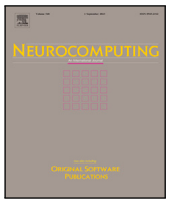
Neurocomputing 650 (2025) 130816

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



LexiSem: A re-ranker balancing lexical and semantic quality for enhanced abstractive summarization☆

Eman Aloraini [a,*], Hozaifa Kassab [b,c], Ali Hamdi [b], Khaled Shaban [a]

[a] Qatar University, Computer Science and Engineering Department, Qatar
[b] MSA University, Department of Computer Science, Egypt
[c] AiTech AU, Melbourne, Australia

ARTICLE INFO

Communicated by D. Cavaliere

Dataset link: https://github.com/MIRAH-Official/LexiSem

Keywords:
Abstractive summarization
Re-ranking
Lexical quality
Semantic quality
Deep learning

ABSTRACT

Sequence-to-sequence neural networks have recently achieved significant success in abstractive summarization, especially through fine-tuning large pre-trained language models on downstream datasets. However, these models frequently suffer from exposure bias, which can impair their performance. To address this, re-ranking systems have been introduced, but their potential remains underexplored despite some demonstrated performance gains. Most prior work relies on ROUGE scores and aligned candidate summaries for ranking, exposing a substantial gap between semantic similarity and lexical overlap metrics. In this study, we demonstrate that a second-stage model can be trained to re-rank a set of summary candidates, significantly enhancing performance. Our novel approach leverages a re-ranker that balance lexical and semantic quality. Additionally, we introduce a new strategy for defining negative samples in ranking models. Through experiments on the CNN/DailyMail, XSum and Reddit TIFU datasets, we show that our method effectively estimates the semantic content of summaries without compromising lexical quality. In particular, our method sets a new performance benchmark on the CNN/DailyMail dataset (48.18 R1, 24.46 R2, 45.05 RL) and on Reddit TIFU (30.37 R1,RL 23.87).

1. Introduction

Summarization is a key task in Natural Language Processing (NLP), enabling users to efficiently grasp complex textual content. Summarization techniques are broadly categorized into two approaches: extractive and abstractive [1]. The abstractive approach, which frames summarization as a sequence-to-sequence text generation task, has gained prominence due to advancements in pre-trained models such as PEGASUS [2], BART [3], and others [4,5]. These models have consistently achieved state-of-the-art results across multiple datasets, reducing reliance on extractive methods.

Typically, abstractive summarization models are trained using Maximum Likelihood Estimation (MLE), where the model learns to maximize the predictive probability of the reference summary, conditioned on preceding gold-standard sub-sequences. However, during inference, these models often suffer from exposure bias, a problem where generated sequences are influenced by previously predicted — potentially erroneous — tokens [6,7]. To mitigate this, re-ranking mechanisms have been proposed [8,9], aiming to improve the quality of generated summaries.

Re-ranking systems in abstractive summarization typically optimize through two primary objectives: contrastive learning and multi-task learning. Contrastive learning-based methods, such as SimCLS [10] and BRIO-Ctr [8], use margin-based losses to align candidate summaries with their quality, often measured by ROUGE scores [11]. In contrast, multi-task learning approaches like SummaReranker [9] and BRIO-Mul [8] combine multiple loss functions to enhance performance by leveraging both contrastive learning and cross-entropy-based optimization. Despite these advancements, existing re-ranking approaches face significant limitations [12,13]:

1. Overemphasis on Lexical Overlap: Current methods often prioritize ranking summaries based on lexical overlap, potentially neglecting deeper semantic content.
2. Inconsistent Quality Assessment: Many candidate summaries

☆ This work was made possible by NPRP13S-0112- 200037 grant from Qatar National Research Fund (a member of Qatar Foundation). The statements made here are solely the responsibility of the authors.

* Corresponding author.
E-mail addresses: ea1805307@qu.edu.qa (E. Aloraini), hozaifa@aitech.net.au (H. Kassab), ahamdi@msa.edu.eg (A. Hamdi), khaled.shaban@qu.edu.qa (K. Shaban).

https://doi.org/10.1016/j.neucom.2025.130816
Received 19 January 2025; Received in revised form 4 June 2025; Accepted 17 June 2025
Available online 2 July 2025

sharing identical ROUGE scores (see Fig. 1), yet previous studies train models to assign different ranks to these summaries, leading to inaccurate quality evaluations.

These limitations suggest that current abstractive models are not being exploited to their full potential, necessitating better methods for identifying the best summary candidate. To address these issues, we investigate whether it is possible to train a second-stage summarization model that learns to select the best summary among a set of candidates generated by a base model while balancing both lexical and semantic quality. Based on a two-stage framework, our model, named LexiSem Re-Ranker, is trained using multi-task learning. We directly incorporate ROUGE score differences into a ranking loss to preserve the lexical quality. Additionally, we use a contrastive loss with hard negative mining to identify summaries whose meanings are closely aligned with the original document. To improve contrastive learning, our approach divides candidate summaries into positive and negative samples based on their cosine similarity scores. Contributions:

1. We introduce a summary candidate ranker as a second-stage approach to abstractive summarization.
2. We demonstrate the effectiveness of our method on CNN/DM, XSum, Reddit TIFU and MeQSum datasets, achieving superior performance in ROUGE metrics on CNN/DM and Reddit TIFU and comparable results on XSUM.
3. We provide empirical evidence showing that the LexiSem method uses sentence embeddings to identify the most similar parts of the reference summary, thereby enhancing the efficiency of the framework by eliminating the need to be apply the metric pair-wise to every set of sentences in the reference and candidate summaries, which ultimately improves downstream evaluations such as summarization quality.
4. We explore how constructing positive and negative samples impacts the training of the ranker model within a contrastive learning framework.
5. In the hope of fostering research in abstractive summarization, we release the code of our ranker and our preprocessed data for our experiments on CNN-DM, XSUM, Reddit TIFU and MeQSum at https://github.com/MIRAH-Official/LexiSem

This study highlights the potential of re-ranking methods to enhance summarization quality, advancing the state-of-the-art in abstractive summarization research. Unlike previous models such as SimCLS or BRIO, LexiSem introduces a novel integration of proposition-level scoring and contrastive learning. It uniquely combines these elements with a hard negative mining strategy that distinguishes semantically close but factually inconsistent summaries, achieving a finer balance between lexical overlap and semantic fidelity. This architecture is particularly beneficial for improving factual consistency in abstractive summarization.

2. Background and related work

2.1. Two-step summarization

Recent advancements in abstractive summarization have led to significant improvements through the adoption of second-stage methods. These methods employ external models to align system outputs with evaluation metrics, facilitating the re-ranking of candidate summaries. This approach addresses the issue of "exposure bias" [7], which often arises from standard MLE training with teacher forcing. By re-ranking diverse candidate summaries, these methods enhance the faithfulness of summaries [14,15] or improve their relevance as measured by ROUGE scores [8,10,16]. Some approaches have also integrated ranking into the training process by incorporating contrastive loss alongside traditional MLE loss, achieving a multi-task learning objective [8,17]. While this work is related to ours, we specifically focus on re-ranking candidates to balance both lexical and semantic quality, ensuring closer alignment with reference summaries.

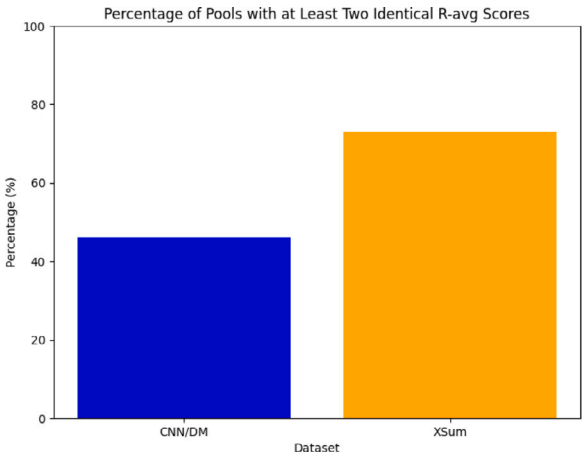


Fig. 1. Number of pools with at least two identical R-avg scores. A pool consists of 16 diverse beam search candidates generated on different datasets (CNN/DM, XSum) with different base models (PEGASUS, BART). R-avg is the average of ROUGE-1/2/L scores.

Several notable works have contributed to this area. GSum [18] introduced discrete guidance signals, such as salient sentences predicted by an extractive model, to better direct abstractive summarization systems. Unlike traditional abstractive models that optimize token-level MLE, second-stage methods generally operate at the sequence level. Models like ConSum [19] and SeqCo [20] fine-tuned their base models using contrastive loss, aiming to boost the confidence in higher-quality summary candidates. RefSum [21] adopted a meta-learning framework to re-rank summaries generated by multiple base systems.

Re-ranking methods such as SummaReranker [9] and SimCLS [10] employed RoBERTa-based models for re-ranking. SummaReranker utilized a multi-label binary cross-entropy loss, while SimCLS leveraged contrastive learning and a ranking loss. BRIO [8] enhanced performance by fine-tuning the base model a second time, combining cross-entropy loss and candidate-level ranking loss. PGA [13], another two-step method, employed an autoregressive model for plan generation, followed by plan-guided abstraction and re-ranking using the BRIO method.

These approaches collectively demonstrate the effectiveness of two-stage summarization strategies, providing more faithful and relevant summaries through re-ranking mechanisms and multi-task learning objectives.

2.2. Re-ranking

Re-ranking has been extensively studied in various branches of NLP and applied to conditional generation tasks for many years [22–24]. In syntactic parsing, Collins and Koo [25] pioneered re-ranking for outputs of a base parser. Charniak and Johnson [26] advanced this approach with a Maximum Entropy re-ranker. Passage re-ranking plays a crucial role in question-answering systems, aiding in retrieving relevant passages that potentially contain the answer [27,28]. Recent question-answering models have also integrated answer re-ranking techniques to improve answer selection precision [29].

In neural machine translation, checkpoint re-ranking [30] generated multiple translation candidates from different model checkpoints, leveraging the observation that the oracle across these checkpoints often outperforms the final checkpoint. Bhattacharyya et al. [31] employed an energy-based model on top of BERT to select translation candidates with higher BLEU scores.

In abstractive summarization, second-stage approaches like re-ranking remain relatively underexplored. RefSum [21] introduced a

second-stage framework to mitigate the train–test distribution mismatch observed in such models. Using the base GSum model [18], the authors achieved a state-of-the-art ROUGE-1 score of 46.18 on the CNN/DM dataset. SimCLS [10] trained a second-stage model using contrastive learning and a ranking loss to select the best summary from a pool of 16 diverse beam search candidates, reaching a ROUGE-1 score of 46.67 on CNN/DM. This approach underscores the potential of large pre-trained sequence-to-sequence (Seq2Seq) models as quality estimation mechanisms [32].

2.3. Contrastive learning

As a fundamental approach in representation learning, contrastive learning has been widely recognized as an effective method to enable models to differentiate the quality of various samples [33,34]. Recently, it has shown promising performance in natural language generation tasks such as text summarization [35] and machine translation [36,37]. Contrastive examples (or "negative" samples) can be crafted through either rule-based or model-based methods, with the latter typically producing outputs that are closer to human-generated text and therefore more natural for contrastive schemes. In addition, contrastive learning can be applied in latent space or in discrete space. For example, Gao et al. [38] propose a contrastive framework to improve sentence embeddings, achieving state-of-the-art results in representation quality. On the other hand, Liu et al. [39] perform discriminative re-ranking on generated summaries in a discrete space, following a line of research that leverages contrastive methods for re-ranking outputs [40–43]. More recently, several works have integrated contrastive learning directly into text summarization. Fang et al. [44] show that contrastive learning improves language models by strengthening their representation capabilities. Similarly, Cao and Wang [35] propose a novel formulation of contrastive learning based on reference summaries, creating negative examples through deletion, replacement, rearrangement, and hallucinations. Their approach helps the summarization model better distinguish low-quality or factually incorrect summaries. In a related effort, Wan and Bansal [45] generate negative samples by applying rule-based transformations (e.g., content replacement and sentence negation) to the source document; contrastive learning then assists the model in distinguishing factual summaries from hallucinated ones. Such advances underscore the versatility and efficacy of contrastive learning for improving both the fidelity and coherence of generated text. Kwon at al.[46] propose a contrastive attention mechanism for summarizing implicit datasets, incorporating conventional and adversarial attention to better distinguish important information. Hard negative sampling plays a crucial role in contrastive learning, influencing the effectiveness of learned representations. A key challenge is selecting negative samples that provide meaningful contrast to the anchor while avoiding false negatives. Robinson et al. [47] propose two guiding principles for effective hard negative sampling: (1) negative samples should be "true negatives", meaning they belong to a different class or have distinct labels from the anchor; and (2) the most informative negatives are those that the model currently considers similar to the anchor in the embedding space. This ensures that contrastive learning benefits from samples that challenge the model's current understanding, providing strong gradient signals during training. In metric learning, the availability of labeled negative pairs inherently satisfies the first principle, while the second principle emphasizes the need for negatives that are difficult to distinguish under the current representation. Drawing inspiration from these existing studies, we propose a contrastive learning-based method to score and rerank candidate summaries using the hard negative sampling to construct positive and negative samples and the formulation of the contrastive loss function.

Fig. 2. Example of candidate summary propositions/segmentation.

2.4. Exposure bias

Exposure bias is a well-documented issue in seq-2-seq models for abstractive summarization. It arises due to discrepancies between the training and inference phases. During training, models are typically optimized using MLE with teacher forcing, where the ground truth tokens are always provided as input at each step. However, during inference, the model generates tokens based on its own predictions, which may accumulate errors and propagate them to subsequent steps. This mismatch between training and inference conditions often degrades the quality of the generated summaries.

To address this, several strategies have been proposed. Bengio et al. [6] introduced scheduled sampling, a training method, where the model gradually transitions from using ground truth tokens to relying on its own generated tokens during training. This strategy helps the model become more robust to its own errors. Reinforcement learning (RL)-based approaches have also been explored to mitigate exposure bias. For example, Paulus et al. [48] and Bahdanau et al. [49] proposed optimizing summarization models by maximizing task-specific rewards, such as ROUGE scores, which are non-differentiable and cannot be directly optimized through MLE. By using RL, these models learn to produce summaries that better align with evaluation metrics, thus reducing the adverse effects of exposure bias. While these techniques have shown promise, they often introduce challenges such as increased training complexity and sensitivity to hyperparameters. Recent research continues to explore alternative methods, including contrastive learning frameworks and hybrid loss functions, to further mitigate exposure bias and improve the robustness of summarization models during inference.

3. Method

Our method follows a two-stage framework. Given a source document D, a base model B, and a function g, the first stage generates a pool of m candidate summaries $C = \{C_1, C_2, \ldots, C_m\}$:

$$\mathbf{C} \leftarrow g(D) \tag{1}$$

In the second stage, a scoring function $CosScore$ assigns a similarity score to each candidate summary. The best summary $C*$ is selected as the one with the highest score:

$$\mathbf{C}^* = \arg\max_{C_i \in C} \{CosScore(C_i, S)\} \tag{2}$$

The goal is to train the ranking model $CosScore$ to identify the most accurate summary from the candidate pool generated by g.

3.1. Model architecture

The proposed LexiSem re-ranker (Fig. 3) evaluates candidate summaries at the proposition level using sentence embeddings computed via SimCSE/BERT-base [50], which measures cosine similarity scores between propositions in the reference summary and candidate summaries. It then ranks them using a RoBERTa-base backbone neural network.

Inspired by [12], our approach focuses on capturing rich semantic units at the level of sentence propositions. The LexiSem re-ranker

evaluates each sentence in a predicted candidate summary against the most similar sentence in the reference summary, using both cosine similarity and ROUGE metrics. This ensures fine-grained evaluation at the proposition level while maintaining computational efficiency through sentence embeddings.

For each candidate summary C_i, the individual similarity score is calculated as follows:

$$\text{Similarity}(E(C_i), E(S)) = CosineSimilarity(\text{seg}_m, \text{seg}_n) \tag{3}$$

where $E(C_i) = \text{seg}_m$ and $E(S) = \text{seg}_n$ are the embeddings of propositions (or segments) in the candidate summary C_i and the reference summary S, respectively.

To compute $CosScore(C_i, S)$, both the reference summary $S = \langle s_i, i \in I \rangle$ and each generated candidate summary $C^* = \langle C_j, j \in J \rangle$ are split into lists of propositions, i.e., self-contained atomic units of meaning within sentences [51], as shown in Fig. 2. Splitting sentences into propositions has been shown to be effective in prior works [12, 52,53]. Embeddings are generated for these propositions using the SimCSE/Roberta-large model [50]. For each sentence in the candidate summary, the cosine similarity is calculated between its proposition embeddings and those of the reference summary. The overall score is then computed as:

$$\text{CosScore} = \text{avrg max similarity}((E(C_i), E(S))) \tag{4}$$

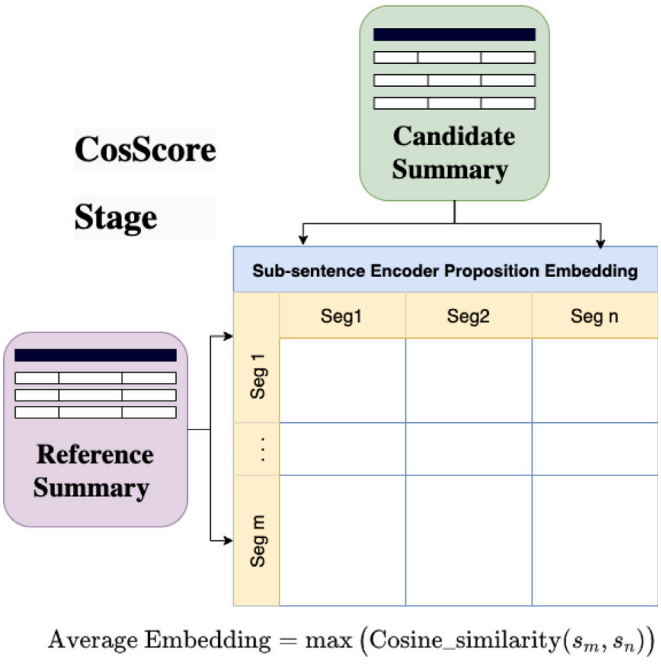


Fig. 3. LexiSem Re-Ranker model architecture.

3.2. Training objective

Ranking Loss The ranking loss is designed to ensure that the higher-quality candidate summaries receive higher scores. The loss function is defined as:

$$\mathcal{L}_{rank} = \sum_i \sum_{j>i} \max\big(0,\ \text{CosScore}(C_j, S) - \text{CosScore}(C_i, S)\big) + \Big(\overline{\text{R}}(C_i, S)\ +\ \overline{\text{R}}(C_j, S)\Big) \times \lambda \tag{5}$$

Here:

- $\overline{R}$ = 1- ROUGE, penalizes deviations from the reference summary.
- λ is a hyperparameter[1] that balances the ranking loss and the ROUGE penalty. This approach addresses inconsistencies in prior methods [8,10], where identical ROUGE scores for different summaries led to varying evaluation margins.

Contrastive Loss The contrastive loss is designed to differentiate between positive and negative candidates effectively. The construction of positive and negative pairs is the critical point in contrastive learning. As shown in Fig. 3, we computing the cosine similarity scores using the SimCSE model for the 16 candidate summaries generated from each document. Based on the resulting similarity scores:

- The top 10 summaries with the highest cosine similarity scores are selected as positive candidates.
- The remaining 6 summaries are treated as negative samples.
- Additionally, to further diversify the negative samples, 4 random irrelevant summaries from other documents are added.

How This Negative-Sample Strategy Improves Re-Ranking:

The effectiveness of contrastive learning hinges on how well positive and negative examples capture subtle variations in summary quality. By selecting the top 10 most semantically similar summaries as positives (via SimCSE) and the bottom 6 as negatives, the model is compelled to discriminate between *closely matched* candidates that might still harbor factual or structural errors. This "hard negative mining" ensures the model does not overlook small but critical discrepancies in otherwise plausible-looking summaries. Meanwhile, adding 4 *irrelevant* summaries from entirely different documents injects additional diversity among the negatives, strengthening the model's ability to reject off-topic or semantically distant outputs.

Together, these two types of negatives (hard negatives vs. irrelevant negatives) enable a more balanced and robust contrastive signal. The model learns to:

- **Refine boundary decisions among highly similar yet lower-quality summaries,** improving its sensitivity to subtle errors or omissions.
- **Generalize to dissimilar, off-topic candidates,** thereby making it less prone to trivializing vastly different summaries as simply "bad" without understanding their content mismatch.

As a result, the re-ranker's learned scoring function better captures minor factual or semantic inconsistencies across a broad spectrum of candidate summaries ranging from near correct paraphrases to completely off-topic text. This comprehensive separation of signals pushes the model to pinpoint the truly *best* summary more effectively, thereby enhancing final re-ranking performance [47,50]. This strategy ensures robust contrastive learning by leveraging hard negative mining [47,50]. Thus, we design a set of candidate summaries C in Eq. (1) as positive and a set of randomly sampled summaries N as negative. To identify summaries whose meanings are close to the document the contrastive loss is defined as:

$$\mathcal{L}_{ctr} = \frac{1}{|C|} \sum_{C_i \in C} -\log \frac{e^{\text{CosScore}(C_i^+, S)}}{e^{\text{CosScore}(C_i^+, S)} + \sum_{C_j^- \in \mathcal{N}} e^{\text{CosScore}(C_j^-, S)}} \tag{6}$$

Combined Objective The final objective combines the ranking loss and contrastive loss:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{rank} + \gamma_2 \mathcal{L}_{ctr} \tag{7}$$

where γ_1 and γ_2 are hyperparameters balancing the two components of the loss.[2]

[1] We set λ to 1.0 on CNN/DM and 0.1 on XSum.

[2] we set $\gamma_1 = 10$ and $\gamma_2 = 0.1$.

4. Experiments

4.1. Datasets

To demonstrate the effectiveness of the proposed model, experiments were conducted using five diverse benchmark datasets with different topics, lengths, and abstractiveness. These datasets provide diverse summarization challenges and allow testing the performance of our LexiSem Re-Ranker on both long-form and single-sentence summaries. The use of propositions as distinct semantic units of meaning in text is inspired by recent studies [51,53,54], which demonstrate their success in representing and evaluating text semantics at a fine-grained level. By breaking text into propositions, the overall semantics of the document can be effectively captured and evaluated.

- **CNN/DM**: This dataset [55] is widely used for news summarization tasks. News articles serve as source documents, and their corresponding highlights act as reference summaries. It contains 287,226 articles for training, 13,368 for validation, and 11,490 for testing. On average, documents 791.6 words long, while summaries are 55.6 words.
- **XSum**: This dataset [56] focuses on highly abstractive, one-sentence summaries sourced from BBC articles. It includes 204,045 articles for training, 11,332 for validation, and 11,334 for testing. The average document length is 429.2 words, and summaries average 23.3 words.
- **Reddit TIFU**: This dataset [57]contains 120k posts from the popular online Reddit forum. As in other summarization works [2], we use the TIFU-long subset, containing 37k posts. As there is no official split, we build a random 80:10:10 split for training:validation:test.
- **MeQSum**: Is a medical question summarization dataset [58]. The summaries in MeqSum are written by medical experts in a formal style. It include 400 question/summary pair for training, 100 for validation and 500 for testing.

4.2. Data preparation: Prop-CNN/DM, prop-XSUM datasets

To train our re-ranker model, we created new proposition-level datasets, namely Prop-CNN/DM and Prop-XSUM, by segmenting summaries from the original CNN/DM and XSum datasets using the SegmenT5-large model [51]. This model segments sentences into sub-sentences or *propositions*, each representing a fine-grained semantic unit. The resulting propositions are concatenated using the special token [SEP] to form proposition-based training sequences. To the best of our knowledge, this is the first reranking-based summarization approach to explicitly operate on proposition-level representations. These datasets enable enhanced semantic granularity, which is crucial for downstream tasks such as summarization, question answering, and factual consistency evaluation. An illustration of candidate summary propositions is shown in Fig. 4. In contrast, we used the original forms of Reddit TIFU, and MeQSum without applying proposition segmentation. This decision was made due to the distinct nature of these datasets: Reddit TIFU summaries are typically informal and user-generated, and MeQSum is constructed from medical QA contexts. All three datasets are characterized by relatively short sentences or utterances that do not benefit meaningfully from further segmentation. Thus, applying proposition segmentation to these datasets was deemed unnecessary and potentially disruptive to their intrinsic discourse structure.

4.3. Implementation details

4.3.1. Model

We implement our model using Huggingface Transformers library [59], utilizing the pre-trained RoBERTa-base model (125M parameters). Experiments are conducted on 2 NVIDIA RTX 3090 GPUs (24 GB memory), and propositions/segmentation process is executed on a Google Cloud Platform instance with 30 GB of RAM and 4 NVIDIA T4 GPUs. We use the PyTorch framework [12] to build our solution. The entire training process takes 33 h on CNN/DM, 22 h on XSum, 16 h for Reddit TIFU and 14 h MeQSum.

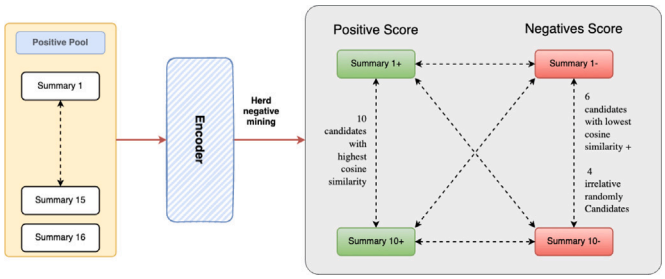


Fig. 4. Overview of the proposed training objective.

4.3.2. Decoding settings

Summaries are decoded using the diverse beam search algorithm [60], generating 16 candidate summaries across 16 diversity groups. We use pre-trained BART-large[3] for CNN/DM and PEGASUS-large[4][5] models for XSum and Reddit TIFU, respectively, as the generation models.

4.3.3. Training settings

The models are trained for 5 epochs using the Adafactor optimizer [61], with a batch size of 4 and a learning rate of 2e−3. Validation is performed every 1000 steps.

4.4. Baseline

Our approach is compatible with any seq-2-seq-based abstractive summarization model. We generate 16 candidates using diverse beam search [60], employing pre-trained and fine-tuned versions of PEGASUS [2] and BART [3] from the Huggingface Transformers library [59].

- **BART**: A Transformer-based pre-training model that uses a bidirectional encoder for input encoding and a left-to-right decoder for summary generation, optimized with a denoising objective.
- **PEGASUS**: A Transformer model trained with a self-supervised objective that masks or removes important sentences from the input document. This enables the model to generate output summaries that effectively capture global sentence-level information.

We compare our results to the following re-ranking approaches:

1. **SimCLS** [10]: Uses a RoBERTa classifier for re-ranking candidate summaries based on a ranking loss.
2. **BRIO** [8]: Calibrates model likelihoods to ROUGE rankings, using multi-task learning.
3. **SummaReranker** [9]: Trains a RoBERTa-based classifier on up to 60 candidates, generated through multiple decoding methods like beam search and top-k sampling.
4. **BalSum** [12]: Focuses on balancing lexical overlap and semantic preservation using joint optimization of ROUGE and cosine similarity.
5. **GAR** [62]: Implements guided abstractive re-ranking to achieve semantic fidelity.

[3] The checkpoint is "facebook/bart-large-cnn" containing around 406M parameters, whose maximum encoding length is 1024.

[4] The checkpoint is "google/pegasus-xsum"fine-tuned with XSum containing around 568M parameters, whose maximum encoding length is 512.

[5] PEGASUS-large fine-tuned on Reddit TIFU from [9].

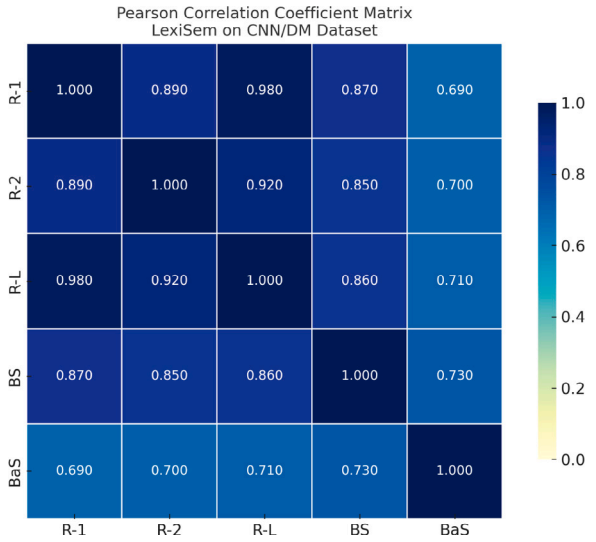


Fig. 5. Pearson correlation coefficient between the five evaluation metrics R-1, R-2, R-L, BS, BaS for a base Bart with beam search on CNN/DM. R-1/2/L denotes ROUGE-1/2/L, BS and BaS denote BERTScore and BARTScore.

4.5. Evaluation metrics

As a comprehensive evaluation suite, we employ the standard ROUGE metrics—ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) to assess lexical overlap, along with two neural evaluation metrics. BERTScore (BS) and BARTScore (BaS), which leverage contextual embeddings from large pre-trained language models. As shown in Fig. 5, we observe a very high correlation between R-1 and R-L (Pearson coefficient of 0.980), indicating that these two ROUGE variants often yield similar rankings of candidate summaries. Likewise, R-2 is also strongly correlated with both R-1 (0.890) and R-L (0.920), suggesting that all three ROUGE metrics tend to agree on the lexical quality of summaries. Neural metrics show moderate to high correlation with ROUGE. Specifically, BERTScore achieves strong alignment with R-1 (0.870) and R-L (0.860), while BARTScore shows relatively lower correlations, with values ranging from 0.690 (with R-1) to 0.730 (with BS). This indicates that BARTScore captures complementary aspects of summary quality not fully aligned with lexical overlap. The relatively lower correlation between BaS and other metrics further supports its role as a semantically sensitive evaluation signal that goes beyond surface-level word overlap. These findings reinforce the importance of combining lexical and semantic evaluation signals in summary quality assessment, and motivate our re-ranker's design which explicitly balances both dimensions.

- **ROUGE-N** [11]: This metric measures lexical overlap between generated and reference summaries. It focuses on unigrams (ROUGE-1), bigrams (ROUGE-2), and longest common subsequences (ROUGE-L).
- **BERTScore** [63]: This metric evaluates semantic similarity using contextual embeddings from pre-trained transformer models like BERT or RoBERTa. It is more robust to tokenization and casing issues compared to ROUGE and correlates better with human judgments.
- **BARTScore** [32]: A learned evaluation metric based on the BART model, which estimates the probability of generating the reference summary given the generated summary (or vice versa). BARTScore captures both fluency and factual consistency in summarization.

ROUGE-N Metric

ROUGE calculates the overlap of N-grams between the generated summaries and reference summaries, capturing lexical similarity. The formula for ROUGE-N is:

$$\text{ROUGE}_N = \frac{\sum_{S \in \text{ReferenceSum}} \sum_{\text{gram}_n \in S} \text{CountMatch}(\text{gram}_n)}{\sum_{S \in \text{ReferenceSum}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{8}$$

For ROUGE-L, which measures the longest common subsequence (LCS) between a generated summary S and a reference summary S^*, the following metrics are used:

- **Recall**:

$$R_{\text{LCS}} = \frac{\text{LCS}(S, S^*)}{N} \tag{9}$$

- **Precision**:

$$P_{\text{LCS}} = \frac{\text{LCS}(S, S^*)}{N'} \tag{10}$$

- **F1 Score**:

$$F_{\text{LCS}} = \frac{(1+\beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}} \tag{11}$$

Here, N is the length of the reference summary, N_0 is the length of the generated summary, and $\text{LCS}(S, S^*)$ denotes the length of the longest common subsequence. β is a weighting hyperparameter, often set to 1 for equal importance of recall and precision. We use the standard PERL ROUGE[6] script for ROUGE scoring, employing PTB tokenization and lowercasing, following [8].

BERTScore Metric

BERTScore measures the semantic similarity between generated and reference summaries using high-dimensional embeddings. It calculates precision, recall, and F1 scores based on the cosine similarity between token embeddings.

Each token in the generated summary S and the reference summary S^* is embedded into a vector space using a pre-trained model (e.g., RoBERTa-large). For token x_i in S, the embedding is represented as $\mathbf{E}(x_i)$.

- **Cosine Similarity**:

$$\text{sim}(x_i, x_j) = \frac{\mathbf{E}(x_i) \cdot \mathbf{E}(x_j)}{\|\mathbf{E}(x_i)\| \|\mathbf{E}(x_j)\|}$$

Using these similarities, BERTScore calculates:

- **Precision**: The average maximum similarity between tokens in the generated summary and the reference summary:

$$\text{Precision} = \frac{1}{|S|} \sum_{x_i \in S} \max_{x_j \in S^*} \text{sim}(x_i, x_j)$$

- **Recall**: The average maximum similarity between tokens in the reference summary and the generated summary:

$$\text{Recall} = \frac{1}{|S^*|} \sum_{x_j \in S^*} \max_{x_i \in S} \text{sim}(x_j, x_i)$$

- **F1 Score**: The harmonic mean of precision and recall:

$$F1_{\text{BERTScore}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

[6] The command parameters are '-c 95 -m -r 1000 -n'. https://github.com/summanlp/evaluation/tree/master/ROUGE-RELEASE-1.5.5.

For evaluation, we use the public bert-score Python package[7] with standard settings. This implementation ensures consistency in tokenization and special token handling.

BARTScore Metric

BARTScore is a learned evaluation metric based on the BART model, a denoising autoencoder that can generate text by reconstructing corrupted input. Instead of relying on explicit word matching (like ROUGE) or cosine similarity (like BERTScore), BARTScore estimates the probability of generating a reference summary given a generated summary (or vice versa), capturing both fluency and factual consistency.

The BARTScore formulation is given as:

$$\text{BARTScore}(S, S^*) = \frac{1}{|S^*|} \sum_{t=1}^{|S^*|} \log P(S_t^* | S_{<t}^*, S) \tag{12}$$

where S^* is the reference summary, S is the generated summary, and $P(S_t^* | S_{<t}^*, S)$ is the probability of the tth token in the reference summary conditioned on its previous tokens and the generated summary.

BARTScore can be used in different modes:

- **Reference-based**: Evaluates how well a generated summary reconstructs a reference summary.
- **Reference-free**: Scores summaries without needing a reference, assessing coherence and factual consistency.
- **Reverse mode**: Measures how likely a reference summary could generate the given summary, useful for factuality assessment.

For evaluation, we use the official BARTScore implementation[8] with pre-trained BART-large settings.

4.6. Results and analysis

Traditional re-ranking and generation-based summarization models predominantly rely on lexical overlap metrics such as ROUGE. While effective in capturing surface-level similarity, these metrics fall short when evaluating abstractive summaries that employ paraphrasing or semantically equivalent phrasing. To address this limitation, our model incorporates proposition-level cosine similarity using sentence embeddings from SimCSE/BERT. This allows for a more nuanced assessment of semantic similarity at the sub-sentence level, identifying the candidate most aligned in meaning with the reference summary. The motivation for this approach is rooted in prior work [12,51,64] demonstrating that proposition-level embeddings can capture fine-grained semantic units more reliably than full-sentence or document-level embeddings. Moreover, in contrastive learning settings, the use of hard negative sampling based on semantic similarity provides more informative gradients, allowing the model to distinguish between subtle factual differences across summaries. Within the two-stage framework, we compare our results to several established re-ranking methods, including SimCLS [10], SummaReranker [9], BRIO [8], BalSum [12] and GAR [62]. We apply the LexiSem re-ranker on top of both BART and PEGASUS base models.

CNN/DM Results:

Table 1 presents our results on the CNN/DM dataset. LexiSem outperforms all competing models, achieving ROUGE-1/2/L scores of 48.18/24.46/45.05, thereby setting a new state-of-the-art. Notably, it also obtains a BERTScore of 89.2 and −3.03 BartScore demonstrating that our model capture semantic meaning with lexical quality. We attribute these improvements to LexiSem's ability to more accurately estimate the underlying semantics of candidate summaries compared to existing re-ranking approaches. We present a case study in Section 5.1.

XSum Results:

[7] https://github.com/Tiiiger/bert_score.

[8] https://github.com/neulab/BARTScore.

Table 1
Results on CNN/DM. R-1/2/L are ROUGE-1/2/L F_1 scores. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore *: results reported in the original papers.†: results from our own evaluation script.†: significantly better than all other models. The best results are bolded.

Model	R-1	R-2	R-L	BS	BaS
BART*	44.16	21.28	40.90	87.95	−3.91
Pegasus*	44.17	21.47	41.11	88.13	−3.83
BRIO-Mul*	47.78	23.55	44.57	–	–
BRIO-Mul†	47.50	23.48	44.01	89.08	–
BRIO-Ctr*	47.28	22.93	44.15	–	–
BRIO-Ctr†	46.08	22.03	43.06	89.03	–
SummaReranker*	47.16	22.55	43.87	87.74	**−2.18**
SimCLS*	46.67	22.15	43.54	66.14	–
BalSum†	46.58	22.33	43.49	89.67	–
GAR*	47.58	23.96	44.36	**89.95**	–
LexiSem (ours)	**48.18**†	**24.46**†	**45.05**†	89.21	−3.03

We also evaluated our method on XSum dataset Table 2. Here, LexiSem delivers a modest but notable improvement over the baseline PEGASUS model. We believe that the relatively smaller gains on XSum are due to the dataset's highly abstractive nature and it shorter, single-sentence reference summaries, which limit the diversity of semantic units that can be leveraged compared to longer summaries in CNN/DM.

Reddit TIFU Results:

Table 3 presents our results on the Reddit TIFU dataset. LexiSem outperforms all competing models, achieving ROUGE-1/L scores of 30.37/23.87, setting a new benchmark for this dataset. Notably, it maintains a R2 of 9.32,BERTScore of 87.25 and a BARTScore of −3.52, demonstrating its ability to balance lexical quality and semantic coherence. We attribute these improvements to LexiSem's ability to effectively integrate lexical and semantic features, leading to more informative and structurally coherent summaries compared to existing re-ranking approaches.

MeQSum Results:

LexiSem was also evaluated on the MeQSum dataset (Table 4). While the model showed slightly lower ROUGE and BERTScore metrics compared to the BART baseline, it maintained strong semantic alignment, underscoring its capability to capture the core meaning of medical question summaries. These results highlight that LexiSem's strength lies in enhancing semantic depth and abstraction, even in highly specialized domains where lexical precision is essential. The findings suggest promising potential for LexiSem to serve as a robust foundation that could be further optimized with domain-specific adaptations, paving the way for future improvements in medical and technical summarization tasks.

MeQSum Results:

LexiSem was also evaluated on the MeQSum dataset (Table 4). While the model showed slightly lower ROUGE and BERTScore metrics compared to the BART baseline, it maintained strong semantic alignment, underscoring its capability to capture the core meaning of medical question summaries. These results highlight that LexiSem's strength lies in enhancing semantic depth and abstraction, even in highly specialized domains where lexical precision is essential. The findings suggest promising potential for LexiSem to serve as a robust foundation that could be further optimized with domain-specific adaptations, paving the way for future improvements in medical and technical summarization tasks. To mitigate the domain-specific performance drop, future work could explore domain-adaptive fine-tuning strategies or the use of biomedical-specific embeddings such as BioBERT [65]. These models are trained on domain-relevant corpora and may enhance LexiSem's ability to better capture terminology and semantic nuance in medical QA settings.

Table 2
Results on XSum. R-1/2/L are ROUGE-1/2/L F_1 scores. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore *: results reported in the original papers. †: results from our own evaluation script. †: LexiSem shows significant improvements over the baseline models.

Model	R-1	R-2	R-L	BS	BaS
BART*	45.14	22.27	37.25	88.17	−4.05
Pegasus*	47.21	24.56	39.25	–	−3.89
BRIO-Mul*	**49.07**	**25.59**	**40.40**	–	–
BRIO-Mul†	48.74	25.38	40.16	92.60	–
BRIO-Ctr*	48.13	25.13	39.84	–	–
BRIO-Ctr†	48.12	25.24	39.96	91.72	–
SummaReranker*	48.12	24.95	40.00	92.14	**−1.90**
SimCLS*	47.61	24.57	39.44	–	–
BalSum††	46.17	23.23	38.09	91.48	–
GAR*	48.93	25.45	40.39	92.67	–
LexiSem (ours)	48.61 †	25.14 †	39.75 †	91.77	−3.08

Table 3
Results on Reddit TIFU. R-1/2/L are ROUGE-1/2/L F_1 scores. BS and BaS refer to the neural model-based metrics BERTScore and BARTScore *: results reported in the original papers. SR refers to SummaReranker †: significantly better than the baseline model. The best results are bolded.

Model	R-1	R-2	R-L	BS	BaS
PEGASUS*	26.63	9.01	21.60	–	–
BART (SR setup)	27.42	**9.53**	22.10	87.43	−3.78
SummaReranker*	29.83	9.50	23.47	87.81	**−3.33**
LexiSem (ours)	**30.37**	9.32†	**23.87**†	87.25	−3.52

Table 4
Average results on MeQSum test set. R-1/2/L is the ROUGE-1/2/L F1 score. BS and BaS are the neural metrics BERTScore and BARTScore respectively. The best results are bolded.

Model	R-1	R-2	R-L	BS	BaS
BART	**46.17**	**29.50**	**44.80**	**87.23**	**−3.43**
LexiSem (ours)	40.52	26.54	37.4	86.97	−4.49

This comprehensive evaluation demonstrates that the LexiSem re-ranker, by balancing lexical and semantic quality through innovative use of proposition-level embeddings and robust contrastive learning, effectively enhances summarization performance across diverse datasets.

In summary, our re-ranking approach introduces necessary complexity to resolve key limitations in existing methods, delivering substantial improvements across both lexical and semantic metrics. The trade-off between computational cost and performance gain is justified, particularly in applications where factual and semantic faithfulness is critical.

5. Ablation study

We examine different model configurations of LexiSem by selectively removing key components and evaluating performance on the CNN/DM test set. As shown in Table 5, removing the `CosScore` component markedly decreases ROUGE-1/2/L to 46.7/22.5/43.7, underscoring the importance of semantic alignment for high-quality summaries. Additionally, we observe that removing contrastive loss primarily affects ROUGE-L, causing a substantial drop in capturing the longest common subsequence with the reference (from 45.05 down to 39.99). This suggests that contrastive learning is instrumental in encouraging the model to retain global coherence and structure across the entire summary, not just local n-gram overlaps. Together, these findings emphasize that both the semantic scoring (`CosScore`) and the contrastive objective are critical to achieving robust, high-fidelity summaries.

Table 5
Ablation study of different model architectures on CNN/DM. R-1/2/L denotes ROUGE-1/2/L.

Model	R-1	R-2	R-L
Without CosScore	46.7	22.5	43.7
Without contrastive loss	48.2	23.2	39.99

5.1. Case study on CNNDM comparing to baseline models

As Tables 6–7–8 show, LexiSem improves the quality of summaries qualitatively by focusing on capturing detailed and relevant information from the original text, making the summaries richer and more informative. We investigate whether all summaries ranked by models satisfy both the lexical and semantic quality. We evaluated models using F1 which measures the cases where the higher-ranked summary has both larger ROUGE and BERTScore than the lower-ranked summary (see table. 8–6–7). They show an examples of summary selected by LexiSem model, along-side its ground-truth (reference) summary compared with BRIO and BalSum models. Here's how LexiSem enhances summary quality:

1. **Content Fidelity and Coverage**: LexiSem includes key details and context that other models may miss, leading to more comprehensive summaries. For example, in the first summary on DIY brain stimulation kits, LexiSem includes specific details, such as mentioning the use of brain stimulation devices at Oxford University to improve speech in individuals with production problems. This level of specificity adds depth and reliability to the summary.
2. **Clarity and Precision**: LexiSem tends to avoid vague language, focusing on precise terms that enhance clarity. For instance, in the loneliness maps case, LexiSem describes the purpose of these maps in clear terms—identifying areas with people at risk of social isolation. This clarity reduces ambiguity, making the summaries more accessible and understandable to readers.
3. **Contextual Relevance**: By providing relevant context, LexiSem creates summaries that maintain the core themes of the original text. In the lizard case, for instance, LexiSem not only describes the swallowing process but also includes extra details about the goanna's feeding adaptations, such as its ability to unhinge its lower jaw, making the summary feel complete and well-rounded.
4. **High ROUGE and BERTScore Metrics**: LexiSem achieves higher scores in ROUGE and comparable BERTScore, indicating that it captures both lexical and semantic similarity with the reference summaries more effectively. These high scores suggest LexiSem's summaries align well with human-written references, both in word choice and meaning [66].

Overall, LexiSem's summaries tend to be detailed, contextually rich, and more aligned with the original text's intent, which collectively improve their qualitative quality compared to other summarization models.

6. Case study on CNN/DM comparing to base BART model

Table 9 presents an interesting pattern we observed when comparing the results of BART and LexiSem. The results demonstrate that LexiSem helps the abstractive model to better align with the reference summaries while filtering out irrelevant or misleading information.

Specifically, we noticed that some of the reference summaries contain crucial contextual details that BART tends to omit or distort. For example, in the first case, BART introduces additional reasoning about "becoming as good as Messi or Ronaldo", which is completely irrelevant to the original article. In contrast, LexiSem maintains the original focus of the summary while capturing key details more accurately.

Table 6
Case study on CNN/DM: Brain Stimulation Kits. R-1/2/L are the ROUGE-1/2/L F_1 scores. BS denotes BERTScore. Irrelevant information is highlighted in blue, while content more consistent with the reference is highlighted in green.

System	R-1	R-2	R-L	BS	Summary
Reference	–	–	–	–	Experts have voiced concerns over DIY brain stimulation kits for children. For a few hundred dollars, one can be purchased online from various sites. It promises to help children with homework and claims to help with ADHD. Professor Colleen Loo from the Black Dog Institute strongly believes this equipment poses a danger to amateurs and children. The equipment is currently being used to treat people with speech impediments but is still very much in trial stages.
BRIO-Ctr	40.10	16.26	19.20	87.11	For a few hundred dollars, you can purchase a brain stimulation kit online. Experts have voiced concerns over the potential side effects. The kits are being sold online for as little as 55 US dollars. One site even advertises how to make your own electrodes using a household sponge.
BalSum	36.92	17.19	27.69	89.90	Parents are buying DIY brain stimulation kits for their kids. The kits are being sold online for as little as $55 US. Experts are concerned about the potential side effects of the equipment. The devices are used to improve speaking in those with speech problems. The equipment is still relatively new and experimental.
LexiSem	56.73	37.41	51.06	90.07	Experts have voiced concerns over DIY brain stimulation kits for children. For a few hundred dollars, one can be purchased online from various sites and delivered to your door. The devices stimulate the brain by applying electrical currents to a small region of the scalp. It is currently being used at Oxford University to improve speaking in those who suffer from speech production problems.

Table 7
Case study on CNN/DM: Loneliness Maps. R-1/2/L are the ROUGE-1/2/L F_1 scores. BS denotes BERTScore. Irrelevant information is highlighted in blue, while content more consistent with the reference is highlighted in green.

System	R-1	R-2	R-L	BS	Summary
Reference	–	–	–	–	Local councils are urged to draw up maps of the residents who are at risk. Essex and Gloucestershire have already made 'loneliness maps'. Experts warn that being lonely can lead to serious health problems.
BRIO-Ctr	50.57	28.24	29.89	90.30	Two county councils have already implemented 'loneliness maps' to target 'danger zones'. Being lonely can lead to health problems including dementia and high blood pressure. Campaigners say councils should draw up maps of the places where pensioners are most at risk. Study by University of Kent and Campaign to End Loneliness recommends maps.
BalSum	50.0	27.91	43.18	91.82	Campaigners say councils should draw up maps of places where pensioners and others are most likely to suffer from social isolation. Two county councils, Essex and Gloucestershire, have already implemented the maps. They allow them to target 'danger zones' of loneliness. Being lonely can lead to health problems including dementia and high blood pressure.
LexiSem	49.38	25.31	34.56	91.50	'Loneliness maps' are being used to identify areas where people are most at risk. Campaigners say councils should draw up maps of the places where pensioners and others are most likely to suffer from social isolation. Two county councils, Essex and Gloucestershire, have already implemented the maps.

Another notable pattern is the inclusion of "click here" in the reference summaries (which appeared in 331 out of 11 490 instances in CNN/DM). BART learned this pattern from the training data and generated this phrase in multiple output summaries. However, LexiSem correctly identified this as a noise pattern demonstrating a more refined ability to discard uninformative content (see Table 9).

7. Conclusion and future work

In this study, we introduce LexiSem, a novel re-ranking approach designed to balance both lexical and semantic quality in abstractive summarization. By leveraging a multi-task learning framework, LexiSem optimizes for lexical overlap metrics while simultaneously evaluating the semantic alignment between candidate and reference summaries. To the best of our knowledge, this is the first work to incorporate a new perspective on contrastive learning through hard negative mining within a summarization re-ranking context, thereby enhancing the model's ability to discriminate between relevant and irrelevant candidates. While previous re-ranking models such as SimCLS, BRIO, SummaReranker, BalSum, and GAR have focused exclusively on news summarization datasets (e.g., CNN/DM and XSum), LexiSem introduces a broader evaluation across domains by including both Reddit TIFU and MeQSum. This highlights its capability to generalize beyond traditional news content to both user-generated and specialized medical summaries. Further research will focus on extending this evaluation to multilingual and scientific domains.

Our experiments demonstrate that LexiSem consistently outperforms strong baselines and existing re-rankers, underscoring its effectiveness in improving summary quality. We hope our findings will inform future explorations of ranking-based approaches in abstractive summarization. Compared to previous work that typically evaluates re-ranking performance using only ROUGE or BERTScore, LexiSem combines all three major evaluation axes: ROUGE, BERTScore, and BARTScore. This enables a more comprehensive view of both surface-level and semantic fidelity. Future work will focus on improving model interpretability by integrating token-level attention mechanisms and attribution methods . Looking ahead, we anticipate refining the two-stage framework by experimenting more advanced Large Language Models for both candidate generation and the re-ranking stage. Additionally, investigating alternative backbones and architectural enhancements

Table 8
Case study on CNN/DM: Lizard Video. R-1/2/L are the ROUGE-1/2/L F_1 scores. BS denotes BERTScore. Irrelevant information is highlighted in blue, while content more consistent with the reference is highlighted in green.

System	R-1	R-2	R-L	BS	Summary
Reference	–	–	–	–	The gruesome lizard was captured in Australia and uploaded last week. The lizard swings its neck back and forth in a bid to swallow the rabbit. Goannas can unhinge their lower jaws, allowing them to swallow large prey.
BRIO-Ctr	51.16	23.81	27.91	88.75	Two-meter long reptile is filmed balancing on top of a power pole to swallow a rabbit. The lizard swings its neck back and forth as it battles to swallow its prey. It finishes the feat in under a minute, and the video was uploaded to YouTube last week.
BalSum	46.91	20.25	34.57	90.72	Two-meter long lizard filmed battling to swallow a rabbit in under one minute. Video shows lizard balance at the top of a power pole while swallowing its prey. Goannas can unhinge their lower jaws when feeding, allowing them to eat oversized prey.
LexiSem	63.52	43.37	61.17	93.54	The footage was uploaded to YouTube last week. The lizard swings its neck back and forth as it battles to swallow the catch. Goannas can unhinge their lower jaws when feeding, allowing them to dispose of some staggering-sized prey, including possums and cats.

Table 9
Case study on CNN/DM: Summaries with Bart baseline model. R-1/2/L are the ROUGE-1/2/L F_1 scores. BS denotes BERTScore.

System	R-1	R-2	R-L	BS	Summary
Reference	–	–	–	–	New pictures show raheem sterling and jordon ibe with shisha pipes. the liverpool pair are dressed in casual clothing and have a pipe each. pictures emerged last week of liverpool star sterling smoking shisha. footage also emerged of him inhaling nitrous oxide from a balloon. the pictures create a fresh problem for liverpool boss brendan rodgers. the images will be a concern for any potential suitors of the liverpool star.
BART	38.1	18.4	31.4	86.3	Raheem sterling and jordon ibe have been pictured smoking from a shisha pipe in the past few days. the images will be a concern to clubs considering parting with £40million to sign the star. becoming as good as lionel messi or cristiano ronaldo — gareth bale even — requires more than ability alone.
LexiSem	63.5	43.3	61.1	93.5	The liverpool star has been pictured smoking from a shisha pipe. jordon ibe was also pictured smoking shisha with sterling. liverpool manager brendan rodgers was due to speak to sterling to remind him of his responsibilities. the images of sterling will be a concern to clubs considering parting with 40million to sign the star.
Reference	–	–	–	–	Chelsea forward tammy abraham nets first-half double for chelsea. dominic solanke adds a third late on as chelsea look set to win trophy. manchester city struggle without injured star thierry ambrose. read: mourinho warns his young chelsea players he can not play them all. **click here** to read our match report from man city's academy stadium.
BART	39.9	15.5	28.5	87.2	Tammy abraham scored twice in the first half to give chelsea the lead. isaac buckley-ricketts levelled the game for manchester city. dominic solanke scored late on to put a gloss on the scoreline. **click here** to read sportsmail's player ratings from the youth cup final.
LexiSem	28.1	10.3	18.1	86.3	Tammy abraham opened the scoring for chelsea with a fierce drive. isaac buckley-ricketts levelled for manchester city. but solanke scored a brace to put chelsea 3-1 up. dominic solanke scored late on to put a gloss on the scoreline.

could further optimize performance, scalability, and adaptability to various summarization domains.

8. Limitations

One notable limitation of our approach is the computational cost and complexity associated with proposition segmentation. Segmenting text into fine-grained propositions requires substantial computational resources, including memory and processing time. This demand is especially pronounced when dealing with large datasets or documents containing complex contextual relationships. Although our multi-GPU infrastructure mitigates some of these challenges, the segmentation process remains a bottleneck, potentially limiting the model's scalability and applicability in resource-constrained settings. Addressing these resource constraints will be a key consideration in future work. Exploring more efficient segmentation methods and developing lightweight models may help alleviate some of these constraints, making the approach more accessible and scalable for broader real-world applications. While recent works such as [67] demonstrate the use of lightweight unsupervised models for efficient quality assessment, LexiSem adopts a more resource-intensive but semantically robust strategy. This trade-off is especially suitable in high-stakes summarization domains (e.g., medical QA), where factual alignment is critical.

Furthermore, our contrastive learning framework relies on cosine similarity-based hard negative sampling, which is effective but not infallible. The selection of negative examples is based on predefined similarity thresholds, which may misclassify certain summaries as negative when they contain valid but paraphrased content. This can lead to unintended penalization of semantically correct summaries and affect overall ranking performance. Cosine similarity itself may also introduce biases in the learning process. Due to the anisotropy of embedding spaces [50], semantic similarity scores may not always reflect true meaning alignment, especially in the presence of syntactic or lexical variance. Embedding-based comparisons can be overly sensitive to surface features such as punctuation or sentence length, potentially skewing the model's judgment of quality. corpora [68], which in turn can propagate through similarity-based contrastive learning, affecting the downstream performance of the re-ranker.

CRediT authorship contribution statement

Eman Aloraini: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Hozaifa Kassab:** Writing – review & editing, Software, Methodology, Conceptualization. **Ali Hamdi:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Khaled Shaban:** Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://github.com/MIRAH-Official/LexiSem.

References

[1] M. Luo, B. Xue, B. Niu, A comprehensive survey for automatic text summarization: techniques, approaches and perspectives, Neurocomputing (2024) 128280.

[2] J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International Conference on Machine Learning, PMLR, 2020, pp. 11328–11339.

[3] M. Lewis, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv preprint arXiv:1910.13461.

[4] A. See, P.J. Liu, C.D. Manning, Get To The Point: Summarization with Pointer-Generator Networks, Association for Computational Linguistics, 2017.

[5] X. Lin, S. Han, S. Joty, Straight to the gradient: Learning to use novel tokens for neural text generation, PMLR, 2021, pp. 6642–6653,

[6] S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, Scheduled sampling for sequence prediction with recurrent neural networks, 2015.

[7] M. Ranzato, S. Chopra, M. Auli, W. Zaremba, Sequence level training with recurrent neural networks, 2016.

[8] Y. Liu, P. Liu, D. Radev, G. Neubig, BRIO: Bringing order to abstractive summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2890–2903.

[9] M. Ravaut, S. Joty, N. Chen, SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4504–4524.

[10] Y. Liu, P. Liu, SimCLS: A simple framework for contrastive learning of abstractive summarization, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 1065–1072.

[11] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[12] J. Sul, Y.S. Choi, Balancing lexical and semantic quality in abstractive summarization, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 637–647.

[13] G. Adams, A. Fabbri, F. Ladhak, K. McKeown, N. Elhadad, Generating EDU extracts for plan-guided summary re-ranking, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2023, 2023, pp. 2680–2697.

[14] Z. Zhao, S.B. Cohen, B. Webber, Reducing quantity hallucinations in abstractive summarization, 2020, pp. 2237–2249.

[15] S. Chen, F. Zhang, K. Sone, D. Roth, Improving faithfulness in abstractive summarization with contrast candidate generation and selection, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5935–5941.

[16] Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh, P.J. Liu, Calibrating sequence likelihood improves conditional language generation, in: The Eleventh International Conference on Learning Representations.

[17] F. Nan, C. dos Santos, H. Zhu, P. Ng, K. Mckeown, R. Nallapati, D. Zhang, Z. Wang, A.O. Arnold, B. Xiang, Improving factual consistency of abstractive summarization via question answering, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 6881–6894.

[18] Z.-Y. Dou, P. Liu, H. Hayashi, Z. Jiang, G. Neubig, GSum: A general framework for guided neural abstractive summarization, 2020, arXiv preprint arXiv:2010.08014.

[19] S. Sun, W. Li, Alleviating exposure bias via contrastive learning for abstractive text summarization, 2021, arXiv preprint arXiv:2108.11846.

[20] S. Xu, X. Zhang, Y. Wu, F. Wei, Sequence level contrastive learning for text summarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, (10) 2022, pp. 11556–11565.

[21] Y. Liu, Z.-Y. Dou, P. Liu, RefSum: Refactoring neural summarization, 2021, pp. arXiv–2104, ArXiv E-Prints.

[22] L. Shen, A. Sarkar, F.J. Och, Discriminative reranking for machine translation, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 2004, pp. 177–184.

[23] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D.A. Smith, K. Eng, et al., A smorgasbord of features for statistical machine translation, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 2004, pp. 161–168.

[24] T. Mizumoto, Y. Matsumoto, Discriminative reranking for grammatical error correction with statistical machine translation, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1133–1138.

[25] M. Collins, T. Koo, Discriminative reranking for natural language parsing, Comput. Linguist. 31 (1) (2005) 25–70.

[26] E. Charniak, M. Johnson, Coarse-to-fine n-best parsing and maxent discriminative reranking, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL'05, 2005, pp. 173–180.

[27] B. Kratzwald, S. Feuerriegel, Adaptive document retrieval for deep question answering, 2018, arXiv preprint arXiv:1808.06528.

[28] R. Nogueira, K. Cho, Passage re-ranking with BERT, 2019, arXiv preprint arXiv:1901.04085.

[29] S. Iyer, S. Min, Y. Mehdad, W.-t. Yih, Reconsider: improved re-ranking using span-focused cross-attention for open domain question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 1280–1287.

[30] V. Pandramish, D.M. Sharma, Checkpoint reranking: An approach to select better hypothesis for neural machine translation systems, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2020, pp. 286–291.

[31] S. Bhattacharyya, A. Rooshenas, S. Naskar, S. Sun, M. Iyyer, A. McCallum, Energy-based reranking: Improving neural machine translation using energy-based models.

[32] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, Adv. Neural Inf. Process. Syst. 34 (2021) 27263–27277.

[33] C.-Y. Chuang, J. Robinson, Y.-Y. Lin, A. Torralba, S. Jegelka, Debiased contrastive learning, in: Proceedings of the 34th Conference on Neural Information Processing Systems, NeurIPS, Vancouver, Canada, 2020, URL: https://proceedings.neurips.cc/paper/2020/hash/....

[34] J. Xie, S. Zhang, X. Zhang, GECSum: Generative evaluation-driven sequence level contrastive learning for abstractive summarization, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 7581–7595.

[35] G. Cao, X. Wang, A contrastive learning approach for reference-based text summarization, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2021, pp. 1234–1244, URL: https://aclanthology.org/2021.acl-long.XXX.

[36] X. Yang, Z. Li, Q. Zhou, Contrastive context-aware learning for machine translation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, Florence, Italy, 2019, pp. 1120–1130, URL: https://aclanthology.org/P19-XXX.

[37] J. Pan, H. Wang, P. Chen, Contrastive regularization for neural machine translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP, Online, 2021, pp. 1333–1344, URL: https://aclanthology.org/2021.emnlp-main.XXX.

[38] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP, Online, 2021, pp. 6894–6910, URL: https://aclanthology.org/2021.emnlp-main.552.

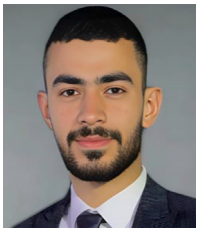
[39] W. Liu, J. Yang, A. Smith, Contrastive re-ranking for abstractive summaries in discrete space, in: Findings of the Association for Computational Linguistics, ACL, Online, 2022, pp. 102–110, URL: https://aclanthology.org/2022.findings-acl.XX.

[40] L. Shen, A. Sarkar, F.J. Och, Discriminative reranking for machine translation, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP, Barcelona, Spain, 2004, pp. 177–184, URL: https://aclanthology.org/W04-XXXX.

[41] F.J. Och, Minimum error rate training in statistical machine translation, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, ACL, Barcelona, Spain, 2004, pp. 160–167, URL: https://aclanthology.org/P04-XXXX.

[42] T. Mizumoto, Y. Matsumoto, Discriminative reranking for grammatical error correction with statistical machine translation methods, in: Proceedings of the 26th International Conference on Computational Linguistics, COLING, Osaka, Japan, 2016, pp. 878–888, URL: https://aclanthology.org/C16-1083.
[43] C. Lee, J. Park, A. Kim, S. Cha, Contrastive decoder for reranking in sequence generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP, Online, 2021, pp. 1462–1473, URL: https://aclanthology.org/2021.emnlp-main.XXX.
[44] H. Fang, X. Zhao, M. Chen, Contrastive learning of stronger language model representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 3420–3430, URL: https://aclanthology.org/2020.acl-main.XXX.
[45] Y. Wan, M. Bansal, Reducing hallucinations in abstractive summarization with contrastive learning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, Abu Dhabi, UAE, 2022, pp. 1236–1247, URL: https://aclanthology.org/2022.emnlp-main.XXX.
[46] S. Kwon, Y. Lee, Enhancing abstractive summarization of implicit datasets with contrastive attention, Neural Comput. Appl. (2024) 1–15.
[47] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, in: International Conference on Learning Representations, ICLR, 2021.
[48] R. Paulus, A deep reinforced model for abstractive summarization, 2017, arXiv preprint arXiv:1705.04304.
[49] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio, An actor-critic algorithm for sequence prediction, 2016, arXiv preprint arXiv:1607.07086.
[50] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6894–6910.
[51] S. Chen, H. Zhang, T. Chen, B. Zhou, W. Yu, D. Yu, B. Peng, H. Wang, D. Roth, D. Yu, Sub-sentence encoder: Contrastive learning of propositional semantic representations, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 1596–1609.
[52] M. Wanner, S. Ebner, Z. Jiang, M. Dredze, B. Van Durme, A closer look at claim decomposition, 2024, arXiv preprint arXiv:2403.11903.
[53] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P.W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023, arXiv preprint arXiv:2305.14251.
[54] R. Kamoi, T. Goyal, J.D. Rodriguez, G. Durrett, Wice: Real-world entailment for claims in wikipedia, 2023, arXiv preprint arXiv:2303.01432.
[55] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, Adv. Neural Inf. Process. Syst. 28 (2015).
[56] S. Narayan, S.B. Cohen, M. Lapata, Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018.
[57] B. Kim, H. Kim, G. Kim, Abstractive summarization of reddit posts with multi-level memory networks, 2018, arXiv preprint arXiv:1811.00783.
[58] A.B. Abacha, D. Demner-Fushman, On the summarization of consumer health questions, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2228–2234.
[59] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[60] A.K. Vijayakumar, M. Cogswell, R.R. Selvaraju, Q. Sun, S. Lee, D. Crandall, D. Batra, Diverse beam search: Decoding diverse solutions from neural sequence models, 2016, arXiv preprint arXiv:1610.02424.
[61] N. Shazeer, M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in: International Conference on Machine Learning, PMLR, 2018, pp. 4596–4604.
[62] J. Zhao, X. Sun, C. Feng, Introducing bidirectional attention for autoregressive models in abstractive summarization, Inform. Sci. 689 (2025) 121497.
[63] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2019, arXiv preprint arXiv:1904.09675.
[64] J. Yang, S. Yoon, B. Kim, H. Lee, FIZZ: Factual inconsistency detection by zoom-in summary and zoom-out document, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 30–45, http://dx.doi.org/10.18653/v1/2024.emnlp-main.3, URL: https://aclanthology.org/2024.emnlp-main.3/.
[65] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinform. 36 (4) (2020) 1234–1240, http://dx.doi.org/10.1093/bioinformatics/btz682.
[66] H. Shakil, A. Farooq, J. Kalita, Abstractive text summarization: State of the art, challenges, and improvements, Neurocomputing (2024) 128255.
[67] E. Rajasekar, H. Chandra, N. Pears, S. Vairavasundaram, K. Kotecha, Lung image quality assessment and diagnosis using generative autoencoders in unsupervised ensemble learning, Biomed. Signal Process. Control. 102 (2025) 107268.
[68] J. Kumarappan, E. Rajasekar, S. Vairavasundaram, K. Kotecha, A. Kulkarni, Siamese graph convolutional split-attention network with NLP based social sentimental data for enhanced stock price predictions, J. Big Data 11 (1) (2024) 154.

Eman Aloraini received the M.Sc. degree from MEU University in 2017. She is currently a Ph.D. student at the Department of computer science, College of Engineering, Qatar University. Her research interests span a variety of areas in Natural Language Processing.



Hozaifa Kassab received the B.Sc. degree in Computer Science from Greenwich University. He is currently working as a Data Scientist and AI Researcher. His research interests include computer vision, particularly in image recognition and generation, as well as natural language processing (NLP). He has authored several publications in these fields, contributing to advancements in deep learning and artificial intelligence.

Ali Hamdi received his PhD from RMIT University, Australia, in 2022, where his research focused on computer vision, graph neural networks, and uncertainty modeling. He obtained his Master's in Computing from the University of Technology, Malaysia, in 2017, with a focus on NLP and text mining. With over 15 years of experience as a data scientist and researcher, Ali has a robust background in teaching and developing software across various domains, including programming, cloud computing, AI, machine learning, deep learning, and data science. He has authored research papers on visual recognition, language understanding and generation models, drone-based object tracking and route optimization, few-shot learning, and spatiotemporal data mining, making significant contributions to computer vision, NLP and deep learning.

Khaled Shaban received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Canada, in 2006. He is currently a Professor at the Computer Science and Engineering Department, College of Engineering, Qatar University, Qatar. His research experience in academic and industrial institutions covers a variety of domains in intelligent systems application and design.