IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5867∼5880
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5867-5880

Bilingual explainable paraphrase generation with transformers and reinforcement learning

Dalia Sameh[1,2], Ahmed El-Sawy[1,3], Hamada Nayel[1,4]
[1]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt
[2]Department of Computer Science, Modern Academy for Computer Science and Management Technology, Cairo, Egypt
[3]Department of Information Technology, Faculty of Technological Industry and Energy, Delta Technological University, Quesna, Egypt
[4]Department of Computer Engineering and Information, College of Engineering, Prince Sattam Bin Abdeluziz University, Al-Kharj, Saudi Arabia

Article Info

Article history:

Received Oct 6, 2025
Revised Aug 21, 2026
Accepted Sep 19, 2026

Keywords:

Arabic paraphrasing benchmark
Generation
PEGASUS
Quora
Reinforcement learning
Shapley additive explanations

ABSTRACT

Paraphrase generation is essential in many natural language processing (NLP) applications, including text summarization, semantic search, and question answering. Arabic paraphrase generation is still understudied because of the scarcity of resources and the linguistic complexity of Arabic, although the current transformer-based language models have greatly improved English paraphrasing. To improve bilingual generation, rule-based synonym substitution, transformer-based generation, cross-lingual back-translation, reinforcement learning (RL), and a hybrid approach incorporating Shapley additive explanations (SHAP)-based feedback are employed. The Arabic paraphrasing benchmark (APB) and the Quora question pairs (QQP) benchmarks were used to evaluate the proposed models. The proposed results demonstrate that, although SHAP-enhanced pipelines achieve competitive generation quality while offering greater interpretability, traditional rule-based models are outperformed by cross-lingual and RL–based approaches. The RL-based model achieved the highest English semantic preservation with a bidirectional encoder representations from transformers score (BERTScore) of 0.9234, while the hybrid model achieved the best Arabic performance with a BERTScore of 0.9020. These findings demonstrate that integrating RL with explainable transformer-based models provides an effective and trustworthy solution for bilingual paraphrase generation.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Dalia Sameh
Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University
Benha 13518, Egypt
Email: d.amin50431@fci.bu.edu.eg

1. INTRODUCTION

Text generation is a crucial task in natural language processing (NLP), it combines artificial intelligence and computational linguistics to generate outputs that are grammatically correct, coherent, and aware of context. Among its various forms, paraphrase generation is essential for answering questions, retrieving information, summarizing, and conducting semantic searches in both low-resource, morphologically rich languages like Arabic and high-resource languages like English [1]. The goal of paraphrase generation is to reiterate a statement while maintaining its original meaning by utilizing new language and structures. Paraphrasing enhances training data in machine translation and aids in systems' ability to generalize to

unknown language variances. It allows computers to identify semantically similar questions that are stated in various ways, which enhances user involvement and retrieval accuracy in question answering and information retrieval. By providing numerous methods to convey the same concept, paraphrase generation in educational technology facilitates automated essay grading, text simplifying, and language acquisition [2].

Additionally, paraphrasing improves data augmentation for low-resource languages, allowing for the establishment of more efficient models in fields with a lack of annotated data [3]. Paraphrase generation is regarded as a crucial job in promoting human–machine communication and guaranteeing semantic variety in NLP due to its wide range of applications spanning fields including conversational agents, summarization, plagiarism detection, and cross-lingual systems. Conventional techniques have shown limited scalability and generality, such as statistical machine translation, rule-based replacement, and lexical transformations utilizing tools like WordNet. On the other hand, the fluency, semantic retention, and domain adaptation of generated paraphrases have been greatly improved by recent developments in transformer-based architectures, including text-to-text transfer transformer (T5), pre-training with extracted gap-sentences for abstractive summarization (PEGASUS), bidirectional and auto-regressive transformers (BART), finetuned language Net–T5 (FLAN-T5), and Arabic bidirectional encoder representations from transformers (AraBERT) [4], [5].

Arabic is one of the most commonly spoken languages in the world, with more than 400 million speakers in more than 20 countries, making it an especially significant emphasis. Arabic is also used in Islamic liturgies, which makes it essential in religious, cultural, and educational contexts all across the world. In addition to facilitating better NLP applications for native speakers, the development of strong Arabic paraphrase generation systems advances cross-lingual technologies, making digital information easier to access [6]. Although English has been the subject of the majority of studies, there are particular difficulties in applying these methods to Arabic, such as the language's rich morphology, limited resources, dialectal variances, and absence of extensive parallel corpora [7]. To fill this need, the Arabic paraphrasing benchmark (APB) was created, allowing for the methodical assessment of Arabic-specific paraphrase-generating systems [8].

Furthermore, assessment is still a challenge since, despite their widespread usage, surface-level measures like bilingual evaluation understudy (BLEU) [9] and recall-oriented understudy for gisting evaluation–longest common subsequence (ROUGE-L) [10] sometimes do not accurately indicate semantic integrity. These metrics mostly use word-overlap, which frequently penalizes legitimate paraphrases that employ distinct syntactic or lexical structures. Additionally, they suffer from semantic blindness, which causes them to generate outputs with comparable surface shapes but inaccurate or conflicting interpretations in high rankings. In morphologically rich languages like Arabic, where minor changes in word forms or affixes can disproportionately lower assessment scores, the issue is made worse. Furthermore, these surface-based metrics are less accurate for assessing paraphrase generation as they frequently don't match human assessments of readability, fluency, and meaning retention bidirectional encoder representations from transformers score (BERTScore) [11], which provides better semantic sensitivity and more alignment with human judgment—especially in multilingual environments—has been introduced to address these issues.

Recent advances in large language models (LLMs) have further improved multilingual text generation by leveraging large-scale pre-training and instruction tuning. Models such as generative pre-trained transformer (GPT)-based architectures and multilingual transformer variants have demonstrated strong paraphrasing capabilities across multiple languages. However, these approaches generally emphasize generation quality without explicitly incorporating explainability or reinforcement learning (RL) into the optimization process, particularly for low-resource languages such as Arabic. Although lexical substitution, neural generation, back-translation, and RL have all been explored for paraphrase generation, most existing studies focus on a single generation strategy, a single language, or a limited evaluation perspective.

Moreover, explainability is rarely integrated into bilingual paraphrase generation frameworks, particularly for morphologically rich languages such as Arabic, where understanding model decisions is essential for improving transparency and trustworthiness. Consequently, there remains a need for a unified bilingual framework that combines transformer-based generation, RL, and explainable artificial intelligence (XAI) while supporting both English and Arabic. To address this gap, this study proposes an explainable bilingual paraphrase generation framework that systematically compares multiple generation strategies and integrates Shapley additive explanations (SHAP)-based explainability with transformer-based models and RL to improve both generation quality and interpretability.

The main contributions of this work are as follows: i) a systematic implementation and comparison of lexical, neural, cross-lingual, explainable, and RL–based paraphrase generation strategies; ii) a comprehensive evaluation is conducted using both surface-level and semantic-aware metrics, including BLEU, ROUGE-L, and BERTScore; and iii) development of a hybrid and explainable paraphrase generation framework that integrates SHAP-based interpretability with semantic similarity models such as robustly optimized BERT pretraining approach (RoBERTa) and AraBERT, combined with RL, bidirectional long short-term memory (BiLSTM), and fast text representation and classification (FastText) components; and an empirical analysis of controllability, interpretability, and domain adaptation in bilingual paraphrase generation was conducted. This is how the rest of the paper is organized: section 2 describes the proposed methodology and experimental setup. Section 3 presents the experimental results and discussion. Section 4 concludes the paper and outlines future research directions.

2. METHOD

2.1. Task definition and problem formulation

The goal of paraphrase generation, which is designed as a sequence-to-sequence learning challenge, is to generate a sentence that maintains the input's semantic meaning while displaying lexical or syntactic diversity. When given an input sentence x, the model generates a paraphrase y that maximizes the semantic similarity sim(x, y) under grammatical and fluency restrictions. This work focuses on generating bilingual paraphrases in Arabic and English. A single-generation approach is inadequate due to the linguistic differences between morphologically rich, low-resource languages (Arabic) and high-resource languages (English). To balance fluency, diversity, semantic integrity, and interpretability, a multi-strategy framework is adopted.

2.2. Mathematical formulations

Given an input sentence $x = (x_1, x_2, \ldots, x_n)$, the paraphrase generation task aims to generate an output sentence $y = (y_1, y_2, \ldots, y_m)$ such that the semantic similarity between x and y is maximized while preserving fluency and grammatical correctness. The generation objective can be expressed as in (1).

$$y^* = \arg\max_y P(y \mid x; \theta) \tag{1}$$

Where θ represents the parameters of the paraphrase generation model.

2.3. Overall system architecture

Figure 1 illustrates the general design of the suggested framework. The system uses a modular pipeline that includes explainability analysis, assessment, hybrid selection, different paraphrase-generating algorithms, and preprocessing. First, a sentence is preprocessed using tokenization and normalization. Several paraphrase-generating modules, including lexical-based, transformer-based, and cross-lingual generation techniques, receive the processed sentence. Semantic similarity models are used to filter and rank the collection of potential paraphrases generated by these modules. The final outputs are assessed using both surface-level and semantic-aware metrics, and explainability is integrated using SHAP to investigate token-level contributions.

2.4. Text preprocessing

Text preprocessing includes lowercasing (for English), normalization of Arabic characters, removal of diacritics, and tokenization using the tokenizer associated with each transformer model. Stop-word removal is applied only in lexical-based generation. Detokenization is performed using the Treebank detokenizer for English and rule-based formatting for Arabic.

2.5. Paraphrase generation strategies

To capture diverse paraphrasing behaviors, five complementary generation strategies are employed. These strategies are discussed in the following subsection, including lexical-based generation, transformer-based generation, and cross-lingual back-translation. The subsequent subsections cover hybrid and context-aware selection and the explainable RL framework.

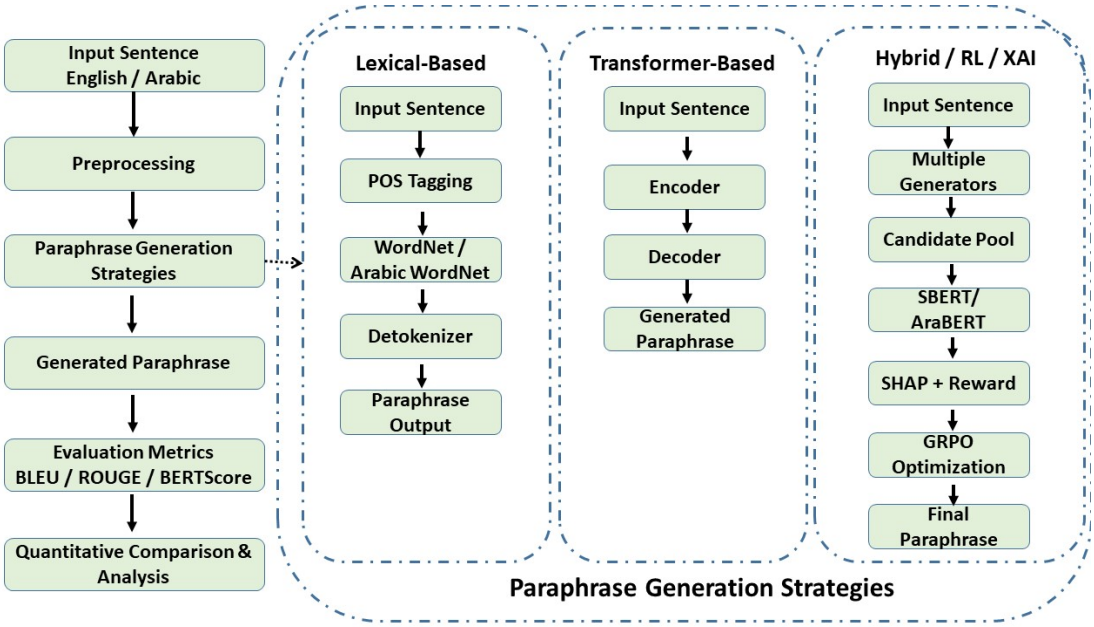


Figure 1. The overall experimental pipeline and the paraphrase generation strategies

2.5.1. Lexical-based generation

Lexical-based approaches are interpretable and computationally efficient [12], but they may introduce semantically incorrect replacements and frequently lack contextual awareness. They are therefore regarded as baseline generators rather than ultimate solutions. WordNet was used in lexical-based generation as follows: i) English: syntactic integrity is preserved while lexical variety is guaranteed using WordNet-based [13] synonym substitution in generation English; and ii) Arabic: to maintain grammaticality, synonym replacement is done using Arabic WordNet (AWN) [14], which is consistent with tokenization norms.

2.5.2. Transformer-based generation

Transformers are a type of neural network architecture designed to process sequential data, such as text, by using self-attention mechanisms to model relationships between all elements in the sequence simultaneously [15]. Transformers allow for effective modeling of long-range dependencies and contextual relationships in text. For transformer-based generation, various models have been used as follows: i) English: uses models that have already been trained for paraphrasing, such as Parrot [16], FLAN-T5 [17], and PEGASUS [13], [18]; and ii) Arabic: rewrites sentences using Arabic T5 [4] paraphrases. However, transformer outputs might occasionally diverge semantically, which is why explainability and downstream semantic filtering are used.

2.5.3. Cross-lingual back-translation

Back-translation [19] is especially useful in low-resource contexts, such as Arabic, because it adds structural variety while preserving semantic equivalency. Since Marian machine translation (MarianMT) [20] is efficient and provides high-quality pretrained models, it is utilized for neural machine translation. MarianMT was utilized in this work as follows: i) English: MarianMT enhances diversity via language transformation by translating English to Arabic and back to English; and ii) Arabic: reformulates Arabic phrases both structurally and semantically by using the Arabic-English translation and vice versa.

2.6. Hybrid and context-aware selection

A candidate pool is generated by combining the results of all generation techniques. Then, depending on the language, a hybrid selection process is used to rank and filter candidates utilizing semantic similarity models, such as sentence bidirectional encoder representations from transformers (SBERT) [21], RoBERTa, and AraBERT. Furthermore, in situations when controllability is required, generation is guided by context-aware and style-conditioned models (such as mT5 [22] with formal or informal prompts). Robustness and generalization across domains and language structures are enhanced by this hybrid method.

2.7. Explainable reinforcement learning framework

To improve the model transparency, an XAI layer was added. In both languages, explainable paraphrasing: enhances model transparency and guarantees content consistency during paraphrase generation by incorporating SHAP to highlight and retain semantically significant tokens. The explainable RL pipeline is illustrated in Figure 2.

The framework begins by preprocessing the input sentence before generating candidate paraphrases using transformer-based generation models. The generated paraphrases are then evaluated by a semantic similarity model to measure semantic preservation with respect to the input sentence. Simultaneously, SHAP is applied to identify the most influential input tokens, providing explainability information that is incorporated into the reward function as a penalty term for less informative token contributions.

The resulting reward, which combines semantic similarity and explainability feedback, is used by the group relative policy optimization (GRPO) optimization loop to update the generation policy iteratively. After each optimization step, token filtering and re-weighting are performed based on the SHAP importance scores. This iterative process continues until the explainability constraint is satisfied, producing a final paraphrase that balances semantic fidelity, generation quality, and interpretability.

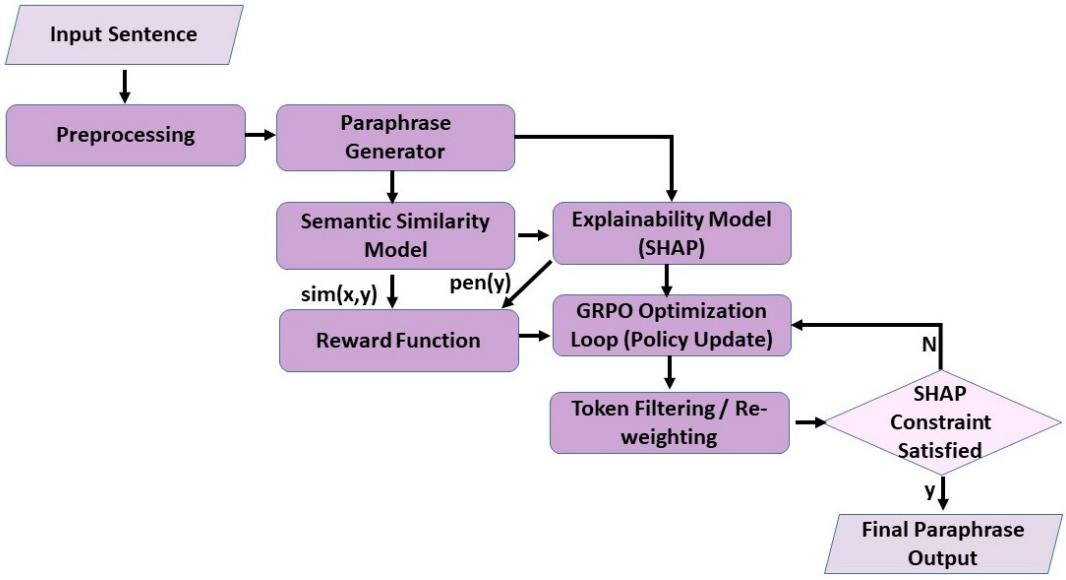


Figure 2. Explainable RL architecture for paraphrase generation integrating semantic similarity, SHAP-guided reward shaping, and GRPO policy optimization

2.7.1. Reward function design

A semantic similarity model is used to assess each produced paraphrase candidate. Semantic integrity between the input and output paraphrase is encouraged by using the resultant similarity score (such as cosine similarity of SBERT embeddings or BERTScore) as a reward signal [23]. The reward function $R(x, y)$ is defined in (2) using a semantic similarity score between the input sentence x and the generated paraphrase y.

$$R(x, y) = \alpha \cdot \text{sim}(x, y) - \beta \cdot \text{pen}(y) \tag{2}$$

Where $\text{sim}(x, y)$ denotes semantic similarity computed using SBERT or AraBERT cosine similarity, and $\text{pen}(y)$ represents a penalty term applied to paraphrases that alter semantically critical tokens identified by SHAP. α and β are weighting coefficients empirically tuned during validation.

The semantic similarity term encourages preservation of the original meaning, whereas the SHAP-based penalty discourages modifications to semantically important tokens identified by SHAP. By adjusting the weighting coefficients α and β, the optimization balances semantic preservation, paraphrase quality, and explainability. This design enables the GRPO optimization process to improve generation quality while maintaining faithful semantic representations and interpretable token importance.

2.7.2. Group relative policy optimization

GRPO is used to maximize paraphrase generation. GRPO evaluates many candidate paraphrases in relation to each other and modifies the generation policy as a result, rather than depending just on absolute rewards. This relative optimization removes the requirement for an explicit evaluation network and stabilizes training. By prioritizing outputs with greater semantic similarity and fluency, GRPO is utilized to fine-tune the paraphrase generator and enable direct optimization toward task-specific objectives.

GRPO updates the generation policy by ranking sampled paraphrase candidates within a group. Given a set of K generated paraphrases $(y_1, y_2, \ldots, y_k)$, the advantage for a candidate y_i is computed as in (3).

$$A_i = R(x, y_i) - \frac{1}{K} \sum_j R(x, y_j) \tag{3}$$

The policy parameters θ are updated by maximizing as shown in (4).

$$\mathcal{L}(\theta) = \mathbb{E}[A_i \cdot \log P(y_i \mid x; \theta)] \tag{4}$$

This relative ranking stabilizes optimization and removes the need for a separate value function.

2.7.3. Shapley additive explanations additive explanations-guided explainability

Token-level contribution scores are calculated by applying SHAP [24] to the semantic similarity model to improve interpretability. The tokens that contribute most to semantic preservation are identified by these scores. Two methods are used to incorporate SHAP explanations into the pipeline: i) token filtering, where paraphrases that alter highly important tokens are penalized; and ii) reward shaping, where GRPO rewards are adjusted based on SHAP importance to preserve semantically critical content. The SHAP-guided GRPO training process is presented in Algorithm 1. This integration maintains high generation quality while making paraphrase generation clear and understandable.

Algorithm 1. GRPO training with SHAP-guided reward for paraphrase generation

Require: Training dataset $D = \{x_1, \ldots, x_N\}$
Require: Paraphrase generator policy π_θ
Require: Semantic similarity model S
Require: SHAP explainer E
Require: Number of samples per input K
Require: Reward weights α, β
Require: Learning rate η
Ensure: Optimized policy π_θ^*

```
Initialize policy parameters θ
for each training epoch do
    for each input sentence x ∈ D do
        Sample K candidates {y_1, ..., y_K} ~ π_θ(· | x)
        for each candidate y_i do
            sim(x, y_i) ← S(x, y_i)
        end for
        Compute SHAP scores Φ ← E(x, {y_i})
        for each candidate y_i do
            R(x, y_i) ← α · sim(x, y_i) − β · pen(y_i)
        end for
        R̄ ← (1/K) Σ_i R(x, y_i)
        for each candidate y_i do
            A_i ← R(x, y_i) − R̄
        end for
        Update θ using Σ_i A_i log π_θ(y_i | x)
    end for
end for
return π_θ*
```

Experiments are conducted using two benchmark datasets, showing the characteristics in Table 1: i) English: more than 400,000 question pairs with semantic similarity labels are included in the Quora question pairs (QQP) dataset [25]; and ii) Arabic: a relevant tested for assessing Arabic paraphrase generation and detection is the APB dataset, which consists of sentence pairs classified as paraphrase (p) or non-paraphrase (np) [8]. All datasets were randomly split using fixed random seeds to ensure reproducibility. For English, text was lowercased and tokenized using the corresponding transformer tokenizer. For Arabic, normalization, punctuation removal, and Farasa-style tokenization were applied to reduce morphological variation.

Table 1. Dataset characteristics table

Dataset	Language	#pairs	Train (%)	Validation (%)	Test (%)
QQP	English	404K	80	10	10
APB	Arabic	70K	80	10	10

2.8. Hyperparameter configuration

The experimental hyperparameters utilized in all training and evaluation experiments are compiled in Table 2. When appropriate, the same optimization parameters were used to guarantee a fair comparison of various paraphrase generating techniques. The optimizer, learning rate, batch size, maximum sequence length, RL configuration, random seeds, and pretrained models used in the suggested framework are all included in the table. Each experimental parameter is listed in the first column, and the matching value used during training is reported in the second column. The rows collectively depict the whole experimental setup used in every bilingual paraphrase generation experiment. All experiments were trained using adaptive moment estimation with weight decay (AdamW) with a learning rate of $2e^{-}5$ and a batch size of 8, as indicated in Table 2. A maximum sequence length of 128 tokens was chosen in order to achieve a compromise between contextual representation and computing performance. To increase repeatability, three random seeds (42, 52, and 62) were utilized.

Table 2. Experimental hyperparameter

Parameter	Value
Batch size	8
Learning rate	$2e^{-}5$
Optimizer	AdamW
Max sequence length	128
RL samples per input	5
Epochs	3
Random seeds	42, 52, 62
Training time (per run)	25–30 minutes
Base models	PEGASUS, FLAN-T5, Parrot, MarianMT, and mT5
Embedding model	SBERT (paraphrase-multilingual-mpnet-base-v2)

2.9. Software and hardware requirements

All experiments were conducted using Python 3.10 and the HuggingFace Transformers library. Training and inference were performed on Google Colab with NVIDIA Tesla T4 GPUs (16 GB VRAM). Evaluation was conducted using BLEU, ROUGE, and the official BERTScore implementation. SHAP explanations were computed using the SHAP library for transformer-based models.

2.10. Statistical rigor

Each experiment was repeated three times using different random seeds. The mean and standard deviation are reported for all evaluation metrics. Statistical significance of improvements is assessed using paired t-tests with $p < 0.05$. Three commonly used measures were employed to thoroughly assess the fluency and semantic quality of the resulting paraphrases. i) BLEU: the n-gram overlap between produced and reference texts is measured by the BLUE [9]; ii) ROUGE-L: evaluates recall-oriented match by concentrating on the longest common subsequence [10]; and iii) BERTScore: calculates semantic similarity by utilizing contextual embeddings from BERT-like models BERTScore [11]. In section 3, both surface-level (BLEU and ROUGE-L) and semantic-aware (BERTScore) metrics are used to develop, assess, and compare generation quality across all strategies.

3. RESULTS AND DISCUSSION

The suggested paraphrase-generating techniques for Arabic and English are thoroughly evaluated in this section. Generation quality is evaluated using both surface-level and semantic-aware measures. The outcomes are compared with previous methods.

3.1. English paraphrase generation results

The baseline results for English paraphrase generation utilizing WordNet and PEGASUS-based synonym expansion, followed by T5 generation, are shown in Table 3. PEGASUS demonstrates the benefit of contextualized synonym generation over rule-based lexical replacement by consistently outperforming WordNet across all ROUGE variants and BLEU. Here's a summary Table 4 comparing all paraphrase generation strategies for English tested using BLEU, ROUGE-L, and BERTScore.

Table 3. Baseline comparison

Evaluation	Wordnet + T5	PEGASUS + T5
Rouge 1	0.2321	0.5801
Rouge 2	0.0681	0.5479
Rouge L	0.2308	0.5791
Avg. BLEU	0.0087	0.4268

Table 4. English paraphrase generation results on Quora (mean ± SD over 3 seeds)

Strategy	BLEU	ROUGE-L	BERTScore
WordNet-Based	0.0860 ± 0.0008	0.3702 ± 0.0014	0.6565 ± 0.0008
PEGASUS	0.1685 ± 0.0013	0.4544 ± 0.0010	0.7805 ± 0.0004
Parrot	0.2167 ± 0.0005	0.5356 ± 0.0015	0.8653 ± 0.0006
FLAN-T5	0.1228 ± 0.0001	0.3630 ± 0.0015	0.7401 ± 0.0018
Back-translation (EN-AR-EN)	0.1721 ± 0.0020	0.4858 ± 0.0006	0.8286 ± 0.0005
Hybrid (T5-Parrot-SBERT filter)	0.1543 ± 0.0300	0.4486 ± 0.0033	0.9106 ± 0.0099
RL Fine-Tuned (BERTScore reward)	0.1506 ± 0.0066	0.4492 ± 0.0074	0.9231 ± 0.0003
Context-aware (FLAN-T5)	0.0326 ± 0.0068	0.1544 ± 0.0112	0.8325 ± 0.0018
Explainable hybrid (SHAP feedback)	0.1149 ± 0.0062	0.3544 ± 0.0350	0.8829 ± 0.0079

There are noticeable differences between lexical, neural, hybrid, and RL techniques, according to the comparative analysis of the suggested strategies. WordNet and other rule-based approaches performed the worst on all criteria, demonstrating their inability to capture semantic richness beyond simple synonym replacement. Although transformer based models such as PEGASUS and FLAN-T5 showed significant gains, Parrot did better than them in terms of BLEU of 0.2167 $\pm$ 0.0005 and ROUGE-L of 0.5356 $\pm$ 0.0015, suggesting a greater degree of lexical and structural similarity with reference paraphrases. Additionally, cross-lingual back-translation showed competitive outcomes by achieving a balance between semantic faithfulness and variety.

The hybrid framework achieved a BERTScore of 0.9106 ± 0.0099 was reached by hybridization techniques, especially the T5–Parrot–SBERT filtering strategy, highlighting their ability to generate meaning-preserving paraphrases even with less lexical overlap. The greatest semantic consistency of 0.9231 $\pm$ 0.0003 was attained by RL using a BERTScore-based reward, highlighting the advantages of optimization over supervised training. Finally, the explainable hybrid approach increased interpretability through SHAP-based feedback and maintained a high BERTScore of 0.8829 $\pm$ 0.0079, making it appropriate when transparency is needed, while being marginally poorer in BLEU and ROUGE. Overall, hybrid techniques provided a compromise between variety, fidelity, and explainability, RL fine-tuning delivered the strongest semantic preservation, while Parrot excelled in structural similarity. The superior BERTScore achieved by RL fine-tuning validates the reward formulation described in subsection 2.7. and Algorithm 1.

The superior semantic preservation achieved by the hybrid framework can be attributed to the complementary strengths of its individual components. Transformer-based generation produces contextually diverse paraphrase candidates, while SBERT-based semantic filtering removes candidates with lower semantic similarity. Compared with lexical substitution, which is limited to word-level replacements, and back-translation, which may introduce translation artifacts, the hybrid framework explicitly preserves semantic

consistency while maintaining linguistic diversity. This complementary design explains the consistently higher BERTScore obtained by the proposed hybrid approach. Statistical significance testing was used to confirm the stated improvements' robustness. BERTScore values from three random seeds (42, 52, and 62) were subjected to paired t-tests. RL with a semantic reward produces persistent and statistically reliable gains, as seen by the RL fine-tuned model's considerable ($p < 0.01$) outperformance of all non-RL baselines. There were no statistically significant differences between powerful non-RL generators such as Parrot and PEGASUS.

To evaluate the effectiveness of the proposed paraphrase generation approach, its results are compared with those reported in previous studies using the QQP dataset or other comparable datasets. The comparison considers the evaluation metrics reported in the literature, including BLEU, ROUGE-L, and BERTScore, where available. Table 5 summarizes the results of previous studies alongside the proposed approach, highlighting differences in performance while acknowledging that variations in datasets, experimental settings, and evaluation protocols may limit direct comparisons. The proposed results differ significantly from those of previous research in a number of significant points. The Quora dataset showed a modest ROUGE score of 33.61% and 39.55% for [26] and [12], respectively. These results showed gains in surface-level similarity but did not take deeper semantic alignment into consideration.

In the same direction, PEGASUS [18], performed well on ROUGE summarization tasks but was not paraphrasing optimized. The proposed experiments, however, demonstrate that Parrot outperforms other models in terms of lexical and structural overlap, producing the greatest BLEU of 0.2162 and ROUGE-L of 0.5350. Furthermore, the BERTScore-based rewards for RL produced the best semantic preservation with BERTScore 0.9234, clearly outperforming earlier approaches that did not include semantic similarity.

While the SHAP-based explainable hybrid maintained high performance with a BERTScore of 0.8823 with the extra advantage of interpretability, hybrid techniques, especially T5–Parrot with SBERT filtering scored a BERTScore of 0.9006, presented a balanced trade-off between lexical purity and semantic richness. These results demonstrate that the proposed approach adds explainability, which was lacking from previous models, in addition to outperforming previous work in terms of semantic integrity. These findings are consistent with recent studies showing that RL can improve semantic preservation by directly optimizing generation objectives beyond supervised learning. Unlike previous approaches that primarily relied on BLEU- or ROUGE-based optimization, the proposed reward formulation emphasizes semantic similarity, resulting in improved contextual consistency while maintaining paraphrase diversity.

Table 5. Comparison with prior work in English paraphrase generation

Study/model	Dataset	BLEU	ROUGE-L	BERTScore
Zhang *et al*. PEGASUS [18]	CNN/DailyMail	-	0.4111	-
Unsupervised paraphrase generation [26]	Quora	-	0.3361	-
Paraphrase generation by editing [12]	Quora	0.3346	0.3955	-
Proposed parrot	Quora	0.2162	0.5350	0.8650
Proposed Hybrid (T5-Parrot-SBERT)	Quora	0.1218	0.4513	0.9006
Proposed RL with BERTScore reward	Quora	0.1433	0.4432	0.9234
Proposed explainable hybrid with SHAP	Quora	0.1143	0.3627	0.8823

3.2. Arabic paraphrase generation results

Table 6 presents a comparative evaluation of the paraphrase generation strategies applied to the APB. The comparison uses BLEU, ROUGE-L, and BERTScore to assess lexical similarity, text overlap, and semantic similarity between the generated paraphrases and the reference texts. The results highlight the relative strengths and limitations of each strategy and provide a basis for identifying the most effective approach for Arabic paraphrase generation within the evaluated experimental setting. Results are reported as mean $\pm$ standard deviation over three random seeds (42, 52, and 62). Due to the metric's emphasis on longest common subsequence matching, which is extremely sensitive to word order and morphological variance, ROUGE-L scores are 0 in all Arabic trials. Paraphrases in Arabic frequently include significant inflectional and syntactic rearrangements that minimize apparent overlap while maintaining meaning. As a result, BLEU and ROUGE underestimate the true quality of paraphrases, while BERTScore offers a more trustworthy semantic assessment in this context.

Particularly in back-translation or mT5 techniques, many of the produced outputs depended on structural changes or semantic rephrasings that went beyond straightforward lexical overlap. Further difficulties

arise from the fundamental complexity of Arabic morphology and tokenization, as even little changes in word order or phrasing can lead to extremely low ROUGE-L scores. Therefore, unless lemmatization or more precise matching methods are used, conventional surface-level measures like BLEU and ROUGE may understate the actual quality of paraphrases in morphologically complex languages like Arabic.

Table 6. Arabic paraphrase generation results

Method description	BLEU	BERTScore
Arabic WordNet synonym replacement	0.354	0.898
MarianMT back-translation (AR–EN–AR)	0.071	0.799
Hybrid (Back-translation + synonym substitution + SBERT filter)	0.378	0.902
mT5 with <formal> / <casual> tokens	-	0.566
SHAP-guided filtering using RL and AraBERT classifier	0.099	0.820

BLEU is 0.000 for style-controlled mT5 as: without any Arabic-specific fine-tuning, the mT5 model was used in a zero-shot scenario using / style tokens. As a result, a large number of the produced outputs exhibited insufficient vocabulary and sentence structure alignment with the reference pairs. This mismatch resulted in scores that were nearly zero since BLEU mostly records n-gram overlap. Overall, the model frequently produced poor or even off-topic outputs and had trouble producing coherent paraphrases in Arabic.

With the greatest BERTScore of 0.902 and a competitive BLEU of 0.378 in Arabic, cross-lingual generation with synonym adaptation and SBERT filtering distinguished itself, demonstrating efficient semantic retention and structural variety. Although AWN produced a good lexical overlap BLEU of 0.354, fluency was a weak point. Lower recall resulted from integration of SHAP-based feedback, which assigned priority to semantic integrity while filtering out a large number of potential outputs. These results highlight the importance of explainable and hybrid systems that incorporate domain-adapted elements and semantic-aware filtering, particularly in low-resource languages.

The hybrid cross-lingual model significantly outperformed both back-translation-only and lexical baselines ($p < 0.01$), confirming that semantic filtering and cross-lingual augmentation lead to statistically reliable improvements. To account for Arabic morphological variation and inflectional richness, this study additionally report chrF++, a character-level metric that captures surface similarity beyond token-based overlap. Arabic WordNet achieves the highest chrF++ due to minimal lexical reordering, while the hybrid approach exhibits very low chrF++ despite achieving the highest BERTScore, indicating substantial structural divergence while preserving semantic meaning. This confirms that character-level and semantic metrics capture complementary aspects of paraphrase quality. chrF++ is not reported for the mT5 model due to unstable zero-shot outputs and insufficient character-level alignment.

To assess the effectiveness of the proposed Arabic paraphrase generation approach, its results are compared with those reported in previous studies using the APB or comparable datasets. The comparison considers the evaluation metrics reported in the literature to examine the quality and semantic similarity of the generated paraphrases. Table 7 summarizes the results of previous studies alongside the proposed approach, highlighting the relative performance of the methods while acknowledging that differences in datasets, experimental settings, and evaluation metrics may limit direct comparisons.

Table 7. Comparison with prior work in paraphrase generation for Arabic

Study/model	Dataset	BLEU	ROUGE-L	BERTScore
Arabic paraphrase generation [3]	Synthetic Arabic paraphrase corpus	0.3581	0.6367	0.8998
GPT-4 for Arabic paraphrase generation [5]	ARGEN dataset	0.50	0.60	–
Proposed hybrid (back-translation + synonym substitution + SBERT filter)	APB	0.378	0.000	0.902

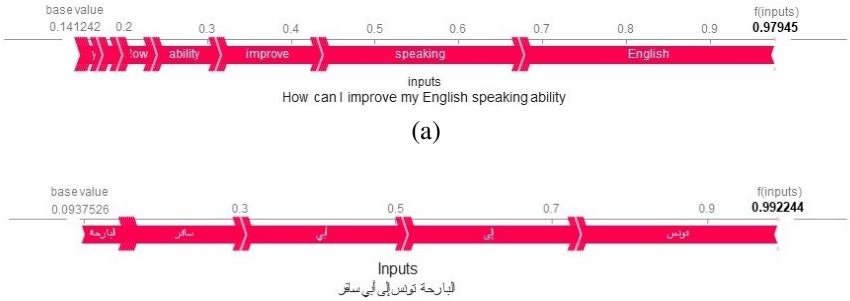
The comparison focuses attention to the relative advantages of several methods for generating Arabic paraphrases. With balanced BLEU 0.3581, ROUGE-L 0.6367, and BERTScore 0.8998, the transformer-based models refined by Al-Shameri and Al-Khalifa [3] had high overall performance, showing both good lexical overlap and semantic preservation. Although embedding-based semantic similarity BERTScore was not specifically reported, Alsulami and Almansour [5] used GPT-4 in a zero-shot context and reported the highest BLEU of 0.50 and ROUGE-L of 0.60 scores, proving the efficacy of large-scale generative models. However,

despite ROUGE-L dropping to zero because of structural divergence from reference sentences, the proposed hybrid approach (back-translation–synonym substitution–SBERT filter) achieved the highest semantic score with BERTScore 0.902 and a competitive BLEU of 0.378, demonstrating strong meaning preservation. This highlights a larger trend: hybrid and cross-lingual approaches capture deeper semantic similarity that may not be captured by n-gram-based measures, whereas large models and fine-tuned transformers succeed in surface overlap.

3.3. Qualitative explainability analysis

SHAP-based token attribution for sample paraphrase instances in Arabic and English is shown in Figure 3. Strong contributions to semantic similarity are shown by high SHAP values for tokens like English, speaking, and ability in English, as well as **تونس** in Arabic. Function words, on the other hand, have little effect. This explains the high BERTScore attained despite low surface-level overlap metrics in Arabic by confirming that the suggested SHAP-guided technique correctly finds semantically important tokens and enforces their retention during paraphrase generation.

SHAP-based token-level explanations for representative paraphrase generation examples in Arabic and English presented in Figure 3. The suggested multilingual paraphrase generation methodology is made easier to understand by the visualizations, which show how specific input tokens affect the similarity score between the source and generated sentences. The explanation for an English example is shown in Figure 3(a), whereas the explanation for an Arabic example is shown in Figure 3(b).



(a)

(b)

Figure 3. SHAP-based token-level explanations for (a) English and (b) Arabic paraphrase generation, showing each token's contribution to the semantic similarity score used for reward shaping

3.4. Qualitative paraphrase examples

Table 8 presents qualitative paraphrase examples for both English and Arabic generated using the proposed hybrid approach. Although the generated paraphrases exhibit limited surface-level n-gram overlap with their references, they preserve the original semantic meaning. This semantic equivalence is effectively captured by BERTScore, whereas BLEU and ROUGE underestimate paraphrase quality due to lexical variation, flexible word order, and morphological richness particularly in Arabic.

Table 8. Qualitative paraphrase examples in English and Arabic

Language	Type	Sentence
English	Original	How can I improve my English speaking skills quickly?
	Reference	What is the fastest way to get better at speaking English?
	Generated (hybrid)	What are effective methods to rapidly improve English speaking ability?
Arabic	Original	سافر أبي إلى تونس أمس
	Reference	ذهب والدي إلى تونس يوم أمس
	Generated (hybrid)	سافر والدي إلى تونس في اليوم السابق

3.5. Error analysis

Even though the performance was excellent generally, several mistake patterns were noted. Occasional lexical over-substitution in the English paraphrase generation trials generated paraphrases that retained meaning but marginally decreased fluency or stylistic naturalness, resulting in tiny drops in BLEU and ROUGE ratings.

Errors in Arabic paraphrase generation were mostly caused by the language's flexible syntax and complex morphology. Word-based measures like ROUGE underestimated quality despite strong semantic similarity, as demonstrated by BERTScore, because valid paraphrases frequently diverged significantly from reference sentences at the surface level.

The model mainly focuses on semantically significant elements, such as verbs and important nouns, according to SHAP-based explainability analysis. Excessive focus on certain tokens occasionally resulted in less-than-ideal lexical choices in less-than-ideal outputs, underscoring SHAP's value in identifying generation behavior. Due to practical limitations, a complete human review was not carried out; instead, this study focuses on automated metrics and explainability analysis.

4. CONCLUSION

This study used lexical, neural, cross-lingual, and XAI methodologies to provide a thorough assessment of paraphrase-generating methods for both English and Arabic. The techniques under investigation included transformer-based neural generators like FLAN-T5, PEGASUS, and Parrot, rule-based lexical replacement using WordNet and Arabic WordNet, and cross-lingual paraphrase via English–Arabic–English back-translation using MarianMT. Furthermore, to improve semantic preservation and interpretability, explainability-driven pipelines were suggested by integrating SHAP-based token attribution into hybrid and RL-based generation frameworks. Experiments employing BLEU, ROUGE-L, BERTScore, and chrF++ on the APB datasets and QQP showed that hybrid and RL-based approaches consistently produced better semantic validity. Specifically, BERTScore was statistically significantly improved by RL optimized with semantic incentives, and hybrid pipelines successfully balanced variety, fluency, and meaning preservation. For Arabic, where limited resources and complicated morphology present difficulties for solely neural or lexical techniques, cross-lingual generation has been particularly helpful. Additionally, SHAP-based explainability increased trust in model outputs and supported their implementation in sensitive or high-stakes applications by offering clear insights into token-level contributions. Despite these beneficial results, the performance differences between Arabic and English show that morphologically rich and low-resource languages continue to face difficulties. While English models obtained better surface-level scores because of more pretraining materials, Arabic paraphrase generation profited more from semantic-aware filtering and hybrid techniques. Subsequent research will concentrate on expanding the suggested framework to other low-resource languages, enhancing controllability using sophisticated prompt- or prefix-tuning techniques, and integrating human-in-the-loop evaluation for more complex quality assessment. Additionally, adding new XAI methods, such as local interpretable model-agnostic explanations (LIME) or attention-based visualizations, might improve paraphrase generation systems' transparency and user control even more.

FUNDING INFORMATION

This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2026/R/1448).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dalia Sameh	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓
Ahmed El-Sawy				✓	✓	✓				✓	✓	✓	✓	
Hamada Nayel	✓	✓		✓	✓	✓				✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created in this study.

REFERENCES

[1] G. Inoue, S. Khalifa, and N. Habash, "Morphosyntactic tagging with pre-trained language models for Arabic and its dialects," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1708–1719, doi: 10.18653/v1/2022.findings-acl.135.

[2] M. A. Ahmed, F. AlGhamdi, and A. Hawwari, "Constructing and augmenting a bidirectional paraphrases dataset from an English-Arabic subtitling parallel corpus," *The Egyptian Journal of Language Engineering*, vol. 11, no. 2, pp. 1–12, Oct. 2024, doi: 10.21608/ejle.2024.308019.1070.

[3] N. A. Al-Shameri and H. S. Al-Khalifa, "Arabic paraphrase generation using transformer-based approaches," *IEEE Access*, vol. 12, pp. 121896–121914, 2024, doi: 10.1109/ACCESS.2024.3450931.

[4] C. Raffel *et al*., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[5] H. R. Alsulami and A. A. Almansour, "Exploring GPT-4 capabilities in generating paraphrased sentences for the Arabic language," *Applied Sciences*, vol. 15, no. 8, Apr. 2025, doi: 10.3390/app15084139.

[6] K. Darwish *et al*., "A panoramic survey of natural language processing in the Arab world," *Communications of the ACM*, vol. 64, no. 4, pp. 72–81, Apr. 2021, doi: 10.1145/3447735.

[7] Warto *et al*., "Systematic literature review on named entity recognition: approach, method, and application," *Statistics, Optimization & Information Computing*, vol. 12, no. 4, pp. 907–942, Feb. 2024, doi: 10.19139/soic-2310-5070-1631.

[8] M. Alian, A. Awajan, A. Al-Hasan, and R. Akuzhia, "Building Arabic paraphrasing benchmark based on transformation rules," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 4, pp. 1–17, Jul. 2021, doi: 10.1145/3446770.

[9] A. M.-Cediel, J.-A. D.-H.-Gabaldon, E. G.-Lopez, A. G.-Cabot, and D. D.-F.-Dominguez, "Evaluating the performance of multilingual models in answer extraction and question generation," *Scientific Reports*, vol. 14, no. 1, Jul. 2024, doi: 10.1038/s41598-024-66472-5.

[10] W. M. Aly, T. H. A. Soliman, and A. M. AbdelAziz, "An evaluation of large language models on text summarization tasks using prompt engineering techniques," Jul. 07, 2025, *arXiv:2507.05123*.

[11] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020.

[12] A. Kazemnejad, M. Salehi, and M. S. Baghshah, "Paraphrase generation by learning how to edit from samples," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6010–6021, doi: 10.18653/v1/2020.acl-main.535.

[13] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 1875–1885, 2018, doi: 10.18653/v1/N18-1170.

[14] A. A. ElSabagh, S. S. Azab, and H. A. Hefny, "A comprehensive survey on Arabic text augmentation: approaches, challenges, and applications," *Neural Computing and Applications*, vol. 37, no. 10, pp. 7015–7048, Apr. 2025, doi: 10.1007/s00521-025-11020-z.

[15] S. Islam *et al*., "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Systems with Applications*, vol. 241, May 2024, doi: 10.1016/j.eswa.2023.122666.

[16] P. Damodaran, Parrot: Paraphrase generation for NLU. (Oct. 09, 2026), github.com. Accessed: Oct. 19, 2026. [Online]. Available: https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

[17] H. W. Chung *et al*., "Scaling instruction-finetuned language models," Dec. 06, 2022, *arXiv:2210.11416*.

[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning*, Nov. 2020, pp. 11328–11339.

[19] A. Malajyan, K. Avetisyan, and T. Ghukasyan, "ARPA: Armenian paraphrase detection corpus and models," in *2020 Ivannikov Memorial Workshop (IVMEM)*, Sep. 2020, pp. 35–39, doi: 10.1109/IVMEM51402.2020.00012.

[20] A. Acharekar, A. Bhosle, and M. Dwivedi, "OCR for multilanguage text extraction, translation and summarization," in *OCR for Multilanguage Text Extraction, Translation and Summarization*, 2025.

[21] L.-X. Kong, Y.-G. Chu, Z. Ma, J.-B. Zhang, and J.-J. Chen, "Mix-lingual relation extraction: dataset and a training approach," *Journal of Computer Science and Technology*, vol. 40, no. 1, pp. 42–59, Jan. 2025, doi: 10.1007/s11390-024-4314-y.

[22] L. Xue *et al*., "mT5: a massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.

[23] B. Heleen, M. Abubakar, and S. Dodda, "Deep reinforcement learning for text generation tasks," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5198622.

[24] Z. Shao *et al.*, "DeepSeekMath: pushing the limits of mathematical reasoning in open language models," 2024, *arXiv:2402.03300*.

[25] Quora, "Quora question pairs dataset," kaggle.com, 2024. Accessed: Oct. 20, 2026. [Online]. Available: https://www.kaggle.com/c/quora-question-pairs/data.

[26] C. Hegde and S. Patil, "Unsupervised paraphrase generation using pre-trained language models," Jun. 09, 2020, *arXiv:2006.05477*.

BIOGRAPHIES OF AUTHORS



Dalia Sameh received B.Sc. in Computer Science in 2012 at Faculty of Computer Science and M.Sc. in Computer Science in 2019 at Faculty of Computers and Informatics, Helwan University. She is a lecturer assistant in Department of Computer Science, Modern Academy for Computer Science and Management Technology, Egypt and Ph.D. researcher at Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University. Her research interests include neural network, machine learning, deep learning, pretrained models and NLP. She can be contacted at email: d.amin50431@fci.bu.edu.eg.



Ahmed El-Sawy received the B.S. and M.S. degrees from Cairo University, in 2000 and 2003, respectively, and the Ph.D. degree in Computer Science from Suez Canal University, in 2011. From 2004 to until 2011, he served as a lecturer in the Faculty of Computers and Information Systems, Taif University, Saudi Arabia. He has been a lecturer at Benha University, Egypt, since 2012, where he has been an associate professor in Computer Science, since 2017, and the head of computer science with the Faculty of Computers and Artificial Intelligence, since 2017. He is a dean at Faculty of Technological Industry and Energy, Delta Technological University, Quesna, Egypt. His research interests include using artificial intelligence in computer networks, bioinformatics, chimoinformatics, deep learning, and optimization. He can be contacted at email: ahmed.el_sawy@fci.bu.edu.eg.



Hamada Nayel is an assistant professor at Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University. In 2019, he received his Ph.D. from Mangalore University, India. He is an assistant professor at College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdeluziz University, Al-Kharj, Saudi Arabia. His research interests including biomedical natural language processing (BioNLP), NLP for Arabic language, NLP for cyber security and author profiling. The aim of current research work is introducing tools and methods for analyzing Arabic language. He can be contacted at email: hamada.ali@fci.bu.edu.eg.