Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
6(1): 58–78.
DOI: 10.3934/aci.2026004
Received: 10 February 2026
Revised: 25 March 2026
Accepted: 02 April 2026
Published: 09 April 2026

Research article

Efficient summarization with lightweight LLMs through sparse input activation and adaptive prompting

Srinivasan Subramanian, Ramya Madhuri Narapureddy, Mohana Priya Palanisamy, Kazi Aminul Islam and Md. Abdullah Al Hafiz Khan*

Department of Computer Science, Kennesaw State University, 1100 South Marietta Pkwy SE, Marietta, GA 30060, USA

* **Correspondence:** Email: mkhan74@kennesaw.edu.

Academic Editor: Jidong Yang

Abstract: Large language models (LLMs) are designed to read, reason, and generate natural language, enabling improved access to information, problem solving, and communication. However, their performance depends strongly on prompt design and model scale. Lightweight LLMs (fewer than 5B parameters) often struggle to summarize complex technical documents such as research publications due to domain-specific terminology, dense citations, and mathematical notation, which can increase hallucinations and reduce reliability. We have introduced a hybrid framework that combines Natural Language Processing (NLP) driven preprocessing techniques, sparse input activation, and prompt engineering to enhance the summarization capacity of small-scale models by minimizing hallucinations and maximizing factual accuracy. The pipeline starts with cleaning and segmenting full articles into sections, removing references, citations, and formulas. On this normalized output, salient sentence extraction and keyphrase extraction are performed as part of the sparse input activation module. The activated input and optimized prompt are fed to LLMs of different scales and the summaries are benchmarked with respect to different prompting strategies and model sizes. This workflow significantly enhances the factual accuracy and reliability of summaries generated by lightweight LLMs, making them competitive for complex scientific and technical summarization tasks. Our adaptive prompting and sparse input activation approach significantly boosted summarization quality in lightweight LLMs, improving average ROUGE-1 recall from about 32% to 46% (44% relative gain), human evaluation scores from 66% to 91% (39% relative gain), and coverage ratio from 39% to 60% (53% relative gain) over the baseline.

Keywords: lightweight large language models (LLMs); document summarization; sparse input activation (SIA); adaptive prompting; salient sentence extraction; keyphrase extraction; factual accuracy

1. Introduction

Text summarization enables faster comprehension of large volumes of information by generating concise versions of documents while preserving key ideas. As the pace at which digital information is produced continues to accelerate, individuals and organizations face increasing difficulty in processing long and complex documents. Summarization helps reduce cognitive effort and allows users to focus on essential insights rather than sifting through extensive text. As a result, summarization has become a vital component in fields such as education, research, journalism, and enterprise systems where rapid comprehension is important for decision-making and productivity. Large language models (LLMs) are effective in text summarization and can generate summaries that help the users understand the core concepts without having to read through the entire document [9]. Their ability to understand language structure, identify salient points, and maintain coherence has pushed automatic summarization beyond traditional rule-based or statistical methods [17].

However, this comes with a cost. Their performance heavily depends on the effectiveness of the prompt given and the size of the model. Crafting suitable prompts is often non-trivial, especially when dealing with domain-specific or highly technical inputs, and even small variations can significantly affect output quality [6, 8, 10]. At the same time, the most powerful models, such as state-of-the-art proprietary foundation models, demand massive computational resources that make them impractical to use in edge devices or local hardware. This creates a performance gap between lightweight LLMs and large-scale LLMs. Additionally, large models introduce concerns related to cost, latency, privacy, and accessibility, making it difficult for smaller organizations or personal users to rely on them for daily tasks. Thus, although state-of-the-art summarization quality is achievable, it is not always attainable under realistic computational constraints.

One of the most challenging applications for smaller models is summarizing complex, technical documents such as research papers or technical reports. These texts are long, filled with technical notations and formulas that require deep context understanding. Such technical research papers frequently include multi-layered arguments, domain-specific abbreviations, and complex cross-sectional logical dependencies, which can challenge smaller language models. Prior studies report that lightweight models are more sensitive to prompt design and more prone to factual inconsistencies and hallucination when handling dense or technical inputs [6, 17], often resulting in incomplete or inaccurate summaries as illustrated in Figure 1. Such inaccuracies can be particularly harmful when the information is used for academic study, engineering design, or scientific interpretation. Meanwhile, larger models handle these tasks well because they are trained with billions of parameters, which provide broader context [17]. The gap lies not only in training but also in access to the large-scale computing resources during inference. This makes it clear that improving the way small models utilize their limited capacity is as important as improving the models themselves. If smaller models can be guided to focus their limited attention on the most important parts of the text, we can make them far more capable without increasing their size [3–5]. This idea reflects how humans process information: rather than reading every sentence with equal weight, skilled readers selectively attend to topic sentences, conclusions, keywords, and central arguments [5]. Adopting a similar strategy for LLMs can allow small models to produce summaries closer in quality to those of large models while avoiding unnecessary computational overhead. Such an approach not only benefits resource-constrained environments but also enhances efficiency in cloud-based systems by reducing input size

and inference cost.

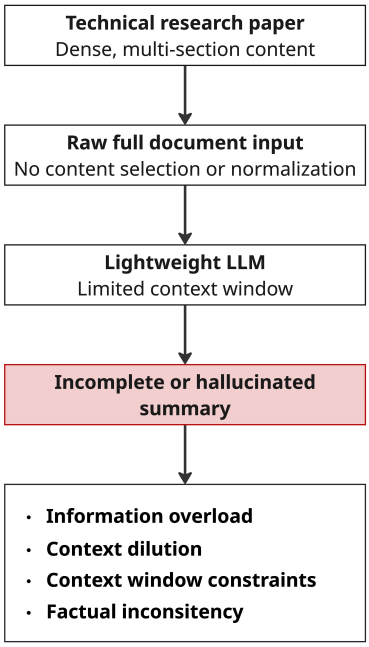

Figure 1. Illustration of the baseline summarization problem in lightweight LLMs.

This project tackles that challenge by introducing sparse input activation combined with adaptive prompt optimization. Before the model begins summarizing, we use NLP techniques to extract the most meaningful sentences, key phrases, and findings from the document. These high-value segments are selectively passed to the model, while the prompt is structured to guide its attention toward what truly matters [8]. By reducing exposure to irrelevant or low-value content, the model devotes more of its limited reasoning ability to understanding central ideas. This architecture also reduces the likelihood of hallucination and off-topic generation. The system is evaluated on open-access academic datasets and benchmarked against larger-scale LLMs, allowing us to measure improvements in both precision and robustness under constrained conditions.

In doing so, this work contributes to a practical way to make small LLMs perform better on summarization tasks by improving accuracy, consistency, and focus without the heavy cost of large-scale training or hardware. It demonstrates how targeted input activation and adaptive prompting can effectively bridge the performance gap, enabling smaller models to handle complex summarization tasks that traditionally require high-capacity LLMs. This approach ultimately supports broader deployment of AI tools in real-world environments where computational resources are limited. Our evaluation shows that the proposed approach improves summarization quality and consistency for small LLMs, narrowing the performance gap with larger models while reducing hallucination and irrelevant outputs. Experiments show consistent gains in ROUGE-1 recall, coverage, and human evaluation scores compared to baseline lightweight summarization.

To address these limitations, we introduce a hybrid framework with the following contributions:

- **Sparse Input Activation as Inference-Time Attention:** We propose sparse input activation (SIA) as an inference-time selective attention mechanism that maximizes information density under limited context constraints. Unlike prior preprocessing approaches, SIA explicitly

compensates for the limited representational capacity of lightweight LLMs by prioritizing high-salience and low-redundancy content.

- **Input Prompt Optimization Strategy:** We design a structured prompting approach incorporating role specification, context injection, and chain-of-thought reasoning to improve factual grounding and coherence.
- **Comprehensive Multi-Metric Evaluation Protocol:** We evaluate the proposed framework on academic datasets using controlled experiments and ablation studies to show that SIA improves factual grounding and coverage, while adaptive prompting enhances lexical retention and coherence, together yielding substantial gains across ROUGE, the coverage ratio, and human evaluation.

2. Related work

Large language models (LLMs) have achieved strong performance on summarization tasks, but their effectiveness often relies on large model sizes and substantial computational resources. Lightweight LLMs are more practical for local and edge environments, yet they struggle with summarizing long, technical documents due to limited context capacity, prompt sensitivity, and hallucinations. Existing research addresses this challenge through content selection techniques [1, 2], long-context transformer architectures such as Longformer and BigBird [3, 4], and inference-time prompting strategies [6, 7, 9, 10]. However, architectural approaches typically require specialized attention mechanisms or modified training procedures [3, 4], while prompting-based techniques remain constrained by input length and context quality, particularly for long technical documents [6, 8, 17].

Prior work on content selection and input reduction aims to identify salient information before summarization. Methods such as TextRank rank sentences based on graph connectivity [1], while KeyBERT extracts semantically important key phrases using contextual embeddings [2]. These techniques reduce input length and highlight important content, and recent work shows that careful content selection and prompt compression can improve small-model summarization efficiency [5]. However, most approaches treat content selection as a preprocessing heuristic rather than as a mechanism to explicitly compensate for the limited attention capacity of small LLMs at inference time. In parallel, architectural approaches such as Longformer and BigBird introduce sparse and global attention patterns to handle long documents efficiently [3, 4], but they require specialized architectures or training pipelines that are impractical for lightweight model deployment.

A complementary line of work focuses on inference-time prompting and robustness. Calibration techniques reduce prompt sensitivity [6], AutoPrompt demonstrates systematic trigger token discovery [7], and structured prompting methods such as chain-of-thought and chain-of-density improve reasoning and information density in generated outputs [8, 10]. While these techniques improve generation quality, they remain constrained by the quality and length of the input context and do not directly address information overload in technical documents. Retrieval-augmented generation (RAG) grounds outputs using external evidence [11], and pretrained summarization frameworks such as BART and PEGASUS show that effective summarization can emerge from partial or compressed inputs [12, 13]. However, RAG typically relies on external retrieval infrastructure and does not explicitly target document-internal content selection for small models. Our work bridges these gaps by combining sparse input activation with adaptive prompting to simulate selective attention at inference

time, enabling lightweight LLMs to generate more accurate, complete, and reliable summaries of technical documents without modifying model architecture or requiring additional training.

3. Proposed framework for efficient summarization with lightweight LLMs

Figure 2 illustrates the proposed efficient summarization framework, designed to improve summarization quality for lightweight LLMs by combining structured preprocessing, sparse input activation, adaptive prompting, and multi-method evaluation. The overall workflow is divided into four sequential stages, each contributing to the factual accuracy, coherence, and completeness of the final generated summary.

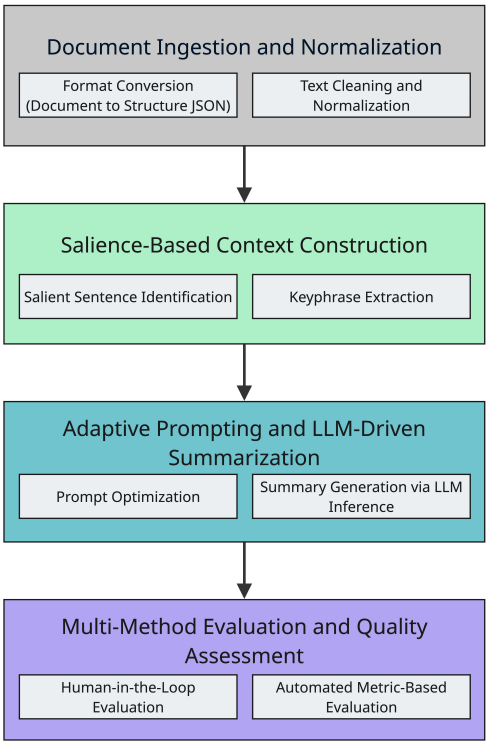

Figure 2. Overview of the proposed four-stage summarization framework.

3.1. Document ingestion and normalization

As shown in Figure 2, the framework begins by transforming raw research papers in PDF format into a structured, machine-readable format that preserves the hierarchical structure of sections and text. We utilized SciPDF Parser [19], an open-source tool designed for scientific documents that preserves section boundaries, paragraph order, and reading flow across two-column layouts. Following this, a rigorous cleaning and normalization stage is applied. This step removes non-essential elements such as in-text citations, section numbers, formulas, and extensive reference blocks, reducing noise that could interfere with smaller language models [14, 16] and ensuring that only meaningful content is retained.

3.2. Salience-based context construction

Once the documents are normalized, the system extracts salient sentences to reduce the text volume while retaining critical information. The process begins with term frequency and inverse

document frequency (TF-IDF) computation, which identifies words that are most representative of the document [20]. Cosine similarity is then calculated pairwise between all sentences to construct a similarity matrix, which is treated as a weighted graph where each sentence is a node. The TextRank algorithm is applied to this graph to assign relevance scores to each sentence, generating an initial measure of importance [1]. To further refine the selection and ensure diversity, maximal marginal relevance (MMR) is employed [21]. MMR iteratively selects sentences that balance high relevance, based on TextRank scores, with low redundancy, using the objective

$$\mathrm{MMR}(i) = \arg\max_{i \in R} \Big[\lambda \cdot \mathrm{STR}(i) - (1 - \lambda) \cdot \max_{j \in S} \mathrm{Sim}(i, j) \Big],$$

where S is the set of already selected sentences, R is the remaining candidate set, $\mathrm{STR}(i)$ is the TextRank score, $\mathrm{Sim}(i, j)$ is the cosine similarity between sentences, and $\lambda = 0.7$ biases the selection toward relevance while penalizing redundancy. This formulation can be interpreted as a constrained optimization problem where the objective is to maximize information density under a limited context budget. By selecting sentences that jointly optimize relevance and diversity, sparse input activation effectively increases the signal-to-noise ratio of the input provided to the LLM. This is particularly important for lightweight models, where limited attention capacity and smaller parameterization make them more sensitive to irrelevant or redundant content. The final set of the K sentences forms the sparse input activation pool. We select the top K sentences (K is determined as 30% of document length with a minimum of 20). These sentences are then passed through BERTSum [22], a pre-trained transformer model, which analyzes the content in context to refine the selection of sentences that best represent the document's main ideas. Keyphrase extraction is performed using contextual embedding–based ranking via KeyBERT, which identifies semantically salient phrases by leveraging transformer-based embeddings, aligning with prior statistical–semantic keyword extraction approaches such as Shah and Fränti [23].

3.3. Adaptive prompting and LLM-driven summarization

After preparing the salient content, the system orchestrates the generation of final summaries using a lightweight locally deployed LLM. For each paper, the system aggregates the normalized abstract, conclusion, BERTSum [22] outputs, and extracted keyphrases from the previous stage.

In this work, adaptive prompting refers to *input-conditioned prompt construction*, where the overall prompt structure remains fixed, but the injected content varies dynamically across documents. Specifically, document-specific elements such as salient sentences, keyphrases, and section-level summaries are incorporated into the prompt, resulting in context-sensitive inputs that adapt to the characteristics of each paper.

The constructed prompt consists of three components: (1) a role instruction that defines the model as an expert academic research assistant responsible for producing accurate and coherent summaries grounded strictly in the provided content, (2) structured chain-of-thought (CoT) guidance that directs the model to analyze the abstract and conclusion, synthesize salient sentences, and integrate keyphrases, and (3) a context injection module containing the document-specific inputs.

Additionally, a one-shot example is included to guide the desired narrative flow, tone, and structure without enforcing content reuse. The model is instructed to generate a cohesive executive summary in a continuous narrative format, avoiding bullet points or sectional headings.

This combination of role conditioning, reasoning guidance, and input-adaptive context injection enables the LLM to generate high-quality, context-aware executive summaries as illustrated in Figure 3 while reducing the chances of hallucinations and preserving factual consistency.

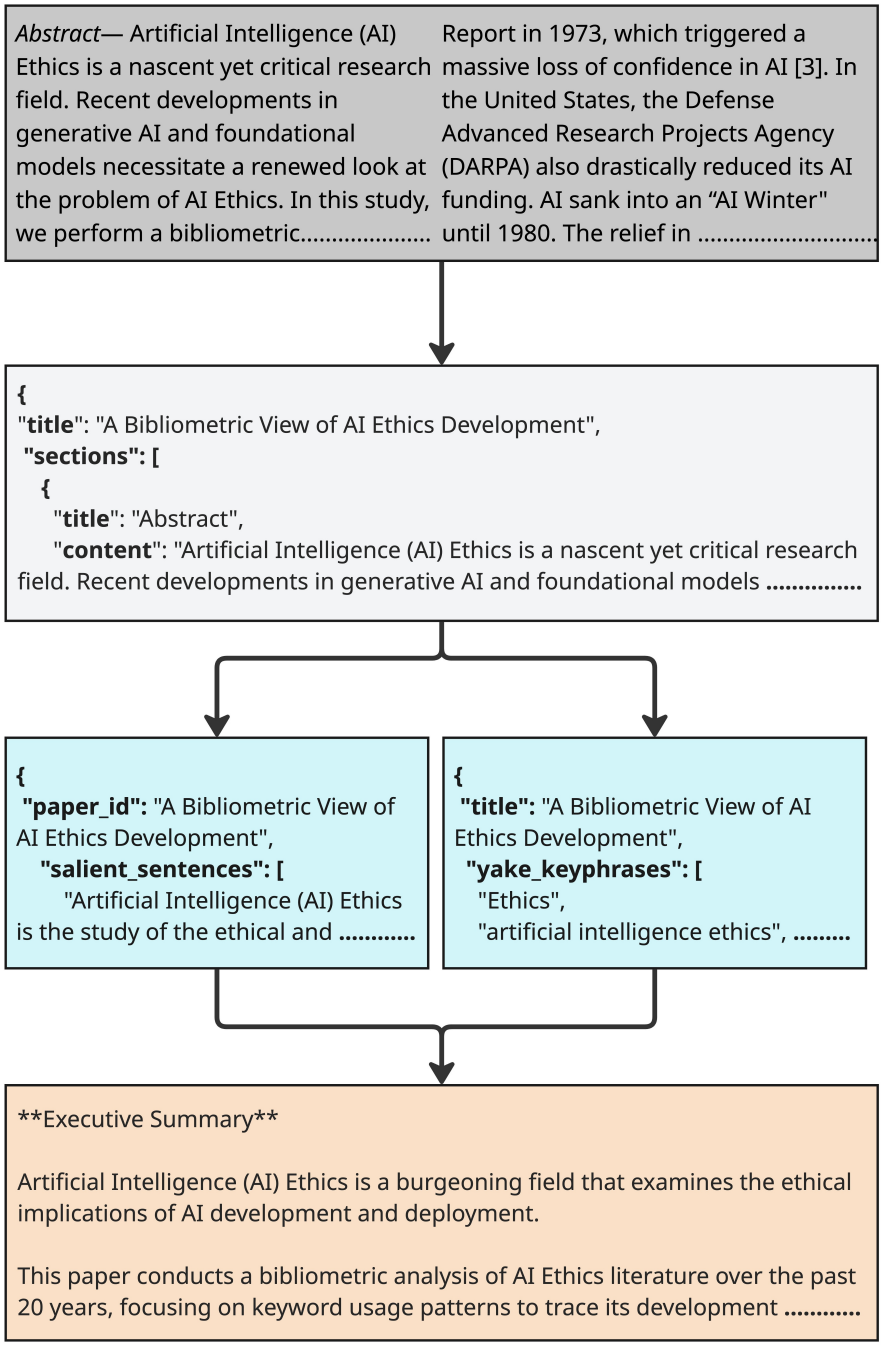

Figure 3. Illustration of stage-wise text transformation across the sequential stages.

The final prompt provided to the LLM follows a structured, multi-component design consisting of role conditioning, reasoning guidance, and context injection. The overall structure is shown below:

System Instruction: You are an expert academic research assistant specializing in

synthesizing complex technical papers into an accessible, accurate, and comprehensive executive summary. The summary must be professional, well-structured, and strictly grounded in the provided contextual data. Do not introduce external information or speculation. The output should be a cohesive, single-block executive summary not exceeding three paragraphs, without sectional headings or bullet points.

User Prompt:
Chain-of-Thought and Context Injection Instructions:

(1) **Analysis:** Review the abstract and conclusion to identify the core research problem, primary gap, and main contribution.
(2) **Synthesis:** Integrate extracted keyphrases to ensure coverage of major thematic elements.
(3) **Grounding:** Use salient sentences as factual anchors to construct a faithful and accurate summary.
(4) **Formatting:** Follow the narrative flow demonstrated in the one-shot example while generating a unique summary for the target paper.

One-Shot Example:
Title: The Role of Quantum Computing in Differential Privacy in IoT Networks

This paper investigates the inherent conflict between high utility data-sharing and robust privacy protection within IoT environments, proposing that traditional cryptographic methods are insufficient against emerging quantum-enabled threats. The authors focus on how differential privacy (DP) mechanisms fail to scale efficiently in high-velocity, decentralized IoT networks, leading to a "utility cliff" where privacy gains erode data usefulness. To address this, the research introduces a novel quantum-resistant differential privacy (QRDP) framework. The methodology combines homomorphic encryption for query processing with a quantum-hardened noise injection protocol guided by machine learning optimization. Experimental results demonstrate a 35% improvement in query response time over classical DP approaches. The primary contribution is a proof-of-concept system that balances quantum-level security with data utility for IoT applications. The work concludes that QRDP protocols provide a viable pathway for regulatory compliance without sacrificing analytical value.

Context Injection:

- **Abstract:** {abstract}
- **Conclusion:** {conclusion}
- **Salient Sentences:** {top-k sentences}
- **Keyphrases:** {keyphrases}

Final Task: Generate the final executive summary for the target paper based on the instructions and context above. Begin the response with the title: **Executive Summary**.

The prompt structure remains fixed across documents, while the injected content (abstract, conclusion, salient sentences, and keyphrases) varies dynamically, enabling input-adaptive summarization. This prompt is instantiated separately for each document, resulting in variations in content composition while preserving a consistent reasoning framework.

3.4. Multi-method evaluation and quality assessment

Finally, the evaluation stage assesses summary quality using both automated metrics and human judgment. The summaries that are generated are evaluated against reference summaries produced by a large language foundation model that serves as a high-quality benchmark. The summaries generated in this study are evaluated against reference summaries produced by the large language model Gemini 2.5 Flash, which serves as a strong benchmark for high-quality summarization. The model was selected due to its strong performance on long-form and technical document summarization tasks. Its large context window enables processing of entire scientific papers without aggressive truncation or chunking. The reference summaries are generated using stochastic decoding, with the temperature set to 1.0, top-p (nucleus sampling) set to 0.95, and top-k sampling set to 40. This configuration enables controlled diversity while maintaining coherence and relevance in summarizing complex scientific content.

The input to the model consists of the full normalized paper text, constructed by concatenating section titles and corresponding content. A zero-shot prompting strategy is employed, where the model is instructed to generate an executive summary of the document without any task-specific prompt engineering, few-shot examples, or adaptive prompting mechanisms. To ensure consistency across evaluations, all reference summaries are generated in a single run and cached. This guarantees that the same set of summaries is used throughout all experiments, avoiding variability during metric computation despite the stochastic decoding process.

It is important to note that these reference summaries are not treated as absolute ground truth. Instead, they function as a high-quality comparative baseline, a widely adopted practice in recent summarization research when large-scale human-authored references are unavailable.

To quantitatively assess the quality of the summaries, multiple evaluation metrics are employed. ROUGE scores are calculated to measure the overlap of n-grams between the generated summaries and the reference summaries, providing an indication of content similarity. Additionally, the BERTScore is computed, leveraging contextual embeddings to capture semantic similarity beyond surface-level word overlap [15]. Beyond comparison with references, coverage rates are also measured to determine how much of the original content from the normalized papers is captured by the summaries. Finally, human evaluation was conducted to assess summary readability, coherence, and factual alignment with the source document.

4. Experimental setup

4.1. Dataset description

The dataset contains 100 IEEE research papers in Computer Science, selected to reflect long, information-dense technical writing suitable for abstractive summarization. Each paper includes standard sections such as the abstract, introduction, methods, results, and conclusion. They also contain domain-specific terms, formulas, figures, tables, and references. This dataset is designed to test small LLMs (around 3 billion parameters) on their ability to maintain factual accuracy, technical coherence, and overall robustness.

Although the dataset is restricted to computer science research papers, it was carefully curated to capture substantial diversity in content and structure. The selected papers span multiple subdomains,

including artificial intelligence, systems, data science, and theoretical computing. Furthermore, the dataset includes a mix of mathematically intensive papers with formal notation, conceptually dense theoretical works, and application-oriented studies with descriptive narratives. Document lengths also varied significantly (mean: 3922 tokens ± 2454) as shown in Table 1, ensuring evaluation across both short and long documents.

Table 1. Dataset statistics.

Metric	Value
Total tokens (all papers)	392,229
Average tokens per paper (mean ± SD)	3922 ± 2454
Median tokens per paper	2926
Average sections per paper	13.69

The dataset is used as a fixed evaluation benchmark rather than being split into training, validation, and test sets, as the proposed framework operates entirely at inference time without model fine-tuning. The selected papers were curated to ensure diversity, which enables evaluation across a wide spectrum of summarization challenges.

This intra-domain diversity ensures that the evaluation reflects a range of document complexities, writing styles, and structural patterns commonly encountered in technical literature. By maintaining a controlled domain, we isolate the impact of sparse input activation and adaptive prompting without confounding effects from domain shift.

However, we acknowledge that broader cross-domain validation (e.g., biomedical, legal, or interdisciplinary corpora) would further strengthen generalizability. We identify this as an important direction for future work.

4.2. Evaluation metrics

Our framework uses a comprehensive evaluation suite with four metrics (see Table 2) to check the reliability, accuracy, and completeness of the summaries generated by lightweight LLMs. First, we use ROUGE to measure lexical overlap, showing how well words, short phrases, and sentence structure match references. We use ROUGE-1 recall, the unigram ($N = 1$) instance of the ROUGE-N family, which quantifies the proportion of reference unigrams retained in the generated summary. Formally, it is defined as

$$\text{ROUGE-1 Recall} = \frac{\text{Number of overlapping unigrams}}{\text{Total number of unigrams in the reference summary}}.$$

Since it operates on token overlap, this metric primarily captures lexical and syntactic similarity rather than deeper semantic equivalence.

Table 2. Evaluation metrics.

No.	Evaluation Metrics	Description	Formulas
1	ROUGE Score	Surface-level lexical overlap	$\text{ROUGE-N}_{\text{recall}} = \frac{\sum \text{match}(n\text{-grams})}{\sum \text{reference}(n\text{-grams})}$
2	BERTScore	Semantic similarity	$\text{BERTScore} = \frac{1}{N}\sum_{i=1}^{N} \max_j \cos(\phi(x_i), \phi(y_j))$
3	Coverage Rate	Completeness	$\text{Coverage} = \frac{\text{Number of key facts preserved in summary}}{\text{Number of key facts extracted from source}}$
4	Coverage Ratio (%)	Relative completeness w.r.t. reference summary	$\text{Coverage Ratio}(\%) = \frac{\text{Coverage}_{\text{result}}}{\text{Coverage}_{\text{reference}}} \times 100$

Within established taxonomies of string similarity measures [18], ROUGE-1 recall is a token-level *n*-gram metric that captures lexical overlap rather than semantic equivalence. To go beyond surface overlap, we additionally evaluated summaries using semantic and content-oriented metrics, including BERTScore, soft recall, coverage ratio, and human evaluation. Soft recall is based on soft cardinality, following the formulation of Fränti and Mariescu-Istodor [24]. In our experiments, soft recall exhibited behavior highly similar to embedding-based evaluation metrics such as BERTScore, with limited additional discriminative power across summaries. Given this redundancy, we retain the BERTScore as the primary semantic evaluation metric. The BERTScore measures the true semantic similarity between the generated summary and the reference, ensuring the meaning is preserved even with different wordings. The coverage ratio measures how many key facts and important data points from the source are included, ensuring the summary is complete.

We note that reference summaries in our evaluation are generated using a large language foundation model rather than human-written gold summaries. While this enables scalable and consistent benchmarking across a large dataset, it may introduce stylistic bias. To mitigate this limitation, we complement automated evaluation with human assessment focusing on factual grounding, completeness, and coherence. The inclusion of human evaluation ensures that improvements are aligned with human judgment rather than solely with model-generated reference patterns.

Human evaluation was conducted by three graduate-level annotators with experience in reading and analyzing technical research papers, ensuring informed judgment of content quality. To ensure a structured and consistent evaluation process, each generated summary was assessed using a predefined rubric across four dimensions: (1) factual accuracy, (2) coverage of key contributions, (3) coherence, and (4) readability as detailed in Table 3. Each dimension was scored on a 0–25 scale, resulting in a total score on a 0–100 scale for each summary.

Table 3. Human evaluation rubric.

Rubric	**Description**
Factual Accuracy	Correctness and faithfulness to the source document
Coverage	Extent to which key contributions and findings are included
Coherence	Logical flow and organization of ideas
Readability	Clarity and ease of understanding

The final human evaluation score for each summary was computed by averaging the total scores across all annotators. To minimize bias, summaries from different methods were anonymized and presented in randomized order, ensuring that annotators were not aware of the underlying generation approach. While formal inter-rater agreement metrics such as Krippendorff's alpha were not computed, we observed consistent scoring trends across annotators. The reported scores reflect the averaged consensus across evaluators. This human-centered evaluation serves as an essential validation layer, ensuring that improvements observed in automated metrics reflect genuine gains in factual accuracy and interpretability rather than alignment with LLM-generated reference styles.

4.3. Baseline configuration

To ensure a fair and controlled comparison, we define a baseline summarization pipeline using the same lightweight LLM architecture (*HuggingFaceTB/SmolLM3-3B*) and identical hardware

configuration (4-bit quantization via BitsAndBytes) as the proposed system. The baseline differs only in the absence of sparse input activation (SIA) and adaptive prompt optimization.

In the baseline setup, the normalized document text is directly provided to the LLM using a standard summarization prompt without structured context injection, chain-of-thought reasoning, or salience-based filtering. No sentence ranking, keyphrase extraction, or selective activation mechanisms are applied.

This configuration allows us to isolate the impact of (1) sparse input activation and (2) adaptive prompt optimization by keeping the model architecture, dataset, and evaluation framework constant. All evaluation metrics, including ROUGE, BERTScore, the coverage ratio, and human evaluation, are computed identically for both the baseline and the proposed methods.

4.4. Implementation

The hybrid summarization framework was implemented entirely in Python, leveraging Google Colab for collaborative development and resource management, including data storage on Google Drive.

The system begins by transforming raw PDF papers into a structured, machine-readable format using the SciPDF library. This library accurately parses multi-column layouts, figures, and other complex academic formatting, producing JSON objects that preserve the hierarchical structure of sections and text. Following this, a rigorous cleaning and normalization stage is applied using Python's native regular expressions. The final set of K sentences, typically around 30% of the document (with a minimum of 20), forms the sparse input activation pool. These sentences are then passed through BERTSum [22], a pre-trained transformer-based extractive summarization model, which performs context-aware sentence scoring and refinement. Unlike earlier stages that rely on statistical importance (TF-IDF, TextRank) and diversity optimization (MMR), BERTSum [22] leverages deep contextual embeddings to model inter-sentence relationships and semantic coherence.

In our framework, BERTSum is used as a refinement layer rather than as a standalone summarizer. It re-evaluates the candidate sentence pool generated by TextRank and MMR, assigning contextual importance scores and selecting sentences that best represent the document's core ideas. This additional contextual filtering step improves the quality of the sparse input activation pool by prioritizing semantically meaningful and globally coherent content.

After preparing the salient content, the system orchestrates the generation of final summaries using a lightweight locally deployed LLM, HuggingFaceTB/SmolLM3-3B, loaded with 4-bit quantization via the BitsAndBytes library to efficiently utilize GPU resources. The same base LLM and hardware configuration were used across all experimental conditions to ensure comparability. For each paper, the system aggregates the normalized abstract, conclusion, BERTSum [22] outputs, and keyphrases extracted in the previous stage. The orchestration loop iterates over all normalized papers, ensuring that even smaller LLMs, when provided with preprocessed, sparsely activated, and contextually enriched input, can produce reliable executive summaries suitable for technical and academic documents.

4.5. Evaluation protocol

All experiments were conducted under a fixed evaluation protocol. Prompt templates and preprocessing parameters were defined prior to full-scale evaluation and were not iteratively tuned

on individual documents within the benchmark dataset. This ensures that results are not influenced by document-specific optimization and avoids data leakage.

The framework operates deterministically with fixed prompts and preprocessing steps, and therefore does not rely on stochastic training procedures. As a result, variability across runs is minimal.

Performance metrics were reported as averages across all documents, and we additionally analyzed per-document trends to confirm that improvements were consistent rather than driven by a small subset of samples.

5. Results

5.1. Performance evaluation

Our evaluation demonstrates that adaptive prompting combined with sparse input activation (SIA) consistently outperforms the baseline summarization approach across all major evaluation metrics.

All reported results represent averages across the dataset. In addition, we observe low variance between documents, indicating stable performance gains. Improvements are consistent across the majority of samples, as shown in the document-level analysis.

On average, our method improves ROUGE-1 recall from approximately 32% in the baseline to 46%, representing a relative gain of over 44%, indicating substantially better retention of salient content from the source documents (as shown in Figure 4). This improvement is consistent across individual documents, as shown by positive per-document ROUGE gains in nearly all cases (Figure 5).

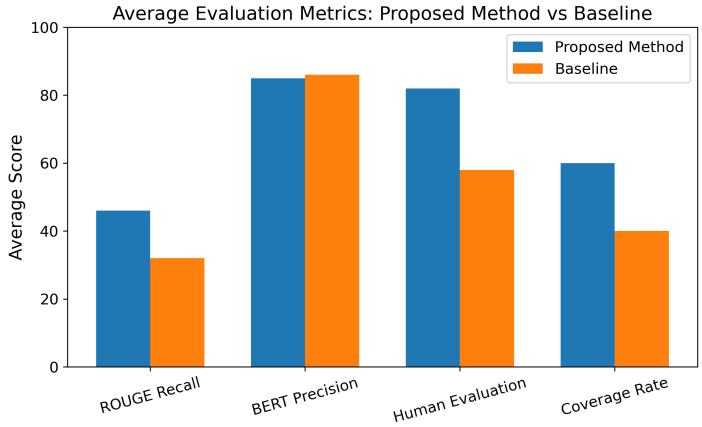

Figure 4. Average metric comparison.

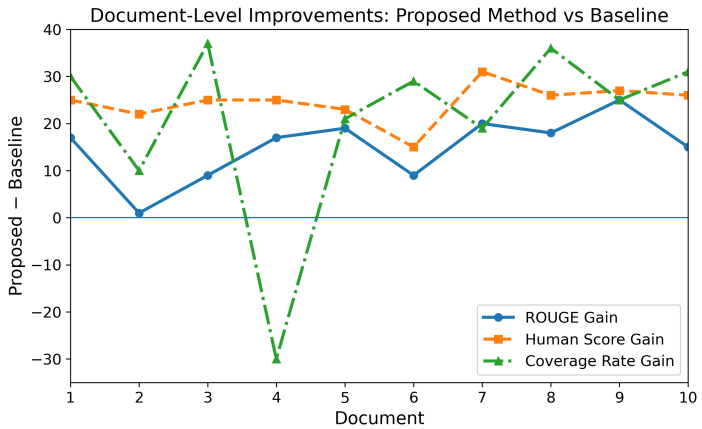

Figure 5. Document-level gains.

Information preservation improves even more significantly. The average coverage ratio (normalized against the reference summary) increases from 39% in the baseline to approximately 60%, representing a 53% relative improvement (Figure 4). Coverage trends closely follow those of the reference summaries, while baseline summaries frequently omit essential technical details (Figure 6). Per-document coverage analysis further confirms that the proposed method maintains stable gains across diverse documents, rather than improving only a small subset (Figure 6).

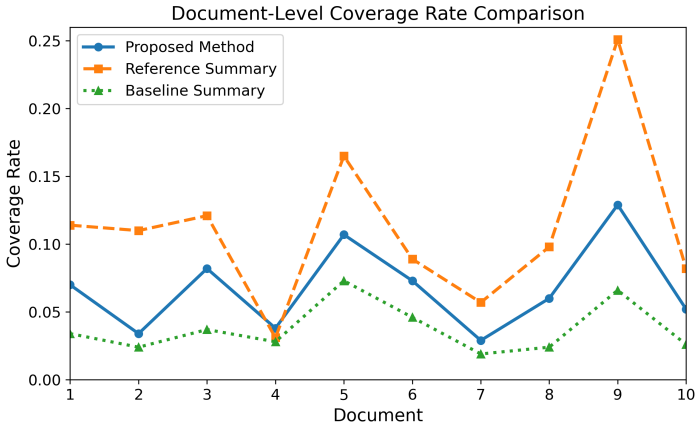

Figure 6. Document-level coverage comparison.

Human evaluation scores show the largest absolute improvement. Baseline summaries receive an average human score of 66%, whereas the proposed approach achieves an average human scores of 91%, yielding an approximate improvement of 25 percentage points (Figure 4). This indicates that summaries produced with adaptive prompting and SIA are not only more complete, but also more readable, coherent, and aligned with human expectations.

Despite these gains, semantic similarity remains stable. BERTScore F1 values stay consistently high across all configurations (85% to 87%), showing that improvements in coverage and recall do not come at the expense of semantic correctness (Figure 7). This stability confirms that the observed gains are driven by better content selection and guidance rather than semantic drift.

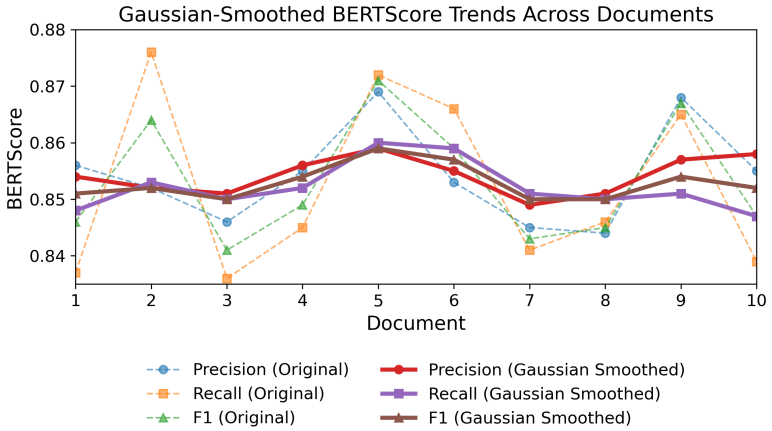

Figure 7. Gaussian-smoothed BERTScore trends.

To further evaluate factual consistency, we employ AlignScore, a reference-free evaluation metric that measures semantic alignment between source documents and generated summaries.

The proposed method achieves a higher AlignScore (0.8671 ± 0.0574) compared to the baseline (0.7827), corresponding to an absolute improvement of 0.0844. This indicates that summaries generated using sparse input activation exhibit stronger alignment with the source content.

The improvement is consistent across documents, suggesting that the proposed approach enhances factual grounding rather than relying on isolated gains. These results complement human evaluation findings and provide additional evidence that the method improves factual consistency while maintaining coherence and readability.

Overall, these results show that careful input activation and prompt optimization allow lightweight LLMs to approach the behavior of much larger models, achieving meaningful improvements in accuracy, completeness, and reliability without increasing model size or computational cost.

5.2. Ablation study and baseline comparisons

To ensure rigorous evaluation, we design a set of controlled baselines that isolate the contribution of each component in the proposed framework. These include direct summarization with a standard prompt, sparse input activation (SIA) without prompt optimization, and multiple prompt-only configurations. This setup enables a fair comparison under identical model, dataset, and hardware conditions, allowing us to attribute performance gains specifically to input activation and prompting strategies.

This ablation study evaluates the contribution of each pipeline component using both an automated lexical metric and human-centered evaluation (see Table 4). Lexical coverage is measured using ROUGE-1 recall, the unigram ($N = 1$) instance of the ROUGE-N family. In the remainder of this section, we refer to ROUGE-1 recall simply as recall. The recall scores indicate that the proposed method preserves a larger proportion of reference content compared to the baselines. Higher recall reflects improved coverage of key information from the source text. However, because recall measures token-level overlap, it does not fully account for paraphrasing or semantic reformulation.

Table 4. Ablation study and baseline comparison across SIA and prompting configurations.

Configuration	**ROUGE-1 (%)**	**Human Score (%)**
Baseline (Direct summarization)	32.0	66.5
SIA only (Sparse input activation)	43.0	82.7
Adaptive prompting only (Chain-of-thought prompting)	48.0	85.3
Adaptive prompting only (Instruction-tuned prompting)	46.5	82.5
Adaptive prompting only (Context injection + reasoning)	40.0	74.6
Proposed approach (SIA + CoT + context injection)	46.0	**90.9**

In addition to lexical coverage, we conduct human evaluation to assess semantic faithfulness, factual consistency, and overall readability. While recall estimates the extent to which reference content is preserved at the token level, the human score reflects qualitative semantic alignment and interpretability. These measures therefore capture complementary aspects of summarization quality rather than directly corresponding to classical precision and recall in information retrieval.

The baseline system achieves an average recall of 32%, reflecting limited overlap with reference

summaries. Introducing sparse input activation (SIA) alone yields a meaningful improvement, increasing average recall to 43%, indicating improved factual grounding and expanded token coverage compared to the baseline.

Adaptive prompting plays a dominant role in lexical retention. Among the prompt variants, the instruction tuning prompting strategy (Strategy 1) achieves an average recall of approximately 46.5%, demonstrating strong retention of salient tokens through structured task guidance. In contrast, the context injection and reasoning tuned prompting strategy (Strategy 2) achieves an average recall of approximately 40%, showing occasional strong peaks but weaker overall stability. When adaptive prompting is applied in its strongest standalone configuration, average recall increases further to approximately 48%, representing the highest lexical overlap observed among individual components.

The proposed pipeline reported in Section 5.1, which integrates chain-of-thought reasoning, context injection, and sparse input activation, achieves an average recall of approximately 46% while producing the highest human evaluation scores. Although adaptive prompting alone achieves the highest ROUGE-1 recall, the integration of sparse input activation prioritizes factual grounding and reduces redundant token overlap, leading to a slight decrease in lexical overlap while significantly improving qualitative human evaluation.

These results confirm that ROUGE improvements are driven primarily by adaptive prompting mechanisms, while SIA enhances factual grounding and contributes to improved qualitative alignment when incorporated within the structured prompting framework rather than functioning as a stand-alone enhancement.

An important observation from the ablation results is the divergence between lexical and human evaluation metrics. One of the prompt-only configurations achieved the highest ROUGE-1 recall, indicating strong token-level alignment with reference summaries. However, the full proposed method (SIA + adaptive prompting) achieved the highest human evaluation scores.

This behavior reflects a trade-off between lexical overlap and semantic quality. One of the prompt-only approaches tend to preserve more surface-level phrasing, which benefits ROUGE scores. In contrast, sparse input activation reduces redundancy and filters less relevant content, encouraging the model to produce more abstract, concise, and semantically coherent summaries.

As a result, the proposed method may slightly reduce n-gram overlap while improving factual grounding, readability, and overall interpretability. This suggests that ROUGE alone does not fully capture summarization quality, particularly for technical documents where abstraction and synthesis are important.

We note that extractive methods such as TextRank and BERTSum, while effective at identifying salient content, produce fragmented outputs that lack coherence and abstraction. In our framework, these methods are used as intermediate components within sparse input activation, and their standalone limitations highlight the necessity of LLM-based synthesis for high-quality summarization.

5.3. Error analysis

Although the proposed framework did not exhibit catastrophic hallucinations or complete failure cases during the evaluation, an in-depth inspection reveals several systematic trade-offs and boundary behaviors.

Sparse input activation (SIA) improves contextual efficiency by selecting salient sentences prior to summarization. However, this filtering stage may occasionally suppress low-frequency yet

conceptually important details. In documents where key contributions are distributed across multiple sections rather than concentrated in abstracts or conclusions, aggressive salience selection can lead to partial omission of nuanced findings. While such omissions did not significantly affect overall summarization, they reduced the fine-grained technical specificity in a few highly detailed papers.

The ablation study results indicate that the full pipeline achieves slightly lower ROUGE-1 recall compared to prompt-only configurations. This suggests that the proposed approach prioritizes semantic coherence and abstraction over lexical overlap. In several cases, summaries rephrased technical claims in more generalized language, improving readability but reducing n-gram alignment with the reference summary. This divergence reflects a metric-level trade-off rather than a semantic error. In addition, the ablation study highlights moderate prompt sensitivity. Variations in prompt structure and context-injection strategies led to noticeable differences in lexical retention and coverage, indicating that lightweight LLM performance remains partially dependent on inference-time prompt formulation. While the integrated pipeline stabilizes this effect, prompting continues to influence summarization quality.

In multi-layered technical sections containing detailed experimental pipelines or multi-step logical reasoning, the model occasionally compressed extended arguments into higher-level generalizations. Although the resulting summaries remained factually consistent, explanatory depth was reduced in comparison to the source.

During the evaluation study, no clear instances of fabricated facts or externally introduced content were identified. Instead, the observed limitations primarily involved selective omissions and abstraction-level compression of factual details, which represent controlled and acceptable trade-offs inherent to abstractive summarization.

6. Discussions and limitations

This work demonstrates that the summarization performance of lightweight large language models can be substantially improved through intelligent input structuring rather than increased model capacity. By integrating NLP-based preprocessing, sparse input activation (SIA), and adaptive prompt optimization, the proposed framework enables small-scale models to generate accurate, focused, and reliable summaries of long and technically dense research papers.

A key contribution of this project is the shift from a model-centric to an input-centric approach. Instead of exposing the LLM to the full document, the system selectively activates high-value content by extracting salient sentences and domain-relevant keyphrases after rigorous normalization. This allows the model to concentrate its limited representational capacity on the most informative segments of the text. From a high level, SIA can be viewed as an externalized attention mechanism that shifts part of the reasoning burden from the model to the input construction process, enabling lightweight LLMs to approximate behaviors typically associated with larger models. The adaptive prompt further guides reasoning by enforcing factual grounding and structured synthesis, reducing off-topic generation and hallucinations.

Our results confirm the effectiveness of this design. Compared to a baseline summarization pipeline, the adaptive prompting + SIA approach yields consistent improvements in ROUGE-1 recall, coverage rate, and human evaluation scores, while maintaining stable semantic similarity as measured by BERTScore. The coverage analysis shows that the proposed method retains a substantially larger

portion of key information, often closely tracking reference summaries produced by large models. These results indicate that targeted preprocessing and prompt engineering can close a substantial portion of the performance gap between lightweight and large-scale LLMs without additional training or architectural changes.

A key insight from this study is that optimizing for lexical overlap does not necessarily lead to better summarization quality. The combination of sparse input activation and adaptive prompting shifts the model toward semantic fidelity and structured abstraction, which aligns more closely with human expectations despite modest reductions in token-level metrics.

The framework is effective for technical and academic documents, where important information is distributed across structured sections such as abstracts, methods, and conclusions. By combining extracted salient sentences with keyphrases and section-aware prompts, the model produces coherent executive summaries that preserve core contributions and findings. The ability to run both locally and via APIs also makes the system practical for environments with constraints on cost, latency, or data privacy.

However, the approach has limitations. The quality of the final summary depends heavily on the accuracy of the preprocessing and sentence selection stages. Errors or omissions during sparse input activation can lead to incomplete summaries, even if the generated text remains fluent. Additionally, aggressive filtering can remove subtle contextual cues that are important for nuanced interpretation. The system is also evaluated primarily on computer science research papers, and its effectiveness in other technical domains remains to be validated. Finally, prompt templates and activation thresholds were designed manually, which may limit adaptability across diverse datasets.

Another limitation lies in the use of LLM-generated reference summaries for automated evaluation. While this enables scalable benchmarking, it may introduce bias toward model-specific summarization styles. However, this risk is mitigated through human evaluation, which confirms that the proposed method improves factual grounding and overall summary quality. Future work will incorporate human-written reference summaries and benchmark datasets to further strengthen evaluation robustness.

Another limitation of the current study is the relatively moderate dataset size (100 documents), which serves as a controlled benchmark rather than a large-scale corpus. However, the diversity of document types and the consistency of observed improvements provide confidence in the generalizability of the results.

Future work can address these limitations by introducing adaptive sentence selection strategies that respond to document length and structure, as well as by incorporating section-aware or retrieval-based weighting to improve coverage. Automating prompt optimization and extending the framework to other domains, such as biomedical or legal text, are promising directions. Overall, this project shows that with carefully designed input activation and prompting, lightweight LLMs can deliver reliable and high-quality summarization, making them viable for real-world deployment under practical constraints.

7. Conclusions

This project shows empirical evidence that sparse input activation and adaptive prompt optimization can significantly improve the summarization capabilities of lightweight LLMs, enabling them to handle complex scientific documents far more effectively within our evaluation setting. Rather than relying on model scale or heavy computation, our method strengthens how small models interpret and

prioritize information. The results indicate substantial gains in accuracy, focus, and factual consistency, showing that effective preprocessing and prompting strategies can close much of the performance gap between lightweight and large-scale models. In conclusion, this work shows that strong research-paper summarization does not require huge models or expensive hardware. With the right techniques, small LLMs can stay accurate, focused, and reliable, making AI summarization easier to use and more practical for everyday environments.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Md Abdullah Al Hafiz Khan is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article. The authors declare no conflict of interest.

References

1. R. Mihalcea, P. Tarau, TextRank: bringing order into texts, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, 404–411.
2. *M. Grootendorst, KeyBERT: a minimal keyword extraction tool using BERT embeddings*, GitHub/Zenodo repository, 2020. Available from: `https://github.com/MaartenGr/KeyBERT`.
3. I. Beltagy, M. Peters, A. Cohan, Longformer: the long-document transformer, arXiv: 2004.05150. https://doi.org/10.48550/arXiv.2004.05150
4. M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, et al., Big bird: transformers for longer sequences, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, 17283–17297.
5. T. Zhang, Y. Wang, D. Z. Wang, SCOPE: a generative approach for LLM prompt compression, arXiv: 2508.15813. https://doi.org/10.48550/arXiv.2508.15813
6. Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate before use: improving few-shot performance of language models, *Proceedings of the 38th International Conference on Machine Learning*, 2021, 12697–12706.
7. T. Shin, Y. Razeghi, R. Logeswaran, E. Wallace, S. Singh, Autoprompt: eliciting knowledge from language models with automatically generated prompts, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, 4222–4235. https://doi.org/10.18653/v1/2020.emnlp-main.346
8. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, et al., Chain-of-thought prompting elicits reasoning in large language models, *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, 24824–24837.
9. T. Kojima, S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, 22199–22213.

10. G. Adams, A. R. Fabbri, F. Ladhak, E. Lehman, N. Elhadad, From sparse to dense: GPT-4 summarization with chain of density prompting, arXiv: 2309.04269. https://doi.org/10.48550/arXiv.2309.04269
11. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, 9459–9474.
12. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, et al., BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
13. J. Zhang, Y. Zhao, M. Saleh, P. Liu, Pegasus: pre-training with extracted gap-sentences for abstractive summarization, *Proceedings of the 37th International Conference on Machine Learning*, 2020, 11328–11339.
14. C. Y. Lin, Rouge: a package for automatic evaluation of summaries, in: *Text summarization branches out*, Barcelona: Association for Computational Linguistics, 2004, 74–81.
15. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: evaluating text generation with bert, *Proceedings of ICLR*, 2020, 1–41.
16. A. R. Fabbri, C. S. Wu, W. Liu, C. Xiong, QAFactEval: improved QA-based factual consistency evaluation for summarization, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022, 2587–2601. https://doi.org/10.18653/v1/2022.naacl-main.187
17. Y. Liu, K. Shi, K. S. He, L. Ye, A. R. Fabbri, P. Liu, et al., On learning to summarize with large language models as references, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, 8647–8664. https://doi.org/10.18653/v1/2024.naacl-long.478
18. N. Gali, R. Mariescu-Istodor, D. Hostettler, P. Fränti, Framework for syntactic string similarity measures, *Expert Syst. Appl.*, **129** (2019), 169–185. https://doi.org/10.1016/j.eswa.2019.03.048
19. *T. Titipata, SciPDF Parser: a Python parser for scientific PDF files*, GitHub repository, 2022. Available from: `https://github.com/titipata/scipdf_parser`.
20. C. D. Manning, P. Raghavan, H. Schütze, *Introduction to information retrieval*, Cambridge: Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511809071
21. J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, 335–336. https://doi.org/10.1145/290941.291025
22. Y. Liu, M. Lapata, Text summarization with pretrained encoders, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, 3730–3740. https://doi.org/10.18653/v1/D19-1387
23. H. Shah, P. Fränti, Combining statistical, structural, and linguistic features for keyword extraction from web pages, *Applied Computing and Intelligence*, **2** (2022), 115–132. https://doi.org/10.3934/aci.2022007

24. P. Fränti, R. Mariescu-Istodor, Soft precision and recall, *Pattern Recogn. Lett.*, **167** (2023), 115–121. https://doi.org/10.1016/j.patrec.2023.02.005

AIMS Press