A Modular Hybrid Transformer Framework for Trustworthy Radiology Report Summarization

A Modular Hybrid Transformer Framework for Trustworthy Radiology Report Summarization

Setareh Pourgholamali[1], Edoardo Patti[2] and Alessandro Aliberti[3]

***Abstract*— Transformer-based language models have demonstrated strong potential for clinical text summarization, yet their applicability in radiology remains limited by hallucination, loss of critical diagnostic content, and inconsistency with reporting conventions. To address these challenges, this work proposes a modular hybrid framework combining extractive sentence scoring, domain-aware clinical entity filtering, and transformer-based abstractive generation. The system is evaluated on the Indiana University Chest X-ray dataset, using paired *Findings* and *Impression* sections as supervised targets. Quantitative results indicate substantial improvements over an extractive baseline, with higher ROUGE and BERTScore values, while qualitative inspection suggests stronger semantic alignment, clearer organization of clinical abnormalities, and reduced speculative phrasing. The findings indicate that architectural modularity and domain filtering provide structured mechanisms for constraining generative behavior and improving controllability in medical summarization. Overall, the approach contributes to ongoing efforts toward safe and clinically coherent Natural Language Processing systems for high-stakes healthcare applications.**

I. Introduction

Transformers and large language models (LLMs) have significantly advanced natural language processing (NLP), demonstrating strong performance in tasks such as summarization, question answering, and controlled text generation [1]. Architectures including BERT [2], T5 [3], and BART [4] have become foundational components for downstream applications requiring contextualized representations and high-quality generation. More recently, domain-specific variants such as BioBERT [5], ClinicalBERT [6], BioGPT [7], and other biomedical LLMs have emerged to better model terminology, structure, and discourse conventions in clinical narratives.

These advances are particularly relevant to radiology report summarization, specifically generating concise *Impression* sections from detailed *Findings*. The Findings-to-Impression task requires not only linguistic fluency but also domain knowledge, semantic reasoning, and preservation of critical diagnostic information [8], [9]. Although fine-tuned transformer-based models can produce fluent summaries, prior work has documented persistent issues with hallucination, factual inconsistency, omission of clinically relevant content, and unsupported diagnostic statements [10], [11]. These challenges are amplified in high-stakes medical settings, where reliability and traceability are as important as linguistic quality.

To overcome these challenges, recent research has investigated mechanisms such as extractive anchoring [12], constrained decoding [13], named-entity filtering [14], and hybrid summarization strategies combining extractive and abstractive modelling [15]. Hybrid pipelines are particularly promising, as they provide a means to preserve factual grounding while enabling fluent abstraction. However, existing approaches are often implemented as monolithic systems and lack systematic analysis of how intermediate processing steps influence performance, interpretability, and factual correctness.

In this work, we propose a controlled hybrid summarization framework that integrates three key stages: i) extractive sentence ranking based on BERTSUM to anchor content selection, ii) domain-aware named-entity filtering using biomedical NLP tools to prioritize clinically meaningful information, and iii) abstractive generation via a fine-tuned BART model. Unlike previous works that treat the pipeline as an indivisible system, our approach explicitly evaluates component contributions through incremental experiments. The goal is not only to improve the quality of the summary in the downstream, but also to provide information on how modular processing strategies affect reliability and hallucination control in medical summaries.

Experimental evaluation of the Indiana University Chest X-ray data set demonstrates that the proposed hybrid architecture improves both lexical and semantic quality over extractive-only and fully abstractive baselines. In particular, incorporating structured intermediate processing reduces hallucination and increases alignment with reference impressions, suggesting that modular design can support more controllable and dependable summarization pipelines. These findings contribute to ongoing efforts to develop safe and deployable NLP systems for clinical decision support and more broadly motivate modular design principles for high-precision text generation in specialized domains.

The remainder of this paper is organized as follows. Section II provides an overview of previous work in clinical NLP, with a focus on transformer-based summarization, factual preservation, and hybrid generation strategies. Section III details the proposed modular framework, including the extractive content selection stage, domain-aware entity filtering, and controlled abstractive generation component. Section IV outlines the experimental setup, characteristics of

*This work was not supported by any organization

[1]Setareh Pourgholamali, Department of Control and Computer Engineering, Politecnico di Torino, Italy setareh.pourgholamali@polito.it

[2]Edoardo Patti, Department of Control and Computer Engineering, Politecnico di Torino, Italy edoardo.patti@polito.it

[3]Alessandro Aliberti, Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, Italy alessandro.aliberti@polito.it

the data set, evaluation metrics, and reports the quantitative and qualitative results obtained from the comparative study. Section V interprets the findings, analyzes the methodological implications, discusses limitations, and considers the potential for generalization within and beyond radiology reporting. Finally, Section VI summarizes the main contributions and outlines directions for future research, including opportunities for broader clinical validation and enhanced controllability mechanisms in medical text generation.

II. Related Work

A. Transformer-Based Medical NLP

Transformer architectures have become the foundation of modern NLP since the introduction of self-attention mechanisms [1]. Large-scale pre-trained language models such as BERT [2], GPT [16], T5 [3], and BART [4] have demonstrated strong performance across a wide range of downstream tasks, including summarization, question answering, and controlled text generation. Their success stems from extensive pretraining on massive unlabeled corpora followed by targeted fine-tuning, enabling flexible adaptation to new domains.

Due to the linguistic complexity, domain-specific terminology, and structured semantic patterns present in biomedical narratives, general-purpose models have been extended through domain-adaptive pretraining. Biomedical variants such as BioBERT [5], ClinicalBERT [6], PubMedBERT [17], BioGPT [7], and PEGASUS-based summarization models [9] incorporate exposure to biomedical literature and clinical corpora. These adaptations significantly improve performance on specialised tasks such as entity recognition, extraction of clinical concepts, and radiology summarization. Recent studies applying fine-tuned BART or PEGASUS models to radiology reporting tasks have demonstrated fluent text generation; however, evaluations also reveal persistent issues with unsupported inferences, omission of critical concepts, and variability in diagnostic phrasing [10], [11]. This highlights a gap between linguistic fluency and clinically dependable reasoning.

B. Factuality and Hallucination Control in Medical Text Generation

Ensuring factual accuracy in neural generation has become a central research challenge, especially in safety-critical domains. Previous work has explored multiple strategies for constraining model behavior. Extractive grounding approaches restrict generative output to content present in the source text [12], while fact-checking and faithfulness scoring mechanisms attempt to detect unsupported statements in post-processing stages. Other work has focused on modifying the decoding process itself, including constrained sampling frameworks such as Metropolis–Hastings decoding [13], rule-guided generation, or ontology-restricted vocabularies.

Domain-aware preprocessing, particularly ontology-based named entity recognition and filtering, has shown promise in ensuring that essential clinical concepts are retained [17]. However, many of these strategies remain ad hoc or partial solutions rather than components of an integrated architectural design. Moreover, most studies evaluate factuality only through small-scale human review rather than controlled component-level experimentation, leaving open questions about how and where factual drift arises within transformer-based systems.

C. Hybrid Extractive–Abstractive Summarization

Hybrid summarization methods aim to balance factual precision with fluent reasoning by combining extractive selection and abstractive reformulation. Prior research in biomedical summarization shows that multi-stage architectures can outperform both extractive-only and end-to-end abstractive systems, particularly in radiology and electronic health record summarization [15], [18]. These approaches typically include a content selection stage followed by a generative rewriting step that compresses or restructures selected text while preserving clinical meaning.

However, many existing hybrid frameworks operate as monolithic pipelines in which intermediate outputs are neither analyzed nor controlled, limiting interpretability. Architectural decisions, such as applying sentence selection before or after semantic filtering, are often empirically motivated rather than grounded in explicit modeling rationale. Consequently, the individual contributions of extractive anchoring, domain filtering, and generative abstraction to factual reliability and stylistic fidelity remain insufficiently characterized [8].

Summary and Research Gap

The literature demonstrates that transformer-based models and hybrid approaches are promising for clinical summarization, yet fundamental limitations remain. Fully generative models lack mechanisms to enforce factual grounding, while existing hybrid architectures treat modular stages as implementation conveniences rather than systematically evaluated design choices. This makes it difficult to determine whether improvements in summary quality arise from extractive anchoring, domain filtering, the generative model itself, or interactions among these components.

This gap motivates the present work, which introduces a modular hybrid framework designed to explicitly separate i) content selection, ii) domain relevance filtering, and iii) generative summarization into distinct processing stages. Unlike previous approaches that treat hybrid summarization as a monolithic pipeline, the proposed architecture enables controlled experimentation on each component to better understand where factual drift emerges and how constraints influence model behavior. By systematically evaluating the impact of extractive sentence selection and biomedical entity filtering before decoder input, this study offers new insight into how architectural modularity affects not only summary accuracy but also the stability, traceability, and clinical interpretability of model outputs. The findings demonstrate that structuring the generation process around interpretable intermediate representations contributes to improved consistency, reduces hallucination frequency, and supports the production

of concise and clinically coherent summaries suitable for radiology reporting and broader medical NLP applications.

III. Methodology

The proposed framework adopts a modular design to ensure controllable and reliable generation of radiology report summaries. Rather than relying on a single end-to-end learning objective, the system decomposes the summarization task into three reasoning stages: i) content selection, ii) semantic filtering, and iii) controlled generation. This structure not only enables fine-grained analysis of system behavior but also aligns with the semantic workflow radiologists implicitly follow when transforming detailed findings into a concise diagnostic impression.

A. Problem Formulation

Let $F = \{s_1, s_2, \ldots, s_n\}$ denote the ordered set of sentences in the *Findings* section, where each s_i represents an individual clinical sentence extracted from the radiology report. The objective is to generate an *Impression* summary $\hat{y}$, where y denotes a candidate summary sequence and $\hat{y}$ is the optimal generated summary, that maximizes the conditional likelihood while satisfying clinical consistency constraints:

$$\hat{y} = \arg\max_{y} P(y \mid F, \mathcal{C}), \tag{1}$$

where $P(y \mid F, \mathcal{C})$ denotes the conditional probability of generating summary y given the source sentence set F and constraint set $\mathcal{C}$. The constraint term $\mathcal{C}$ encodes structural and semantic restrictions derived from extractive sentence scoring and domain-aware filtering, thereby limiting the admissible generation space.

Unlike unconstrained sequence-to-sequence generation, this formulation defines summarization as a staged and controlled mapping:

$$F \xrightarrow{\text{Extractive Scoring}} F' \xrightarrow{\text{NER Filtering}} F'' \xrightarrow{\text{Generation}} \hat{y}$$

where F' denotes the subset of sentences selected through extractive relevance scoring, and F'' represents the filtered representation obtained after applying domain-aware named entity recognition (NER) and semantic constraint filtering. The final generation step produces $\hat{y}$ conditioned on the progressively constrained representation.

This staged formulation enables explicit intervention between processing steps, allowing intermediate representations (F' and F'') to be inspected, validated, and, if necessary, adjusted, thereby supporting explainability and controllability within the overall summarization framework.

B. Dataset

Experiments were conducted using the Indiana University Chest X-ray dataset [19], a publicly available clinical corpus widely used in radiology NLP research. Each record in the dataset contains a radiology report paired with its corresponding chest X-ray image, although only textual components were used in this study. For summarization, the *Findings* and *Impression* sections were programmatically identified, extracted, and paired to form input–output supervision pairs. Reports lacking either section or containing malformed structure were excluded to ensure label quality, resulting in a final corpus of 3,955 usable examples.

Preprocessing included lowercasing, normalization of punctuation and spacing, and removal of boilerplate phrases such as institutional disclaimers. No attempt was made to correct grammatical inconsistencies or stylistic variation, as these characteristics reflect authentic clinical documentation patterns and contribute to task difficulty. The dataset was partitioned into 80% training, 10% validation, and 10% test splits, consistent with previous work on radiology summarization benchmarks [8] to ensure comparative relevance.

The original dataset is fully anonymized, with all Protected Health Information (PHI) removed prior to public release. All additional processing performed in this work adhered to HIPAA-compliant handling guidelines and best practices for clinical NLP experimentation. No external patient-identifying metadata was accessed (or introduced) at any stage of the study.

C. Model Architecture

The proposed summarization framework follows a modular pipeline designed to progressively refine the input text from raw clinical narrative to a concise and clinically coherent *Impression*. This structure allows each stage to contribute functionally distinct constraints on representation, filtering, and generation, supporting interpretability and reducing the likelihood of semantic drift. An overview of the full architecture is shown in Fig. 1.

1) Stage 1: Extractive Sentence Scoring: The first stage performs relevance estimation to identify the most clinically meaningful sentences in the *Findings* section. This is achieved using BERTSUM [20], an encoder-based architecture optimized for extractive summarization tasks. Each input sentence is encoded into a contextualized embedding representation, and a classification token is appended to enable sentence-level discrimination. Relevance is computed using a fully connected scoring layer applied over embeddings.

Rather than selecting a fixed number of sentences, the model employs a dynamic thresholding strategy. Let $F = \{s_1, \ldots, s_n\}$ denote the set of input sentences, where each s_i is encoded by BERTSUM into a contextualized embedding $h_i \in \mathbb{R}^d$. A trainable parameter vector $w \in \mathbb{R}^d$ is applied to compute a relevance score through a linear projection followed by a sigmoid activation function. The extractive subset is defined as

$$F' = \{s_i \in F \mid \sigma(w^\top h_i) > \tau\}, \tag{2}$$

where $w^\top h_i$ denotes the relevance logit for sentence s_i, $\sigma(\cdot)$ is the sigmoid activation mapping scores to the interval $(0, 1)$, and $\tau \in (0, 1)$ is a learned or validation-tuned selection threshold.

This adaptive mechanism allows the model to account for variability in report length and structure, preventing over-selection in verbose reports and under-selection in shorter or information-dense cases.

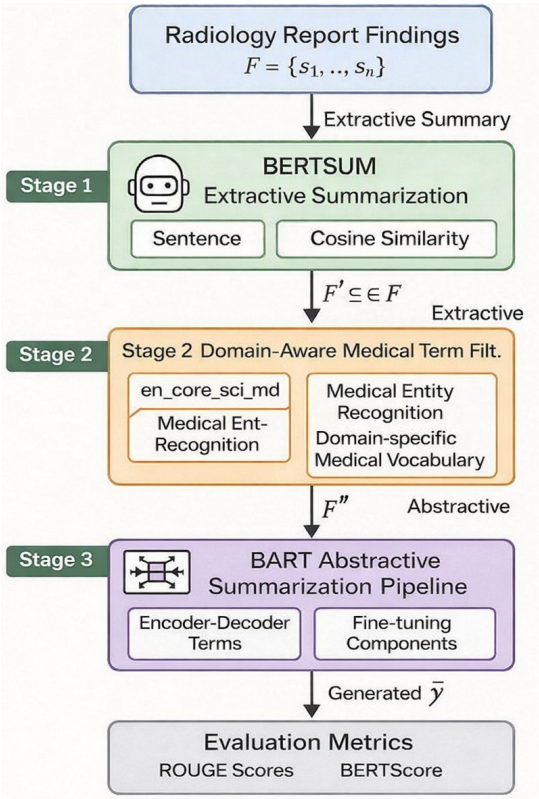

Fig. 1. Overview of the proposed modular pipeline. The *Findings* section $F = \{s_1, \ldots, s_n\}$ is processed through three stages: extractive scoring (Stage 1) yielding F', domain-aware filtering (Stage 2) producing F'', and abstractive generation (Stage 3) generating $\hat{y}$. Evaluation uses ROUGE and BERTScore.

2) Stage 2: Domain-Aware Filtering: The second stage applies clinical filtering to enforce semantic relevance and reduce noise. Extracted text F' is processed using SciSpaCy [21] with UMLS concept linking. Only lexical units corresponding to clinically relevant semantic groups, such as anatomical entities, pathologies, modifiers indicating severity or uncertainty, and radiological descriptors, are retained. Procedural statements, formatting artefacts, and non-actionable descriptive fragments are removed. Specifically, the retained UMLS semantic groups include *Anatomical Site*, *Disease or Syndrome*, *Sign or Symptom*, and *Diagnostic Procedure*. In addition, clinically meaningful modifiers (e.g., severity indicators and radiological descriptors) were preserved to maintain interpretive precision. Negation cues such as "no" and "without", as well as uncertainty expressions including "possible", "suggestive of", and similar hedging formulations, were explicitly retained during filtering to preserve diagnostic nuance and prevent distortion of clinical meaning.

This filtering stage serves two functions: i) it compresses the representation space by removing linguistically fluent but clinically irrelevant tokens, and ii) it reinforces factual alignment by ensuring the generative stage operates on medically meaningful content. As a result, the model receives inputs more consistent with radiology impression style and less influenced by verbose observational phrasing.

3) Stage 3: Abstractive Generation: In the final stage, the filtered content F'' is provided to a fine-tuned instance of BART [4], used as a conditional abstractive generator. BART was selected due to its strong performance in text compression and its bidirectional encoder–decoder structure, which is advantageous when rewriting structured clinical descriptions into shorter narrative summaries.

Decoding employs beam search with a beam width of 4, combined with a no-repeat 3-gram constraint to minimize repetition and reduce speculative reasoning. The model does not receive additional external knowledge bases or reinforcement signals; rather, factual constraints are carried forward implicitly from earlier pipeline stages. The output of this stage is the final *Impression*-style summary.

D. Training and Implementation

All components were trained or fine-tuned using the AdamW optimizer [22] with a learning rate of 2×10^{-5}. Training proceeded for up to 10 epochs with early stopping based on validation ROUGE-L performance to prevent overfitting. Experiments were conducted using PyTorch and the Hugging Face Transformers library on a workstation equipped with a single NVIDIA RTX 4090 GPU. The implementation was containerized to ensure reproducibility and to support potential downstream integration into deployment frameworks.

IV. Experiments and Results

This section presents the evaluation protocol and discusses the results of the proposed hybrid summarization framework. All experiments were conducted on the Indiana University Chest X-ray dataset [19], using the preprocessing, training, and inference settings described in Section III-B. The evaluation assesses whether structuring the summarization process into modular stages yields measurable improvements over a conventional extractive approach.

Evaluation Procedure

To systematically evaluate summarization quality, both lexical and semantic metrics were employed. ROUGE-1, ROUGE-2, and ROUGE-Lsum [23] were used as primary text-overlap measures to quantify n-gram and structural consistency between generated summaries and reference *Impression* sections. Because medical summaries frequently involve paraphrasing rather than token-level matching, BERTScore F1 [24] was additionally included as a contextual semantic similarity measure.

ROUGE scores were computed using the official ROUGE-1.5.5 implementation with Porter stemming enabled. Evaluation was performed exclusively against the reference *Impression* sections of the test split. Tokenization followed the default configuration of the HuggingFace `evaluate` framework to ensure consistency across all model outputs. No additional normalization or post-processing was applied before metric computation.

Evaluation was conducted on the test split withheld during model development to avoid overfitting and bias. No

TABLE I

PERFORMANCE COMPARISON BETWEEN THE EXTRACTIVE BASELINE AND THE PROPOSED HYBRID MODEL.

Model	ROUGE-1	ROUGE-2	ROUGE-Lsum	BERTScore F1
Extractive (BERTSUM)	13.96	4.31	12.59	0.8553
Hybrid (Proposed)	**58.19**	**49.53**	**57.68**	**0.9293**

TABLE II

EXAMPLE COMPARISON OF GENERATED SUMMARIES FOR A GIVEN RADIOLOGY REPORT.

Input (*Findings* excerpt): *"Heart size is normal. No pneumothorax. Mild increased interstitial markings bilaterally. No definite consolidation identified. Minimal blunting of the costophrenic angles may represent trace pleural effusion."*
Extractive Baseline Output: *"Heart size is normal. Mild increased interstitial markings bilaterally. Minimal blunting of the costophrenic angles may represent trace pleural effusion."*
Hybrid Model Output (Proposed): *"Mild interstitial prominence with possible trace bilateral pleural effusion. Heart size is normal. No pneumothorax or consolidation."*

post-processing (e.g., rule-based correction, smoothing, or heuristics) was applied to the generated outputs to ensure the evaluation reflects model behavior rather than handcrafted refinements.

Quantitative Results

Table I reports the comparative performance between the baseline extractive model (i.e., BERTSUM) and the proposed hybrid system. The extractive model demonstrated predictable behavior: although fully grounded in the source text, its summaries were long, minimally condensed, and lacked the stylistic characteristics of radiologist-written *Impression* statements. This resulted in low ROUGE scores and moderate BERTScore performance.

In contrast, the hybrid approach showed substantial performance improvement across all evaluation metrics. ROUGE-1, ROUGE-2, and ROUGE-Lsum increased by more than fourfold compared to the extractive baseline, indicating that the system successfully generated concise summaries while preserving the key semantic content. Similarly, BERTScore increased from 0.8553 to 0.9293, reflecting stronger alignment with the reference summaries in terms of meaning and contextual phrasing. The large performance gap reflects the fact that the extractive baseline does not perform abstraction or structural reorganization, whereas the hybrid model explicitly optimizes for alignment with Impression-style outputs.

Qualitative Observations

While quantitative evaluation provides evidence of performance improvements, qualitative inspection further highlights differences in linguistic structure, information prioritization, and clinical appropriateness between the extractive baseline and the proposed hybrid model. The extractive approach often reproduced long descriptive fragments from the *Findings* section with limited reorganization, producing summaries that were factually accurate but verbose and less aligned with radiology reporting style. In contrast, the hybrid method demonstrated stronger abstraction, generating concise summaries that preserved clinically significant information while maintaining radiological conventions such as uncertainty markers and negation cues.

Table II provides an illustrative example from the test set. As shown, the hybrid approach restructures content into a more diagnostic format, foregrounding abnormalities while retaining negation (e.g., "no pneumothorax") and uncertainty expressions (e.g., "possible trace"), features central to radiological reasoning [11]. The extractive method, although factually grounded, does not consistently reflect the conventional discourse structure of *Impression*-style outputs. These qualitative findings complement the quantitative results, indicating that the proposed modular architecture supports improved factual reliability, stylistic alignment, and clinical interpretability.

V. DISCUSSION

The findings indicate that organizing the summarization task into explicit processing stages yields measurable benefits in output quality and clinical reliability. In line with previous work observing hallucination and factual drift in unconstrained medical text generation [10], [11], the proposed hybrid approach demonstrated improved lexical alignment and semantic fidelity compared to a purely extractive baseline. The consistent gains across ROUGE and BERTScore suggest that the system not only retains diagnostically relevant information but also reformulates it in a manner consistent with radiologist-authored *Impression* sections [8].

A central contribution of this study is the role of modularization in controlling generative behavior. Previous research has shown that combining extractive anchoring with abstractive reasoning can improve factual grounding [12], [15]. In our framework, the extractive stage restricts generation to content explicitly present in the *Findings*, reducing exposure to irrelevant textual noise. The filtering of biomedical entities further reinforces clinically significant constraints, consistent with domain-adaptive approaches [5], [6]. Qualitative observations indicate that the hybrid method better reflects the interpretive structure typical of expert-authored *Impression* statements, prioritizing high-impact abnormalities over descriptive reporting [9]. While the extractive baseline preserved factual correctness, it lacked abstraction and stylistic consolidation. The controlled generative step enables compression and reorganization only after factual constraints are established, echoing recent work on controllable generation in safety-critical NLP applications [7], [13].

Beyond metric improvements, the modular architecture can be interpreted as a structured control mechanism. By progressively compressing and filtering the input representation, the framework reduces the decoding search space and mitigates semantic drift. From a modeling perspective, modu-

lar decomposition may be viewed as reducing representation entropy before generation, contributing to increased stability and coherence in the final summaries.

Despite these encouraging outcomes, the work presents limitations. The evaluation was performed on a single institutional dataset, and radiology language varies across modality, specialty, and region [8]. Broader multi-institutional validation would strengthen claims of robustness. Additionally, the entity filtering rules are heuristic rather than learned, which may limit portability to multilingual or domain-specific corpora. The current evaluation relies primarily on automatic lexical and semantic metrics and does not include clinically grounded factuality measures such as RadGraph-based evaluation, CheXbert-based labelling consistency, or structured human expert assessment. Incorporating such protocols would provide deeper insight into omission and fabrication patterns beyond token-level similarity. Finally, this study does not compare against very large proprietary language models. While such systems represent an important research direction, the present work prioritizes reproducibility, architectural transparency, and deployability within open-weight transformer frameworks. This choice reflects considerations of controllability and real-world clinical integration rather than direct benchmarking against closed commercial models.

VI. Conclusion and Future Work

This study introduces a modular hybrid summarization framework to improve factual reliability in medical text summarization by integrating i) extractive grounding, ii) clinical entity filtering, and iii) transformer-based abstractive generation. The results indicate that structuring the pipeline around interpretable intermediate steps yields consistent gains in lexical accuracy, semantic similarity, and hallucination reduction compared with extractive-only and end-to-end baselines. The findings further suggest that modularity enhances summary quality while providing operational benefits: i) intermediate representations can be inspected and validated, ii) computational demands decrease through input compression, and iii) the system becomes more predictable and controllable. These properties are particularly relevant in clinical settings, where unsupported statements can have significant consequences and explainability is essential for deployment. Overall, the results support hybrid architectures as a dependable pathway for integrating generative models into high-stakes workflows.

Despite these encouraging outcomes, further investigation is needed. The evaluation relies primarily on automatic metrics; incorporating structured radiologist feedback or clinically grounded factuality measures would strengthen real-world assessment. Extending the framework to multimodal settings and transferring it to related high-precision domains, such as pathology reports, operative summaries, cardiology assessments, legal narratives, or aviation safety reports, would further test generalizability.

In summary, controllable hybrid summarization frameworks represent a promising step toward trustworthy deployment of neural text generation in safety-critical environments, where accuracy, traceability, and interpretability are as important as linguistic fluency.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of ACL*, 2020, pp. 7871–7880.

[5] J. Lee, W. Yoon, S. Kim, D. Kim *et al.*, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[6] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," in *Proceedings of the 2nd Clinical NLP Workshop*, 2019, pp. 72–78.

[7] R. Luo, X. Sun, X. Tan *et al.*, "Biogpt: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[8] Y. Peng, S. Yan *et al.*, "A benchmark for radiology report summarization and evaluation," *Journal of Biomedical Informatics*, 2023.

[9] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of ICML*, 2020, pp. 11 328–11 339.

[10] S. Ji *et al.*, "Hallucinations in large language models: A survey," *arXiv preprint arXiv:2309.00113*, 2023.

[11] K. Krishna *et al.*, "Faithfulness in clinical natural language generation," in *EMNLP*, 2023.

[12] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," in *Proceedings of EMNLP*, 2018.

[13] N. Miao, H. Yu, L. Dai, and P. Blunsom, "Cgmh: Constrained text generation via metropolis-hastings sampling," in *ACL*, 2019, pp. 1530–1543.

[14] Y. Gu *et al.*, "Pubmedbert: Domain-specific language model pretraining for biomedical nlp," in *ACL*, 2021.

[15] J. DeYoung *et al.*, "Ms^2: Multi-document summarization for medical evidence," in *ACL*, 2021.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[17] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2009–2022.

[18] A. Nagar, Y. Liu, A. T. Liu, V. Schlegel, V. P. Dwivedi, A. K. Kaliya-Perumal *et al.*, "umedsum: A unified framework for advancing medical abstractive summarization," *arXiv preprint arXiv:2408.12095*, 2024.

[19] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[20] Y. Liu and M. Lapata, "Fine-grained attention for neural text summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5503–5514.

[21] M. Neumann, M. King, I. Beltagy, and W. Ammar, "Scispacy: Fast and robust models for biomedical natural language processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 319–327.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *ICLR*, 2018.

[23] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *ACL Workshop on Text Summarization*, 2004.

[24] T. Zhang, V. Kishore, F. Wu *et al.*, "Bertscore: Evaluating text generation with bert," *International Conference on Learning Representations*, 2020.