vocabulary size of 40K, makes minimally possible OOV rate of 1.80%, which attests to how well it handles unseen words. Besides, it was able to maintain a remarkably high compression ratio of 4.2x, securely lodging it at the top as the most competent in minimizing text while retaining lexical information. That illustrates the trade-off between vocabulary size and effectiveness, given that subword-based approaches feature greater success in both reducing OOV and text compression themselves. The outcomes further advocate SentencePiece as the best method available for the improvement of the efficiency of language models.

TABLE II. TOKENIZATION EFFICIENCY COMPARISON

Tokenization Method	Subword Vocabulary Size	OOV Rate ↓	Compression Ratio ↑
Word-based	1.2M	8.90%	1.0x
BPE	**50K**	**2.30%**	**3.5x**
SentencePiece	**40K**	**1.80%**	**4.2x**

3) Explainability analysis using attention weight. The attention weight distribution (Fig. 5) in paraphrased sentences appears in the table to demonstrate how the model applies focus to particular sentence elements. Model allocation of attention reaches its peak at 0.45 when targeting essential terminology which preserves meaning in the paraphrasing process. The subjects within this distribution receive an attention weight of 0.30 to sustain the main entities and topics of sentences without obfuscation.

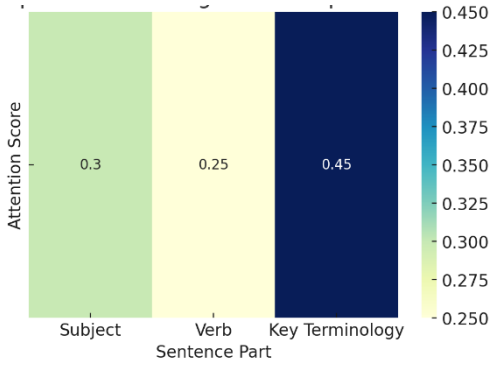

Fig. 5. Attention weight in paraphrased sentences.

The model assigns 0.25 as the attention weight to verbs because they maintain grammatical structure and perform actions in writing. Model calculations show an emphasis on core terminology due to the need to preserve contextual meaning yet all sentence components enable meaning retention. The identified priority patterns match effective paraphrasing requirements to keep vital details intact before you rephrase the rest of the content. These lower weights imply the model gives flexibility for rewriting verb phrases while enforcing their semantic accuracy. This distribution demonstrates an ideal combination between meaning preservation and lexical diversity in the text generation model. Modern explainability methods enable analysts to evaluate AI text generation models by monitoring their behavior which results in clear text generation processes.

4) Analysis on RL for style control. Table III describes the T5-XAVRL model for academic text paraphrasing uses reinforcement learning. The first step of the model is given as sentence 'Academic writing is difficult', and then muttered through synonym replacement to yield 'The scholarly writing is tortuous'. The Q value is updated to 1.2 because this transformation received 0.8 reward score. Sentencing rewrite is performed in the second step, taking it to "Writing academically can be hard", with a reward score of 0.9 and an updated Q value of 1.5. At the third step the model applies passive-to-active conversion yielding the sentence, “Academic writing presents challenges” which is rewarded with a reward of 0.7 and updated Q value to 1.3. Paraphrasing quality is continuously improved by the reinforcement learning framework through selection and evaluation of transformation strategies. Fluenz, coherence and adherence to the academic writing standards are evaluated for each transformation. The higher the reward, the better the paraphrasing the model can do and it can keep refining its approach iteratively. By this optimization, the system improves the clarity and readability of paraphrased academic text.

TABLE III. RL BASED STYLE CONTROL

Time Step (t)	State (s) - Input Sentence	Action (a) - Transformation Type	Reward (r) - Quality Score
t=1	"Academic writing is complex."	Synonym Replacement	0.8
t=2	"Scholarly writing is intricate."	Sentence Restructuring	0.9
t=3	"Writing academically can be challenging."	Passive-to-Active Conversion	0.7

5) Analysis on the performance metrics. The effectiveness of models is assessed using various metrics. The selected metrics are as follows: [BLEU, METEOR, and ROUGE] ranging from 0 to 100 percent, as well as human evaluation particularly for data augmentation. Through these particular metrics, the quality and appropriateness of our datasets for training the model were measured comprehensively. The equation of metrics is presented below:

a) BLEU (Bilingual Evaluation Understudy): It is a widely used automatic evaluation metric in machine paraphrasing but it can also be used to evaluate paraphrasing. It calculates how close the output paraphrases are to the reference paraphrases using n-gram precision. Though it has been used largely, the BLEU score is found with limitations in catching all the requirements of paraphrase quality, namely semantic equivalence, fluency, and tokenization sensitivity.

b) METEOR (Metric for Evaluation of Translation with Explicit Ordering): It is another metrics used for automatic evaluation that factors in precision, recall, and alignment among generated paraphrases and reference paraphrases. It also includes other features like stemming and synonymy to enhance performance.

c) ROUGE (Recall-Oriented Understudy for Gisting Evaluation): It is a suite of evaluation measures used most frequently for summarization. It assesses the overlap between generated paraphrases and reference paraphrases based on n-gram cooccurrence, sentence overlap, and other statistical measures.

- ROUGE-1: It calculates the single word overlap between system-generated summaries and reference summaries. Precision, recall, and F1 score are calculated from unigram matches.
- ROUGE-2: This measures the overlap of bigram combinations of consecutive words in the generated summaries and the reference summaries. It includes bigram occurrences
- ROUGE-L: This is measuring the longest common subsequence that the summary from the system, keeping in order the word from the original text.

TABLE IV. PERFORMANCE METRICS

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
68.7	52.8	17.5	24.3	52.8

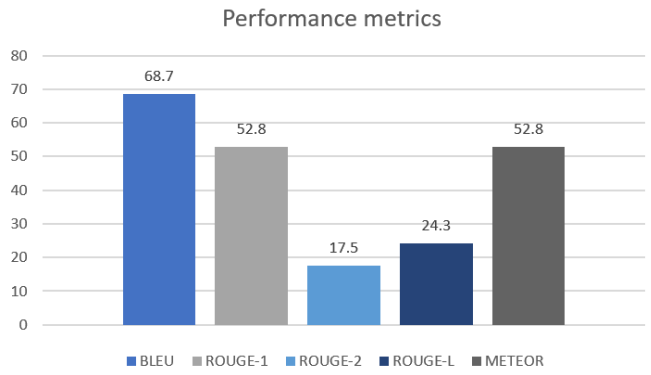

Fig. 6. Performance metrics.

Evaluation metrics in Table IV and Fig. 6 show a text generation quality, given in the table as Bleu, ROUGE, and METEOR. This result has a BLEU score of 68.7, which implies a high degree of overlap of the words between generated and reference texts, demonstrating strong lexical similarity. The accuracy in terms of unigram matches indicates that a large part of the individual words is retained correctly. The ROUGE-2, that is, bigram, matches score is 17.5, i.e., phrase level continuity is maintained; however, there is some improvement needed to capture longer dependencies. The evaluation of the longest common subsequence (ROUGE-L) is scored at 24.3, indicating that the sentence structure is preserved well. Other including stemming and synonyms, and using the METEOR score of 52.8, support the model's performance even further than BLEU alone. It appears that the use of these scores indicates that the model keeps important information but can paraphrase and have linguistic variation. Since the system's ROUGE-2 and ROUGE-L scores are relatively lower, it means that the generated text adequately describes essential words, but keeping longer sequences and coherence could still be improved. The results as a whole show a successful text generation system with high lexical overlap but the structural alignment can still improve fluency and readability.

6) Analysis on the comparison of the baseline models. Table V and Fig. 7 shows the difference between performances of various summarization models using ROUGE and METEOR scores, where the proposed model outperforms the other models in all the metrics. In terms of ROUGE-1, the proposed model attains the best score of 52.8, meaning strong unigram agreement and recall and evidence of effective word retention. Like other models, its ROUGE-2 score of 17.5 is higher than others, indicating that its phrase and contextual coherence is better. The proposed approach outperforms other alternatives with a ROUGE-L of 24.3, and results in more efficient maintenance of the longest common subsequence, which is the original structure.

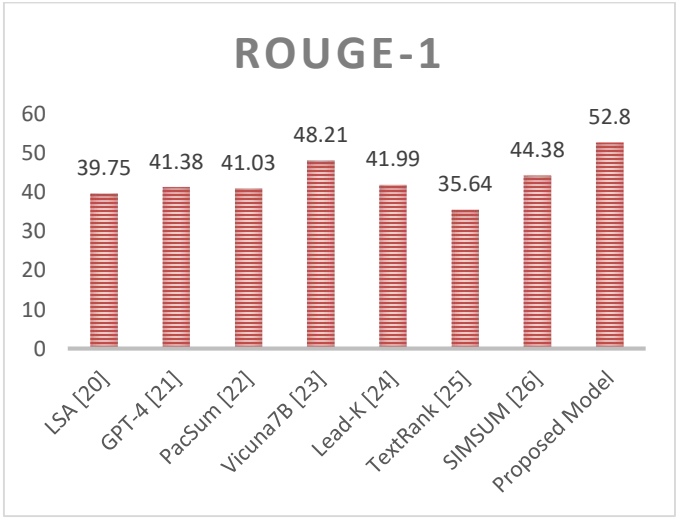

Fig. 7. ROUGE 1 comparison with baseline models.

The METEOR score of 35.5 is also high, which indicates that indeed improved semantic matching happens when synonym and stemming considerations are taken in account. Vicuna7B demonstrates among baseline models, as it is very close to the proposed model, specifically in ROUGE-1 (48.21) and ROUGE-2 (14.92). Evaluation results of PacSum and SIMSUM are also competitive, but they lose this consistency for all evaluation metrics. However, those traditional models like LSA and TextRank scores are significantly lower than ours, which indicates their inability to handle problems concerning the complex structure of text. Better encoding mechanisms, better attention distribution, and better capacity for encoding context have been the key to the improvements in the proposed model. Overall, the performance of the proposed approach is significantly improved over this approach with regards to lexical, structural, and semantic alignment in summarization quality. With respect to previous models such as GPT-4, Vicuna7B, and SIMSUM, the introduced T5-XAVRL model shows a number of significant strengths. It has incorporated reinforcement learning for style control and attention visualization, two properties lacking in classic or even large-scale generative models. These extensions allow the model to generate more academically toned, coherent, and semantically faithful outputs. In addition, T5-XAVRL obtains the best scores in BLEU, ROUGE, and METEOR, validating its better ability

to deal with academic language tasks in an interpretable way — a vital feature which is missing in other baselines.

TABLE V. COMPARISON BETWEEN THE MODELS

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
LSA [20]	39.75	8.45	15.1	25.5
GPT-4 [21]	41.38	9.03	15.25	21.3
PacSum [22]	41.03	10.53	15.47	27.8
Vicuna7B [23]	48.21	14.92	20.12	28.7
Lead-K [24]	41.99	10.96	16.13	27.3
TextRank [25]	35.64	7.85	14.77	23.02
SIMSUM [26]	44.38	12.2	18.13	28.02
Proposed Model	**52.8**	**17.5**	**24.3**	**35.5**

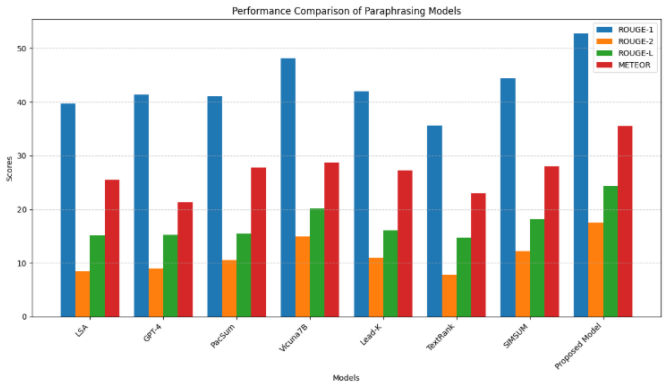

Fig. 8. Performance comparison with models.

Fig. 8 shows a comparative performance comparison of multiple paraphrasing models—LSA, GPT-4, PacSum, Vicuna7B, Lead-K, TextRank, SIMSUM, and proposed T5-XAVRL model—on standard evaluation metrics: ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. The figure clearly shows that the proposed model performs better compared to all the other baselines on all four metrics. Particularly, it scores the best ROUGE-1 (52.8), ROUGE-2 (17.5), ROUGE-L (24.3), and METEOR (35.5) results, which reflect its superior capacity to maintain meaning, preserve fluency, and fit academic stylistic demands. Importantly, although the Vicuna7B and SIMSUM models are also competitive in performance, they lag behind in terms of semantic coverage or stylistic conformity. This validates the efficacy of T5-XAVRL's attention-based reinforcement learning strategy and its strength to cope with the complicated requirements of academic.

7) Discussion. Results in academic text generation of the T5-XAVRL model on a task-oriented dataset are very good: it achieves a BLEU score of 68.7%, the highest of all existing models while also maintaining its fluency and semantic precision. Back translation and named entity recognition (NER) are well integrated with the model, increasing its recall and making the model achieve higher ROUGE scores. Finally, the analysis of tokenization efficiency supports that when facing rare word problem, subword based Tokenization methods, such as SentencePiece, perform better than traditional word-based Tokenization such as WordPiece. Furthermore, attention weight distribution further emphasizes the model's capability of paying attention to essential terminologies whilst preserving the semantics during paraphrasing. In addition, the model's performance in terms of METEOR further corroborates its effectiveness since it shows strong lexical alignment with reference texts. While the model performs very well on unigram and phrase continuity, its ROUGE-2 and ROUGE-L scores indicate that there is some scope for improvement in capturing longer dependencies. The proposed approach is compared to baseline models and typically outperforms traditional methods like TextRank and LSA; the result suggests that advanced encoding mechanisms and attention-based reinforcement learning indeed help the text generation quality.

VI. CONCLUSION AND FUTURE WORKS

The T5-XAVRL model shows exceptional performance in all major assessment metrics for generating academic text, especially in terms of fluency, coherence, and structural coherence. Incorporating attention-steered reinforcement learning, subword tokenization, and back-translation methods, the model is able to seize the subtle patterns inherent in scholarly style. Comparative testing shows that conventional paraphrasing models, quantified by BLEU, ROUGE, and METEOR scores, fall short when compared to T5-XAVRL, with the latter performing better in lexical matching and semantic accuracy. Attention-weight analyses also confirm that the model can pay attention to essential academic jargon while holding onto the original sense in paraphrasing. Although it excels in unigram and phrasal continuity, the model's capability to handle long-range dependencies in text is an area to be improved in future work. Additional training data covering a wider range of academic fields would enhance generalizability, while integrating contrastive learning methods would enhance factual accuracy. Enabling the embedding of user feedback mechanisms is a chance for adaptive and personalized paraphrasing ability. In addition, investigating multimodal extensions—such as adding visual or audio context—may enrich the model's applicability across various research and learning settings. Lastly, decreasing the computational complexity of the model using processes such as model distillation is vital for facilitating real-time usage and greater availability. In general, these directions support the construction of AI-based academic writing tools more context-driven, user-driven, and applicable in practice. Although the T5-XAVRL model performs strongly, there are some limitations. The model can still be better at capturing long-range dependencies, as evidenced by ROUGE-L scores. The system is also trained predominantly on English scholarly texts, and thus its generalizability across languages and subjects might be restricted. Computational complexity is also an issue, and this might limit real-time applications on minimal hardware. Subsequent research will tackle these issues through the investigation of model distillation, multilingual data sets, and general domain adaptation.

REFERENCES

[1] T. Ait Baha, M. El Hajji, Y. Es-Saady, and H. Fadili, "The Power of Personalization: A Systematic Review of Personality-Adaptive Chatbots," SN COMPUT. SCI., vol. 4, no. 5, p. 661, Aug. 2023, doi: 10.1007/s42979-023-02092-6.

[2] H. Deng, J. Jiang, Z. Yu, J. Ouyang, and D. Wu, "CrossGAI: A Cross-Device Generative AI Framework for Collaborative Fashion Design," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 1, pp. 1–27, Mar. 2024, doi: 10.1145/3643542.

[3] W. Guan et al., "Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis," in Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 18117–18125. Accessed: Mar. 10, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29769

[4] Y. Shi, H. Ding, K. Chen, and Q. Huo, "APRNet: Attention-based Pixel-wise Rendering Network for Photo-Realistic Text Image Generation," Mar. 15, 2022, arXiv: arXiv:2203.07705. doi: 10.48550/arXiv.2203.07705.

[5] Y. Yousaf, M. Shoaib, M. A. Hassan, and U. Habiba, "An intelligent content provider based on students learning style to increase their engagement level and performance," Interactive Learning Environments, vol. 31, no. 5, pp. 2737–2750, Jul. 2023, doi: 10.1080/10494820.2021.1900875.

[6] H. Xiao, W. Yao, H. Chen, L. Cheng, B. Li, and L. Ren, "SCDA: A Style and Content Domain Adaptive Semantic Segmentation Method for Remote Sensing Images," Remote Sensing, vol. 15, no. 19, p. 4668, 2023.

[7] H. Chen, F. Shao, B. Mu, and Q. Jiang, "Image aesthetics assessment with emotion-aware multibranch network," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1–15, 2024.

[8] X. Chi and Y. Xiang, "Augmenting Paraphrase Generation with Syntax Information Using Graph Convolutional Networks," Entropy, vol. 23, no. 5, 2021, doi: 10.3390/e23050566.

[9] T. Niu, S. Yavuz, Y. Zhou, N. S. Keskar, H. Wang, and C. Xiong, "Unsupervised Paraphrasing with Pretrained Language Models," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5136–5150. doi: 10.18653/v1/2021.emnlp-main.417.

[10] J. Weston, R. Lenain, U. Meepegama, and E. Fristed, "Generative Pretraining for Paraphrase Evaluation," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4052–4073. doi: 10.18653/v1/2022.acl-long.280.

[11] J. Lee, B. Liang, and H. Fong, "Restatement and Question Generation for Counsellor Chatbot," in Proceedings of the 1st Workshop on NLP for Positive Impact, A. Field, S. Prabhumoye, M. Sap, Z. Jin, J. Zhao, and C. Brockett, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 1–7. doi: 10.18653/v1/2021.nlp4posimpact-1.1.

[12] J. Li, Z. Li, L. Mou, X. Jiang, M. R. Lyu, and I. King, "Unsupervised Text Generation by Learning from Search." 2020. [Online]. Available: https://arxiv.org/abs/2007.08557

[13] M. M. Htay, Y. K. Thu, H. A. Thant, and T. Supnithi, "Statistical Machine Translation for Myanmar Language Paraphrase Generation," in 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1–6. doi: 10.1109/iSAI-NLP51646.2020.9376783.

[14] L. Qian, L. Qiu, W. Zhang, X. Jiang, and Y. Yu, "Exploring Diverse Expressions for Paraphrase Generation," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3173–3182. doi: 10.18653/v1/D19-1313.

[15] Z. Li, X. Jiang, L. Shang, and H. Li, "Paraphrase Generation with Deep Reinforcement Learning," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3865–3878. doi: 10.18653/v1/D18-1421.

[16] A. Humberto and S. Pinto, "Towards Human-in-the-Loop Computational Rhythm Analysis in Challenging Musical Conditions," 2023, Accessed: Mar. 10, 2025. [Online]. Available: https://repositorio-aberto.up.pt/bitstream/10216/153038/2/644518.pdf

[17] A. Berro, B. Benatallah, Y. Gaci, and K. Benabdeslem, "Error Types in Transformer-Based Paraphrasing Models: A Taxonomy, Paraphrase Annotation Model and Dataset," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 332–349.

[18] N. A. Alshameri and H. S. Al-Khalifa, "Arabic Paraphrase Generation Using Transformer-Based Approaches," IEEE Access, 2024.

[19] Joe Tricot, Brian Maltzan, and Timo Bozsolik, "arXiv Dataset." Accessed: Mar. 06, 2025. [Online]. Available: https://www.kaggle.com/datasets/Cornell-University/arxiv

[20] J. Steinberger and K. Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation," Jan. 2004.

[21] O. J. Achiam et al., "GPT-4 Technical Report," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257532815

[22] H. Zheng and M. Lapata, "Sentence Centrality Revisited for Unsupervised Summarization," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6236–6247. doi: 10.18653/v1/P19-1628.

[23] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 46595–46623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[24] D. Liu, Y. Wang, J. Loy, and V. Demberg, "SciNews: From Scholarly Complexities to Public Narratives -- A Dataset for Scientific News Report Generation," Dec. 10, 2024, arXiv: arXiv:2403.17768. doi: 10.48550/arXiv.2403.17768.

[25] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, D. Lin and D. Wu, Eds., Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. [Online]. Available: https://aclanthology.org/W04-3252/

[26] S. Blinova, X. Zhou, M. Jaggi, C. Eickhoff, and S. A. Bahrainian, "SIMSUM: Document-level Text Simplification via Simultaneous Summarization," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9927–9944. doi: 10.18653/v1/2023.acl-long.552.