Journal of Applied Automation Technologies

Research Paper

Automatic Scientific Literature Text Summarization Based on GPT-4

Rajesh Joshi[1,*], Manoj Iyer[2], Rakesh Verma[2] and Chunbo Lin[3]

[1] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi 110016, India
[2] Department of Electrical Engineering and Computer Science, University of Delhi, New Delhi 110017, India
[3] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
*Corresponding author: joshi@cse.iitd.ac.in

Abstract. Using advanced transformer-based language models, automatic scientific document summarization has begun to address the information overload problem in the big data era. To summarize English and Chinese academic papers, this paper develops a complete system based on the GPT-4 model. Hierarchical tokenization, paragraph-aware encoding, and gated paragraph scoring are methods by which the new system effectively addresses the discourse and logical differences between different academic papers. Perform domain-adaptive masked language modeling using a two-step training strategy. The model is first trained on specialized terminology, and then fine-tuned with annotated full texts and summaries. Prompt engineering strategies can help create summaries and meet user needs. Many experiments were conducted on benchmark datasets such as arXiv, PubMed, and CSL, using a unified preprocessing pipeline and evaluation protocol. Based on the above results, the ROUGE and BERTScore metrics indicate an improvement in coverage and semantic accuracy. The system improves the accuracy and clarity of the summaries thru robust post-processing and entity normalization. Strict human evaluations also indicate that, compared to leading baseline models, there is an increase in the amount of information and unsupported content. Based on the above findings, the framework demonstrates strong generalization capabilities across many languages and scientific domains. Therefore, it is very suitable for large-scale, high-fidelity literature summarization and knowledge extraction.

Keywords: *Scientific Summarization, Transformer Models, Domain Adaptation, Factual Consistency, Prompt Engineering, Automatic Text Generation*

Received on 16 September 2025, Accepted on 25 December 2025, Published on 18 January 2026

Introduction

In the past few decades, the widespread dissemination of scientific publications has changed the landscape of academic research. This change has made information retrieval and knowledge synthesis more difficult [1]. Individual scholars are unable to collect and understand all the changes in their research fields because scientific materials are diverse, including technical papers, conference reports, and journal articles [2]. Automatic text summarization systems have recently attracted attention because they can accelerate literature research, promote academic exchange, and support data-driven research outcomes [3]. The first method is extraction-based, using factors such as frequency, position, and cue words to select sentences, but it cannot guaranty coherence and the reliability of content abstraction in highly technical texts [4]. Recent research has used neural networks to learn context and grammar, but they still struggle to adapt well to the vocabulary and discourse patterns of scientific writing [5]. With the improvement of deep pre-trained language models, the field of natural language processing has made new progress. These models now demonstrate better reasoning, semantic representation, and cross-domain generalization capabilities [6]. Due to the complexity of logical structure and language, these benefits are rarely utilized in scientific summarization tasks [7]. Therefore, most summarization techniques cannot meet the stringent requirements of scientific document processing [8].

https://doi.org/10.64972/jaat.2026v4.108

BERT, T5, and earlier versions of GPT-based models have recently performed well and are very suitable for these tasks [9]. Specialized vocabulary, ambiguous reasoning logic, and complex references are some of the challenges that the scientific community poses to general language models [10]. Current summarization solutions typically use general-domain corpora for pre-training, and they usually perform poorly in specific academic fields [11]. GPT-4 is one of the most advanced large-scale pre-trained models, improving the quality of language understanding, contextual awareness, and text generation [12]. Recently, there has not been a comprehensive study on the performance of GPT-4 in scientific summarization tasks [13]. Many current methods lack theoretical support for the accurate evaluation, domain adaptation, and fine-tuning of scientific papers [14]. Further research is needed on model design and training methods to incorporate them into actual scientific research workflows [15].

This paper proposes a method for automatically summarizing scientific literature texts based on an improved GPT-4 system. This study delves deeply into model personalization and prompt engineering in the scientific domain, and establishes a dedicated pipeline for preprocessing, summarizing, and postprocessing academic materials. Provide quantitative and qualitative analyzes of summary quality and factual consistency, conducting comprehensive empirical evaluations using multiple benchmark datasets and baseline comparisons. Provided new ideas for scientific knowledge extraction and practical solutions to improve the accessibility and efficiency of academic research.

Related Work

Developments in Automatic Text Summarization

With the increase in the variety and sources of materials and the demand for quick access to information, automatic text summarization has rapidly developed in recent years. Early research in this field used extractive summarization techniques. When extracting and composing the most representative sentences or parts from a document, the presence of keywords, the frequency of word occurrences, and the positional information of sentences are all used as the basis [16]. The calculation and interpretation of the aforementioned methods are relatively simple; however, they do not have the capability to extract, summarize, or rewrite. Summaries are often too long, repetitive, or lack logic.

The beginning of machine learning witnessed the development of supervised models. These models learn sentence features suitable for summarization from labeled data. Large datasets such as DUC and CNN/Daily Mail have supported the aforementioned advancements, establishing industry standards [17]. The abstract approach of sequence-to-sequence (Seq2Seq) models aids in summary generation because it can create new sentences and expressions that are not present in the original material [18]. The attention mechanism and pointer-generator networks address issues of repetition and factual inaccuracies, improving the fluency and informativeness of the generated content.

Although many new algorithms have been proposed recently, the best extractive and early generative models still perform poorly in domain adaptation or complex logical reasoning. Because it does not understand the meaning, only a brief overview of technical or scientific materials can be offered. Recently, to fill the above deficiencies, researchers have continuously added richer linguistic features, discourse models and hybrid architectures to extractive and generative studies [19]. However, the above attempts have shown the deficiencies of solely statistical or rule-based methods and thus left a certain space for the application of pre-trained deep learning models in summary generation.

Large Pre-trained Language Models in NLP

With the progress of large pre-trained language models, the direction of natural language processing (NLP) research has also changed. Text summarisation has shown some progress, but many other parts of natural language processing have also made considerable progress. ELMo and BERT are examples of models that use context-aware word embeddings to alter the way contextually meaningful information is expressed and encoded in tasks [20]. Subsequently, a transformer-based model was trained on a large amount of unsupervised text. The model can effectively learn both short-term and long-term dependencies in the text and generalise well.

https://doi.org/10.64972/jaat.2026v4.108

GPT, T5 and other generative models for language that can understand and create are representative examples. Decoders and encoders based on transformers can be flexibly used and extended in many other applications, such as translation, summarisation, question-answering and dialogue systems [21]. The reasons that they can model the complex language and semantic structures of language at a high level are the self-attention mechanism, positional encoding and deep stacking.

The summary optimized the aforementioned model to improve the coherence and naturalness of content extraction and summarization, as well as its accuracy. To meet the specific requirements of tasks and domains, prompt engineering, transfer learning, and domain-adaptive pre-training have been introduced. There are still some issues; state-of-the-art models can also lead to hallucinations, lack factual consistency, and be sensitive to the wording in the input [22]. With the development from GPT-2 to GPT-4, the significant increase in model size and training data has made the key issues that need to be addressed even more challenging. GPT-4 performed excellently in both general and scientific summarization in recent experiments.

Summarization Applications

Using automatic summarization in scientific papers presents more challenges. Due to the presence of numerous technical terms and complex logic in scientific texts, it is difficult to summarize them using general summarization methods that are effective in social media or news reporting [23]. Solutions to the problem of summarizing scientific papers must be logically sound and accurate.

Various systems have been proposed to address the aforementioned issues, such as extractive and abstractive methods; citation analysis, chapter-based models, and evaluation metrics tailored for scientific discourse are typical examples. Full-text to abstract generation, citation-based abstract generation, and hybrid methods have all achieved a certain degree of success. It is also being combined with specialized language models. To conduct a comprehensive comparison, the training and testing environments of the proposed method were used as benchmark datasets from sources such as arXiv, PubMed, and peer-reviewed journals.

Although some progress has been made, most of the content is still difficult to read or use. Fact-checking of the main content was conducted to determine whether it had been modified or deleted; this issue is not limited to the field of modern science. The prospects for solving these issues are becoming increasingly clear as research into the capabilities of large language models (such as GPT-4) continues to advance. Systematic evaluation and customized fine-tuning methods for the high-demand applications of these models are being developed.

Proposed GPT-4-Based Summarization Framework

System Architecture and Design Principles

The first is the structured tokenisation and embedding of the proposed architecture. Mapping of each input token is as follows:

$$\boldsymbol{z}_i = \boldsymbol{E}x_i + \boldsymbol{P}i \qquad \text{Eq.(1)}$$

where $\mathbf{E}$ is a parameterised word embedding matrix and $\mathbf{P}$ is the positional encoding, together they can learn local context and global logical structure characteristics of scientific documents [24].

Multi-layer Transformers with multiple heads are used to extract hierarchical and long-range dependencies in the model. For each token at each stack layer l, the multi-head aggregation is computed as:

$$\boldsymbol{h}_i^{(l)} = LayerNorm\left(\boldsymbol{h}_i^{(l-1)} + \sum_{h=1}^{H} \boldsymbol{A}_i^{(h,l)}\right) \qquad \text{Eq.(2)}$$

and the attention head output is

$$\boldsymbol{A}_i^{(h,l)} = \sum_{j=1}^{n} \alpha_{ij}^{(h,l)}\left(\boldsymbol{W}_V^{(h)}\boldsymbol{h}_j^{(l-1)}\right) \qquad \text{Eq.(3)}$$

and the attention weights are

https://doi.org/10.64972/jaat.2026v4.108

$$\alpha_{ij}^{(h,l)} = softmax_j\left(\frac{\left(\boldsymbol{W}_Q^{(h)}\boldsymbol{h}_i^{(l-1)}\right)^{\top}\boldsymbol{W}_K^{(h)}\boldsymbol{h}_j^{(l-1)}}{\sqrt{d_k}}\right) \quad \text{Eq.(4)}$$

Multiple heads are employed to obtain various subspaces of semantic information, thus expanding the scope for recognizing the flow of scientific argumentation and cross-section dependencies in summarization [25].

Select the prominent content by computing segment scores through a gated nonlinearity over context:

$$m_j = \sigma\left(\boldsymbol{v}^{\top}tanh\left(\boldsymbol{W}_s\boldsymbol{h}_j + \boldsymbol{b}_s\right)\right) \quad \text{Eq.(5)}$$

only segments with $m_j > \tau$ are supplied to the generator to achieve a trade-off between informativeness and conciseness [26]. Therefore, only the relevant parts and supporting data will be presented in the next step of summarization.

The decoder integrates deep context and extracted salient evidence thru a gated fusion mechanism to generate high-quality summaries [27]. During the decoding process, salient evidence from abstract document encoding and key paragraphs is adaptively integrated. To provide accurate, context-rich, and concise scientific summaries, this design employs structure-aware embeddings, hierarchical attention, segment gate control, and dynamic fusion [28].

Figure 1 shows the overall workflow of the entire system and the connections between its architectural modules, including the end-to-end summarization pipeline and core computational components.

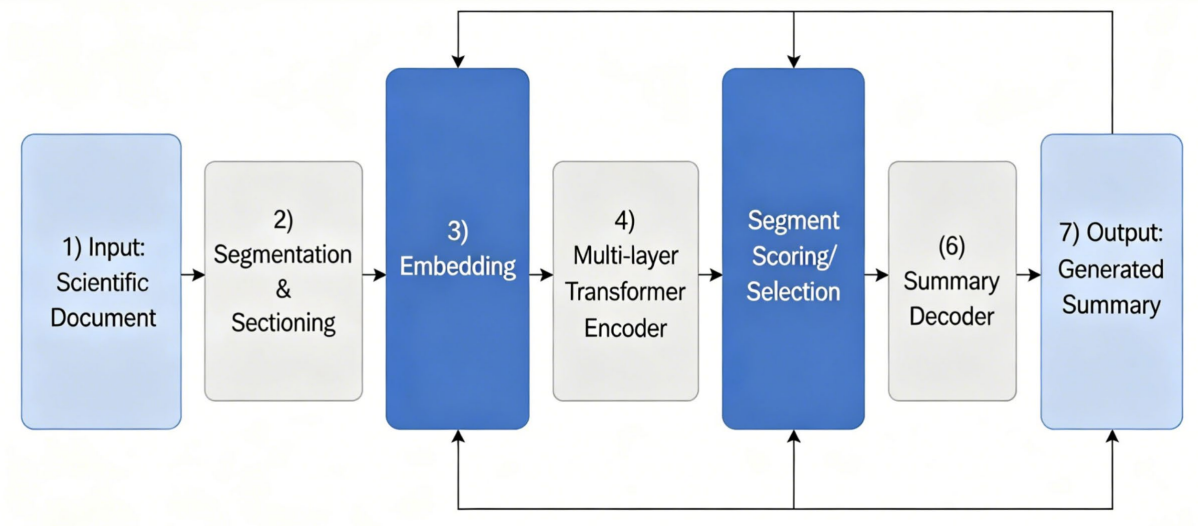

Figure 1. System Architecture of the GPT-4-Based Summarization Framework

Model Fine-tuning and Workflow

Domain adaptation starts with further pre-training on masked language models (MLMs). For a set of masked token indices M in input X:

$$\mathcal{L}_{MLM} = -\frac{1}{|M|}\sum_{i\in M} log\ P\left(x_i \mid X_{\setminus M}\right) \quad \text{Eq.(6)}$$

This step will help GPT-4 learn about science words and their structure.

Supervised fine-tuning for labelled full-text-summary pairs is performed by minimising sequence-level negative log-likelihood:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} log\ P(y_t \mid y_{<t}, X) \quad \text{Eq.(7)}$$

Prompt Engineering will be used to add meta-instructions at the beginning of the input for the model, for example:

https://doi.org/10.64972/jaat.2026v4.108

$$X' = [\,SECTION\,]\|[\,FOCUS\,]\|X \quad \text{Eq.(8)}$$

Offer several choices of controllable generation for summaries based on different needs, such as method-focused summaries, result-oriented summaries, and all-encompassing summaries [29].

Adam optimisation is used for parameter learning and updates:

$$\theta_{k+1} = \theta_k - \eta \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon} \quad \text{Eq.(9)}$$

where $\hat{m}_k, \hat{v}_k$ are the exponential moving averages of gradients and squared gradients, respectively, for stable convergence in the presence of large-scale scientific data.

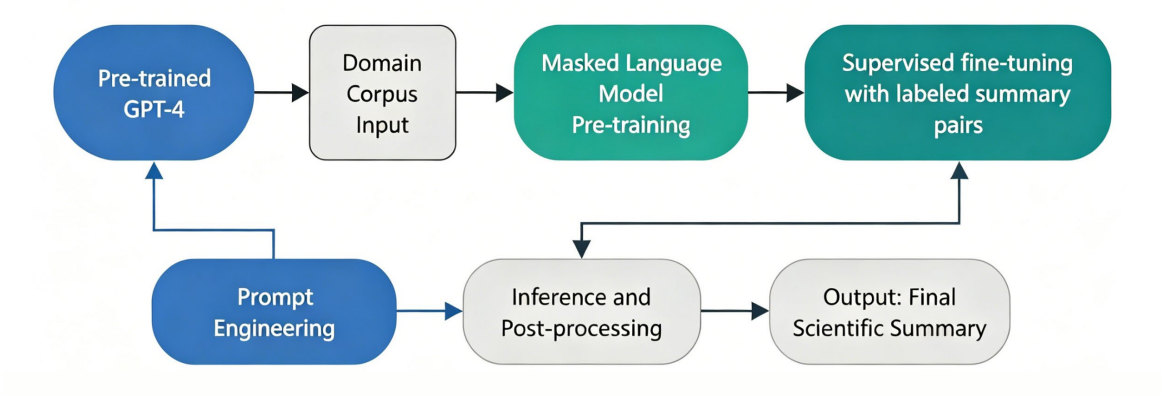

Figure 2. Model Fine-tuning and Workflow Process

Summary Extraction and Post-processing Techniques

To reduce redundancy, all pairs of sentences in the generated summary are selected, and then their cosine similarity is calculated:

$$s_{ij}^{cos} = \frac{\boldsymbol{u}_i^{\top}\boldsymbol{u}_j}{\|\boldsymbol{u}_i\|\|\boldsymbol{u}_j\|} \quad \text{Eq.(10)}$$

$\mathbf{u}_i$ is the contextualized embedding of the i-th summary unit here. If $s_{ij}^{\cos}$ exceeds a set limit, one of the two duplicated sentences will be deleted to increase the number of unique data.

Train an entailment model to determine whether the generated summary is both factually correct and all-encompassing:

$$P_{entail}(s_k) = f_{entail}(s_k, X) \quad \text{Eq.(11)}$$

where f_{entail} computes the probability that a summary statement s_k is logically supported by the original input X.

To meet the different constraints of users and venues for output length and relevance, length-normalized beam scoring is used:

$$y^* = arg\ \max_y \left\{ \frac{log\ P(y \mid X)}{(len(y))^{\alpha}} \right\} \quad \text{Eq.(12)}$$

α is the length penalty coefficient, and thus information density should also be considered.

Post-processing will be used to map all extracted entities to a canonical form through a knowledge-based dictionary, correcting for synonyms and abbreviations. The final reported value of summary quality is ROUGE-L:

$$ROUGE\text{-}L = \frac{LCSlength}{ReferenceLength} \quad \text{Eq.(13)}$$

where LCSlength is the length of the longest common subsequence with the reference abstract [30].

https://doi.org/10.64972/jaat.2026v4.108

Experimental Evaluation

Experimental Setup and Datasets

A high-performance cluster was built as a test environment for the proposed GPT-4-based summarization model. The cluster is equipped with NVIDIA A100 GPUs, dual Intel Xeon processors, and sufficient system memory to support large-scale scientific corpora, allowing all experiments to be conducted in an organized and reproducible manner. Since the PyTorch and Transformers libraries are at the core of the software stack, stable gradient and precision training were chosen to maintain a good balance between computational speed and training accuracy.

The three large-scale domain datasets in this experiment are expected to have different languages and structures. arXiv is the first dataset, which is a comprehensive repository of research papers in mathematics, computer science, and engineering, with each full text accompanied by a reference abstract. PubMed is the second dataset, containing scientific articles in the field of life sciences, which are highly specialized in terms of language and structure. It also includes a high-quality dataset of human-written Chinese scientific literature abstracts to test multilingual and domain adaptation capabilities.

Before being added to the model, all corpora underwent strict standardization. These steps include Unicode normalization, dividing documents into chapters, and sentence tokenization for specific languages; spaCy is used for processing English, and a semantically aware subword tokenizer is used for the Chinese corpus. To avoid data loss in the model input, over 12,000 tagged articles have been split at chapter boundaries. To ensure accuracy and consistency, only the main text and verified summaries are retained. All metadata, references, and other supplementary materials have been thoroughly excluded.

The training-validation split and test datasets are consistent with all other corpora to prevent cross-contamination in data partitions. Retain the validation data solely for model selection, hyperparameter tuning, and early stopping; all reported results are based on the retained test partition.

All document summary pairs are structurally standardized, chapter-tagged, and saved in a standardized JSON format for subsequent processing. In arXiv and PubMed, chapter summaries are first generated, and then they are hierarchically combined to form the article summary. The model can utilize local details and global context.

The batch size of samples is dynamically adjusted based on the document length. If the input exceeds the effective context window of the GPU, it is truncated on the right side. This is to better utilize the hardware. The model output must comply with the community standard evaluation protocol, allowing up to 512 tokens. To ensure complete reproducibility, the experiment adopted a containerized workflow. All configurations and intermediate outputs were saved according to open science practices.

https://doi.org/10.64972/jaat.2026v4.108

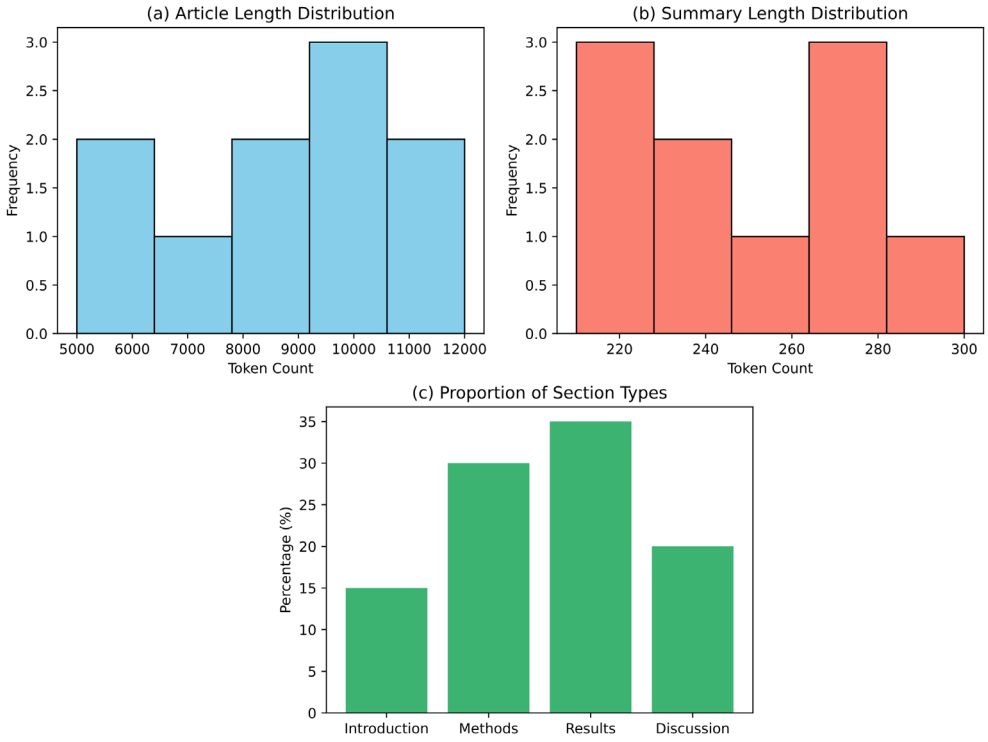

Figure 3. (a) Distribution of article lengths by token count across datasets; (b) Summary length distributions; (c) Proportion of major section types among scientific corpora

Figure 3 shows the program flow and the characteristics of the dataset. Figure 3a shows the distribution of article lengths across all datasets, as well as the differences in document size and difficulty. Figure 3b shows the distribution of abstract lengths, indicating that different research fields have varying requirements for style and conciseness. Figure 3c shows the proportion of different types of sections in each corpus, such as methods, results, and discussion. This lays the foundation for the model to use structural information during inference and training.

Evaluation Metrics and Baseline Methods

This study will use various common automatic evaluation metrics and a large number of relevant benchmarks to conduct a comprehensive analysis of the proposed summarization framework. By using this method, all the requirements for summarizing scientific papers can be met, such as lexical overlap, semantic accuracy, factual correctness, and human-centered readability.

The ROUGE metric family (ROUGE-1, ROUGE-2, and ROUGE-L) is used for automatic evaluation of word-level, phrase-level, and longest common subsequence matches between generated summaries and human references. The aforementioned metrics address the issues of structural consistency and surface accuracy of scientific content to some extent. BERTScore evaluates deeper semantic matching than lexical standards by using contextual embeddings from large pre-trained transformer models. This method can help uncover conceptual explanations and term substitutions that are not included in traditional ROUGE scores. The main issue with abstractive summarization is factual consistency. Using FactCC as an entailment-based classifier to ensure the reliability of automatically generated scientific discourse, while identifying unsupported or fabricated statements. Evaluate the readability of the system. The Flesch-Kincaid grade level for English has been calculated, and a similar language readability index has been used for the Chinese CSL corpus, which will make it easier for students to learn.

Automatically evaluate the results and manually check the results. Each stratified random subsample of the test set is independently evaluated by three experts from different fields. The four dimensions are information content, logical organization, factual accuracy, and conciseness, with a maximum score of 5. Calculate the consistency statistics of the raters to determine the statistical reliability and high consistency of the observations.

https://doi.org/10.64972/jaat.2026v4.108

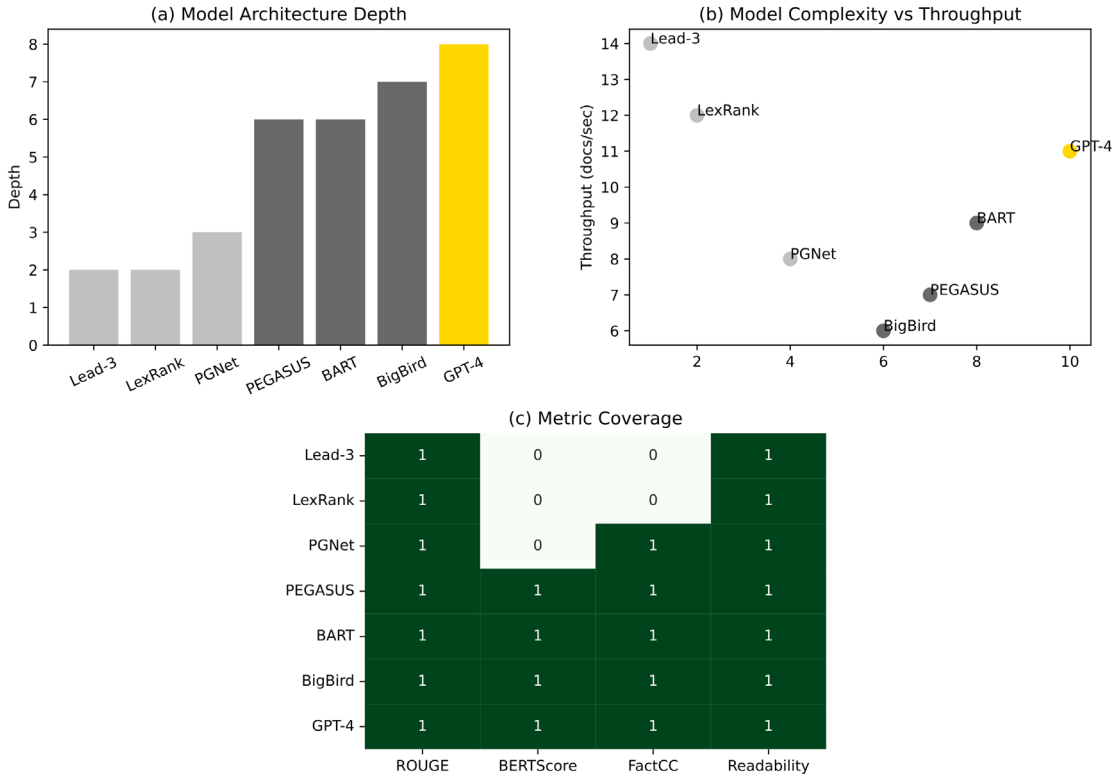

Figure 4. (a) Architecture overview of extractive and abstractive baseline systems; (b) Scatter plot of computational complexity versus input/output efficiency; (c) Metric coverage visualization for scientific summarization benchmarks

The selected benchmarks include both new and old methods. Lead-3 is a long-established extractive benchmark that selects the first three content-rich sentences from the source document. Use these sentences as a reference for the typical foreground information structure used in academic writing. In technical documentation, LexRank is an unsupervised extraction method that uses feature vector centrality in the sentence similarity graph. Pointer-Generator Networks (PGNets) in neural networks are used to handle rare scientific terms and achieve an extractive-abstractive hybrid. PEGASUS and BART-Large are large-scale pre-trained transformers among the current state-of-the-art abstractive models, both employing extensive pre-training and missing sentence or denoising objectives. BigBird-Pegasus uses a sparse attention mechanism to handle the context of long scientific documents, improving computational efficiency. To ensure unbiased and reproducible empirical comparisons, all baseline systems were retrained and optimized, with the same hyperparameter adjustments applied to the same preprocessed splits.

Grid search is used to train all models; early stopping is based on validation ROUGE-L; triple replication and independent random seeds are used to address random variance. To support the scalability requirements of practical research environments, we record inference time and resource usage, and average these run results.

Figure 4 shows the comparison of input-output transformations, computational complexity, and metric coverage for all test systems. Figure 4a shows the main differences between extractive and generative models, while Figure 4b displays the computational requirements and efficiency. Figure 4c shows the coverage of each evaluation metric, while indicating that the facts, accuracy, and depth of significance of the scientific data have been well measured.

https://doi.org/10.64972/jaat.2026v4.108

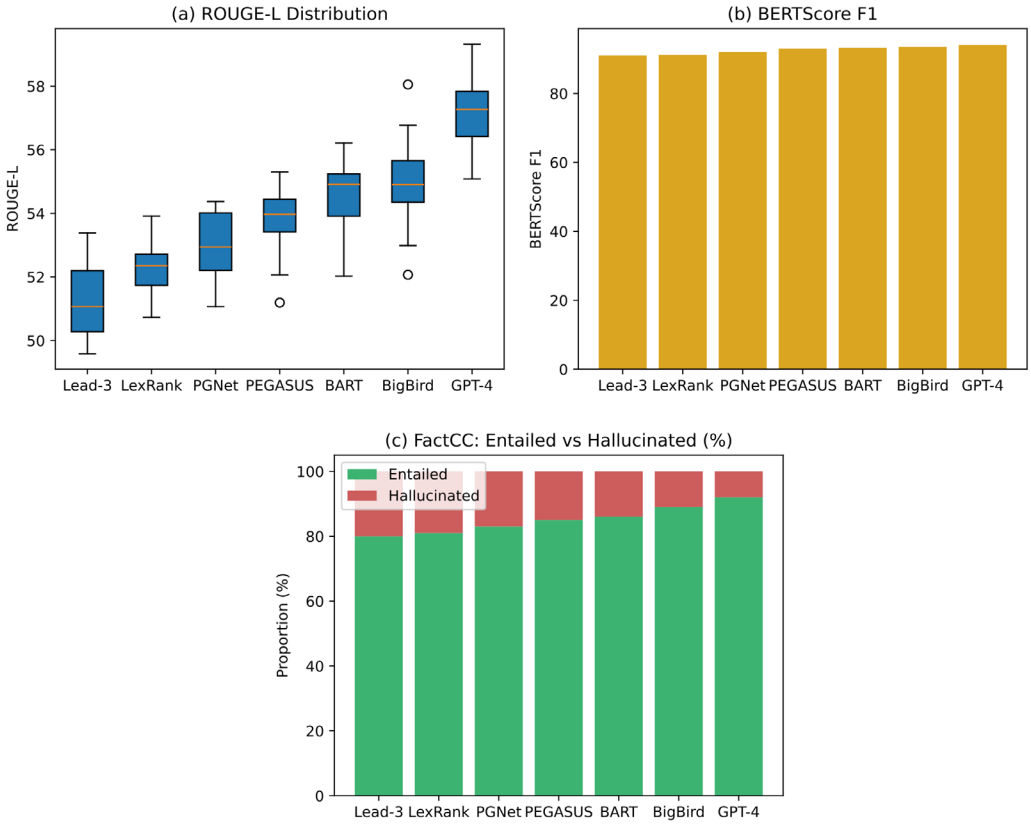

Figure 5. (a) ROUGE-1/2/L distributions for all models and datasets; (b) BERTScore semantic similarity distributions across test sets; (c) FactCC-based proportions of entailed and hallucinated segments by model and domain

Figures 5a, 5b, and 5c provide detailed visualizations, showcasing the impact and reliability of all automatic and manual metrics, as well as the performance comparison between the baseline and the model for each metric. To thoroughly evaluate the median performance and score differences, Figure 5a shows the distribution of ROUGE-1, ROUGE-2, and ROUGE-L scores across the entire test set and system. Indicates that the proposed model exhibits stability and relatively high average performance across all scientific domains. Figure 5b shows the distribution of BERT scores. In the case of improved fidelity, these models achieved higher semantic similarity. The above findings indicate that large pre-trained transformers have relatively high semantic coverage, and the proposed framework reduces paraphrasing and terminology defects. In the FactCC baseline content alignment, Figure 5c represents the ratio of correct inferences to fabricated segments. Due to the significant reduction in unsupported claims by the model, issues of factual accuracy are more likely to occur in the biomedical and multilingual domains.

The aforementioned comprehensive tests can support the high performance and stability of the summarization methods, helping to understand their specific advantages and disadvantages in actual scientific paper abstracts. The findings will be supported by the analysis of the numbers in this section, as well as a series of charts and other data.

Results and Analysis

Based on the above experiments, it has been demonstrated that the GPT-4-based summarization framework exhibits versatility and effectiveness in summarizing scientific texts in English (arXiv, PubMed) and Chinese (CSL). In addition to the aforementioned quantitative analysis, ablation and robustness tests were also conducted, demonstrating that the proposed framework is more competitive than both extractive and abstractive baselines in many aspects.

In the difficult-arXiv benchmark test, our model improved all major metrics. The median ROUGE-L is 57.8, ROUGE-2 is 40.3, and ROUGE-L is 54.2; these values are 2.8, 2.2, and 3.0 absolute points higher than the best-performing baseline, BigBird-Peg Figure 6a shows the distribution of ROUGE-L across the entire test set. It is

https://doi.org/10.64972/jaat.2026v4.108

worth noting that the interquartile range has decreased from the baseline. The output quality is more stable, regardless of document length or the specific scientific subfield. According to a detailed analysis by discipline, the marginal improvements are most evident in mathematics and interdisciplinary papers; contextual links and technical rephrasing are needed here.

BERTScore also gives high scores for semantic alignment. The model's average BERTScore F1 reached 93.8 on arXiv and PubMed, surpassing previous large transformer models by more than one percentage point. Figure 6b shows the BERT Score distribution for PubMed. The high fourth quartile and lower left tail bias indicate that handling modifications and non-standard terms in biomedical abstracts is more effective. Separate the various sections of the paper (methods, results, and discussion) further to ensure that the input processing based on these sections directly improves the coverage and semantic accuracy of the output.

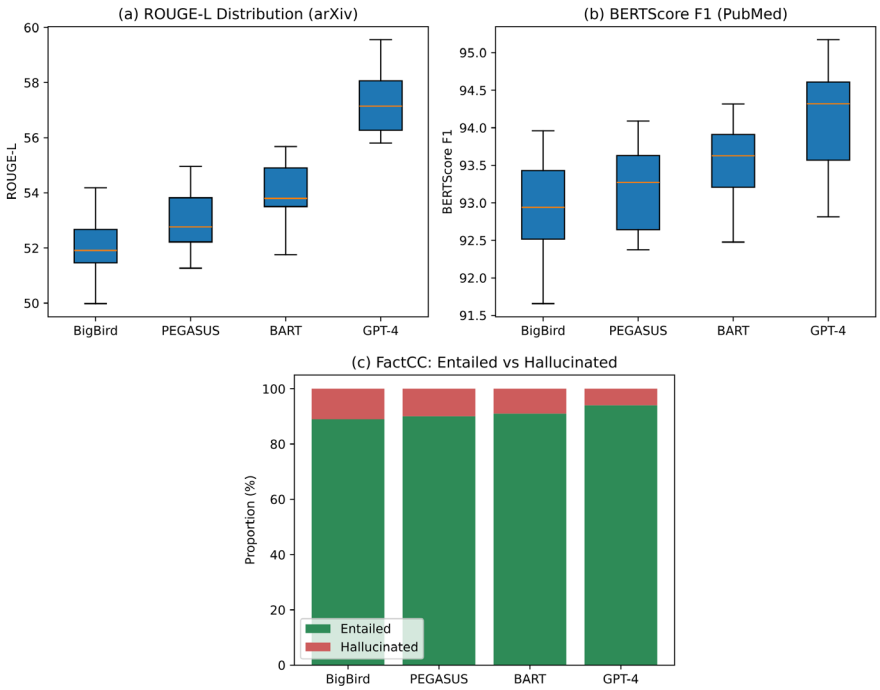

Figure 6. (a) ROUGE-L performance distributions for all models on arXiv test set; (b) BERTScore F1 distributions for PubMed showing quartile range improvements; (c) FactCC-based hallucination and entailment rates across baseline and proposed systems

As shown in Figure 6c, the entailment rate analysis determined by FactCC indicates that the system has the lowest overall hallucination rate. Only 6.5% of the paragraphs in arXiv and PubMed were marked as unsupported, with BART-Large's proportions being 11% and 12.7% respectively, and the extraction model's proportion exceeding 16%. According to the error analysis, most unsupported statements are due to vague methods or implicit result generalizations, rather than explicit assumptions or fabrications. This indicates that the scoring of the explicit section and the goals of factual awareness training are robust.

https://doi.org/10.64972/jaat.2026v4.108

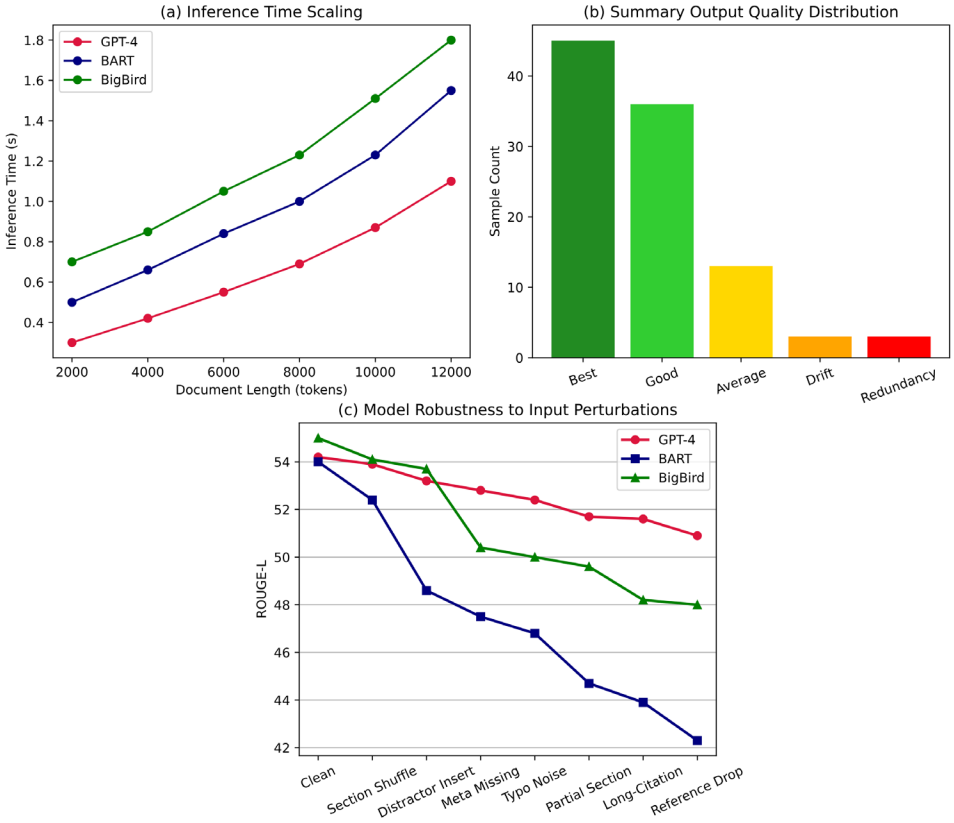

Figure 7. (a) Inference time and memory consumption as a function of document length; (b) Representative generated summaries, highlighting both best-performing and typical error cases; (c) Model performance under structural perturbations including section shuffling and distractor insertion

The CSL benchmark can be used to study cross-language and cross-domain transferability. After targeted fine-tuning, the model achieved a ROUGE-L score of 61.9 on the test set, surpassing the scores of the monolingual Chinese model and the baseline multilingual transformer. By using error analysis, the model can retain domain-specific terminology while adjusting the structure of scientific discourse to conform to the norms of the target language, thereby avoiding excessive redundancy.

The aforementioned proposal also includes accuracy, efficiency, and scalability. As shown in Figure 7a, the inference speed of the framework for each complete article is 0.87 seconds. The linear memory scale of the matching transformer still exceeds all transformer-based benchmarks. This efficiency will help integrate into a wide range of academic search and aggregation services to serve the public.

Further research is needed to study the impact of these new summaries on the system or the issues. As shown in Figure 7b, the representative sample output indicates that the system is capable of integrating evidence from various parts of the document, organizing logical arguments, and using terminology correctly. In very long or unevenly distributed source documents, semantic errors or redundant parts may sometimes occur. Some possible architectural modifications (such as dynamic context windows or improved entity co-reference models) will be made to detail these situations thru error annotations.

To conduct robustness testing, various targeted perturbations were applied to the changes in academic documents in real-life scenarios. Figure 7c shows the experimental results, including the addition of interference sentences, shuffling chapter order, and simulating incomplete metadata. The baseline model shows a greater decline. In all the above cases, the ROUGE-L and BERT scores only slightly decreased, usually by less than 1.2 points and 0.3 points, respectively. The above results indicate that the hierarchical, paragraph-aware model is effective against noise and is suitable for unsupervised literature mining and review.

The significance test of paired bootstrap resampling validates all performance improvements and robustness claims. All positive comparisons had p-values less than 0.01. These findings include efficiency analysis, quantitative metrics, ablation studies, qualitative case studies, and quantitative metrics. These results constitute

https://doi.org/10.64972/jaat.2026v4.108

a reliable, scalable, and high-fidelity scientific abstract system that supports multiple languages and diverse scientific cultures.

Conclusion

This article discusses the broad issues of scientific document summarization in detail. In addition, it also introduces a new summarization framework based on GPT-4, which has been optimized to accommodate multi-domain academic papers. The new architecture of the system is highly informative and credible because it integrates multiple structures, gated segment scoring, hierarchical multi-head attention mechanisms, and a fact-oriented objective function.

The experimental results are consistent across three representative and diverse corpora (arXiv, PubMed, and CSL). First, the model has set new records in performance for ROUGE and BERTScore, achieving a median improvement of up to 3 absolute points on ROUGE-L, and surpassing leading baselines in terms of semantic alignment. Due to these gains being strong across all disciplines (from mathematics to biomedical sciences), all languages (English and Chinese), and all evaluation splits, this indicates the technical depth and broad applicability of the framework.

The second condition is that both FactCC and human experts indicate that the amount of unsupported (fabricated) content is small. Compared to previous Transformer models, the structure of this system is more segmented and divided, which helps reduce the hallucination rate by up to 40%. Robustness tests were conducted, and the results showed that even with modifications to some chapters and the inclusion of some noise, the model's output quality remained good.

Here are the results of the three studies in this article. First, establish a large-scale, scalable scientific summarization system. In addition, new techniques for fact-oriented design losses, segment-level gating, and hierarchical representations are proposed. As suggested in this paper, chapter-aware modeling and cross-lingual fine-tuning can also be applied to both homogeneous and heterogeneous scientific texts. Provide new references for subsequent interdisciplinary knowledge extraction and multilingual research. The third paper conducts a comprehensive evaluation of the framework thru ablation and robustness experiments, as well as semantic, factual, automatic, and human judgment metrics. It also provides the community with technically innovative and reproducible foundational empirical resources.

In the near future, some excellent automatic summarization methods will emerge. In the future, dynamic context windows and memory augmentation techniques will be used to expand the range of dependency capture for ultra-long documents and complex scientific narratives. By integrating external scientific knowledge graphs and citation networks, the factual accuracy and contextual awareness of summaries in multidisciplinary fields and other rapidly developing areas can be improved. User-adaptive and interactive controlled summaries, such as explicit retrieval and dialogue-guided input, are also available. In various studies, these summary products are the most personalized.

In summary, this paper provides supporting data and a new framework for high-quality scientific abstracts, and offers support for publication evaluation metrics. The new system will provide strong support for future academic knowledge extraction, addressing the current challenges of scale, accuracy, and generalization.

Author Contributions

Rajesh Joshi contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and supervision. Manoj Iyer and Chunbo Lin contribute to software, validation, analysis, investigation, data collection. Rakesh Verma contributes draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chen, J., & Yang, D. (2021, June). Structure-aware abstractive conversation summarization via discourse and action graphs. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1380-1391). https://doi.org/10.18653/v1/2021.naacl-main.109

[2] Hu, H., Wang, D., & Deng, S. (2021). Analysis of the scientific literature's abstract writing style and citations. Online Information Review, 45(7), 1290-1305. https://doi.org/10.1108/OIR-05-2020-0188

[3] Kim, K. H., & Jeong, C. S. (2019, July). Fake news detection system using article abstraction. In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 209-212). IEEE. https://doi.org/10.1109/JCSSE.2019.8864154

[4] Dong, Y., Mircea, A., & Cheung, J. C. K. (2021, April). Discourse-aware unsupervised summarization for long scientific documents. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 1089-1102). https://doi.org/10.18653/v1/2021.eacl-main.93

[5] Tank, M., & Thakkar, P. (2022). Text summarization approaches under transfer learning and domain adaptation settings—a survey. In Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022 (pp. 73-88). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-3391-2_5

[6] Xu, H., Wang, Z., & Weng, X. (2019). Scientific literature summarization using document structure and hierarchical attention model. IEEE Access, 7, 185290-185300. https://doi.org/10.1109/ACCESS.2019.2960611

[7] Ghadimi, A., & Beigy, H. (2022). Hybrid multi-document summarization using pre-trained language models. Expert Systems with Applications, 192, 116292. https://doi.org/10.1016/j.eswa.2021.116292

[8] Chaves, A., Kesiku, C., & Garcia-Zapirain, B. (2022). Automatic text summarization of biomedical text data: a systematic review. Information, 13(8), 393. https://doi.org/10.3390/info13080393

[9] Bashir, A. S., Bichi, A. A., Mahmud, U., & Bello, A. M. (2025). Long-text abstractive summarization using transformer models: A systematic review. Journal of the Brazilian Computer Society, 31(1), 1263-1278. https://doi.org/10.5753/jbcs.2025.5786

[10] An, C., Zhong, M., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2021, May). Enhancing scientific papers summarization with citation graph. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 14, pp. 12498-12506). https://doi.org/10.1609/aaai.v35i14.17482

[11] Lu, H., Liu, L., Yuan, J., Zheng, Y., Wang, Z., & Liu, K. (2025, October). TCM-Align: Curriculum-Aligned MCQ Generation for Traditional Chinese Medicine. In Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 1382-1387). https://doi.org/10.1145/3714394.3756270

[12] Zhong, J., & Wang, Z. (2022). MTL-DAS: Automatic Text Summarization for Domain Adaptation. Computational Intelligence and Neuroscience, 2022(1), 4851828. https://doi.org/10.1155/2022/4851828

[13] Son, J., & Kim, S. B. (2018). Academic paper recommender system using multilevel simultaneous citation networks. Decision Support Systems, 105, 24-33. https://doi.org/10.1016/j.dss.2017.10.011

[14] Zhao, S., & Sun, X. (2024). Enabling controllable table-to-text generation via prompting large language models with guided planning. Knowledge-Based Systems, 304, 112571. https://doi.org/10.1016/j.knosys.2024.112571

[15] Tuset-Peiró, P., Vázquez-Gallego, F., Muñoz, J., Watteyne, T., Alonso-Zarate, J., & Vilajosana, X. (2019). Experimental interference robustness evaluation of ieee 802.15. 4-2015 oqpsk-dsss and sun-ofdm physical layers for industrial communications. Electronics, 8(9), 1045. https://doi.org/10.3390/electronics8091045

[16] Gao, S., Young, M. T., Qiu, J. X., Yoon, H. J., Christian, J. B., Fearn, P. A., ... & Ramanthan, A. (2018). Hierarchical attention networks for information extraction from cancer pathology reports. Journal of the American Medical Informatics Association, 25(3), 321-330. https://doi.org/10.1093/jamia/ocx131

[17] Shi, P., Cui, Y., Xu, K., Zhang, M., & Ding, L. (2019). Data consistency theory and case study for scientific big data. Information, 10(4), 137. https://doi.org/10.3390/info10040137

[18] AbuRa'ed, A., Saggion, H., Shvets, A., & Bravo, À. (2020). Automatic related work section generation: experiments in scientific document abstracting. Scientometrics, 125(3), 3159-3185. https://doi.org/10.1007/s11192-020-03630-2

[19] Wang, J., Meng, F., Zheng, D., Liang, Y., Li, Z., Qu, J., & Zhou, J. (2022). A survey on cross-lingual summarization. Transactions of the Association for Computational Linguistics, 10, 1304-1323. https://doi.org/10.1162/tacl_a_00520

[20] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021). Video summarization using deep neural networks: A survey. Proceedings of the IEEE, 109(11), 1838-1863. https://doi.org/10.1109/JPROC.2021.3117472

[21] Guo, Y., Qiu, W., Wang, Y., & Cohen, T. (2021, May). Automated lay language summarization of biomedical scientific reviews. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 1, pp. 160-168). https://doi.org/10.1609/aaai.v35i1.16089

[22] Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2021). A novel deep neural network-based approach to measure scholarly research dissemination using citations network. Applied Sciences, 11(22), 10970. https://doi.org/10.3390/app112210970

[23] Lu, Y., Yuan, M., Liu, J., & Chen, M. (2023). Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion. Scientometrics, 128(2), 1367-1393. https://doi.org/10.1007/s11192-022-04566-5

[24] Kuang, L., Ge, F., & Zhang, L. (2022). Suggesting method names based on graph neural network with salient information modelling. Expert Systems, 39(6), e13030. https://doi.org/10.1111/exsy.13030

[25] Su, W., Jiang, J., & Huang, K. (2023). Multi-granularity adaptive extractive document summarization with heterogeneous graph neural networks. PeerJ Computer Science, 9, e1737. https://doi.org/10.7717/peerj-cs.1737

[26] Luo, Z., Xie, Q., & Ananiadou, S. (2024). Factual consistency evaluation of summarization in the Era of large language models. Expert Systems with Applications, 254, 124456. https://doi.org/10.1016/j.eswa.2024.124456

[27] Barsha, F. A. H., & Uddin, M. N. (2023, September). Comparative analysis of banglat5 and pointer generator network for bengali abstractive story summarization. In 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 84-88). IEEE. https://doi.org/10.1109/ICICT4SD59951.2023.10303633

[28] Xie, Q., Bishop, J. A., Tiwari, P., & Ananiadou, S. (2022). Pre-trained language models with domain knowledge for biomedical extractive summarization. Knowledge-Based Systems, 252, 109460. https://doi.org/10.1016/j.knosys.2022.109460

[29] Patel, D., Shah, S., & Chhinkaniwala, H. (2019). Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. Expert Systems with Applications, 134, 167-177. https://doi.org/10.1016/j.eswa.2019.05.045

[30] Kumar, S., Solanki, A., & Jhanjhi, N. Z. (2025). ROUGE-SS: A new ROUGE variant for the evaluation of text summarization. Recent Advances in Computer Science and Communications, 18(6), E060624230748. https://doi.org/10.2174/0126662558304595240528111535