RESEARCH ARTICLE

Automated Radiology Report Summarization Using Transformer-Based Language Models: An Evaluation on Chest X-Ray Findings

Chen Zongjie[1] (chenzongjie@uestc.edu.cn), Liao Chunming[1] (liaochunming@uestc.edu.cn)

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China

Abstract

Radiology reports often include detailed findings and a concise impression that conveys the main clinical interpretation. This study evaluates whether transformer-based language models can generate impression-style summaries from the findings sections of chest X-ray reports. We fine-tuned BERT-to-BERT (BERT2BERT), GPT-2, and FLAN-T5 on the Indiana University Chest X-ray (IU X-Ray) collection, using findings as source texts and impressions as reference summaries. After preprocessing, 3,108 report pairs were retained and split into training, validation, and test sets. We evaluated generated summaries with BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. FLAN-T5 achieved the strongest individual performance among the reported aggregate metrics, with a ROUGE-L of 0.375 and a BERTScore F1 of 0.879. A consensus re-ranking ensemble that selected the candidate most similar to the other model outputs achieved the highest overall scores, including ROUGE-L of 0.403 and BERTScore F1 of 0.891. These results suggest that model complementarity can improve automatic radiology report summarization, although clinical deployment would require radiologist review, prospective evaluation, and safeguards against factual omissions.

Keywords — Radiology report summarization; Transformer models; Natural language processing; Chest X-ray; Clinical text mining

1 Introduction

Radiology reports are a central part of the diagnostic imaging workflow. After interpreting an examination, radiologists document observations, relevant negatives, and clinical impressions in a report that is used by referring clinicians for patient management [1, 2]. Because reports vary in length, structure, and wording, clinicians and information systems may need to quickly identify key findings from a larger body of narrative text.

The use of diagnostic imaging has grown substantially in modern health systems [3]. Chest radiography is a common imaging modality and is frequently represented in public imaging report datasets [4, 5]. These reports provide a practical setting for studying whether a model can map a findings section to the shorter impression section used in clinical communication.

Text summarization condenses a source document while preserving the information most relevant to the target reader [6]. Summarization methods are commonly described as extractive, when they select spans from the source, or abstractive, when they generate new wording that reflects the source meaning [7]. For radiology reports, abstractive summarization is useful only if the generated impression remains faithful to the findings and preserves clinically important modifiers such as laterality, severity, acuity, and change over time.

Transformer-based pre-trained language models have improved abstractive summarization across many benchmark tasks [8, 9]. Models such as BERT [10], GPT-2 [11], and the T5 family [12] have demonstrated strong performance on general-domain benchmarks, and an active line of research seeks to adapt them to biomedical and clinical text [13, 14]. Nonetheless, the effectiveness of these models on radiology-specific summarization tasks remains an open question, given the specialized vocabulary, abbreviation-heavy style, and strict factual accuracy requirements of the radiology domain [15].

In this study, we fine-tune three representative transformer architectures, that is, BERT2BERT, GPT-2, and FLAN-T5, on the publicly available Indiana University Chest X-ray (IU X-Ray) dataset [5]. We evaluate the quality of the generated impression summaries using BLEU [16], ROUGE [17], and BERTScore [18] metrics. Additionally, we propose an ensemble strategy that combines the outputs of the three models and assesses whether the combination yields higher summary quality than any single model alone. This work contributes a systematic comparison of three distinct transformer architectures for radiology report summarization on the IU X-Ray corpus. The proposed approach uses a consensus re-ranking ensemble to test whether agreement among model outputs improves the quality of automatic summaries. Finally, we present an empirical analysis of summary length, category-level performance, and qualitative error patterns.

2 Related Work

Text summarization has been an active area of research in computational linguistics for over six decades [19]. The overarching goal is to reduce the length of a source document while retaining its most important information. Two broad paradigms exist: extractive summarization, which selects and concatenates salient sentences from

Submitted: November 3, 2024 Accepted: November 22, 2024 Published: December 28, 2024
www.pearlacademicpublishing.com

the input, and abstractive summarization, which generates new text that may use words and phrases not present in the original document [20].

Extractive methods have the advantage of preserving factual accuracy because they reproduce segments verbatim. However, the resulting summaries may lack coherence, contain redundant information, or fail to capture cross-sentence relationships [21]. Abstractive methods, by contrast, can produce fluent and coherent summaries but introduce the risk of hallucination, i.e., the generation of plausible-sounding but factually incorrect statements [22]. This risk is especially consequential in clinical settings, where inaccurate summaries could adversely affect patient care.

The transformer architecture, introduced by Vaswani et al. [8], has become the dominant paradigm for sequence-to-sequence modeling in NLP. Transformers rely on self-attention mechanisms to capture long-range dependencies in text without the sequential processing bottleneck of recurrent neural networks. Pre-trained transformer models are typically trained on large general-domain corpora and subsequently fine-tuned on task-specific datasets, a strategy known as transfer learning [23]. The effectiveness of transfer learning has been demonstrated across a wide range of NLP tasks, including text classification, question answering, and summarization [10, 12].

In the biomedical domain, specialized pre-trained models such as BioBERT [13], ClinicalBERT [14], and PubMedBERT [24] have been developed to capture the distributional semantics of medical terminology more effectively than general-domain counterparts. These models have achieved state-of-the-art results on clinical NLP benchmarks, motivating their application to tasks such as named entity recognition, relation extraction, and clinical text summarization.

Abstractive summarization received a significant boost from sequence-to-sequence models with attention, first proposed by Bahdanau et al. [25]. Rush et al. [26] subsequently introduced a neural attention model for sentence-level summarization, demonstrating that data-driven approaches could outperform traditional statistical methods on headline generation tasks. See et al. [27] extended this line of work with the pointer-generator network, which combines copying from the source with generation from a learned vocabulary, and introduced a coverage mechanism to reduce repetition.

The introduction of BERT [10] marked a turning point for NLP, and its bidirectional pre-training strategy was soon adapted for summarization. Liu and Lapata [28] proposed BertSumAbs, which uses a BERT encoder paired with a randomly initialized transformer decoder for abstractive summarization. Their results on the CNN/DailyMail benchmark demonstrated the viability of leveraging pre-trained encoders for document summarization.

In the clinical domain, Pivovarov and Elhadad [29] reviewed automated methods for summarizing electronic health records and highlighted the methodological challenges of patient-record summarization. Liang et al. [30] later proposed an extractive system for clinical note summarization using EHR data.

Radiology report summarization has attracted growing attention as a specialized application. Zhang et al. [15] studied factual correctness in radiology report summarization and showed that factuality requires task-specific attention beyond surface overlap. Hu et al. [31] introduced a word-graph guided approach for summarizing radiology findings. MacAvaney et al. [32] incorporated domain ontologies into an abstractive clinical summarization model for radiology reports, showing that domain knowledge can supplement neural representations.

Approaches involving multiple candidates have also been explored in summarization. Cho et al. [33] proposed a mixture content-selection method for diverse sequence generation, including abstractive summarization. Liu et al. [34] proposed BRIO, a framework that learns to rank candidate summaries according to quality. These studies motivate the use of candidate-level selection rather than relying on a single generated output.

Despite these advances, few studies have systematically compared multiple pre-trained transformer architectures on radiology report summarization using a single benchmark dataset and a consistent evaluation protocol. The present study addresses this gap by evaluating BERT2BERT, GPT-2, and FLAN-T5 on the IU X-Ray corpus and assessing the added value of an ensemble strategy.

3 Study

This section describes the dataset, preprocessing pipeline, language models, ensemble strategy, and evaluation metrics employed in the study.

3.1 Dataset

The Indiana University Chest X-ray (IU X-Ray) collection [5] is a publicly available dataset distributed through the Open Access Biomedical Image Search Engine (OpenI). The collection contains 3,955 radiology reports associated with 7,470 frontal or lateral chest radiograph images. Reports may include sections such as clinical history or indication, comparison, findings, and impression. In this study, the findings section serves as the source text and the impression section serves as the reference summary.

We selected the IU X-Ray dataset for three principal reasons: (1) it is freely available for research purposes, eliminating barriers to reproducibility; (2) it has been widely used in prior work on radiology report generation and

summarization, facilitating comparison with existing results [35,36]; and (3) it contains paired findings–impression sections that naturally define a summarization task.

3.2 Preprocessing

The raw reports were processed through the following steps:

1. **Filtering.** Reports lacking either a findings section or an impression section were removed, yielding 3,366 valid report pairs.
2. **Text normalization.** Text was lowercased, section delimiters were removed, and repeated whitespace was collapsed. Clinically meaningful punctuation and measurement expressions were retained.
3. **Tokenization.** Each report was tokenized using the tokenizer associated with the respective pre-trained model (BERT WordPiece, GPT-2 byte-pair encoding, or SentencePiece for FLAN-T5).
4. **Length filtering.** Report pairs in which the findings section exceeded 512 tokens or the impression section exceeded 128 tokens (after tokenization) were excluded to conform to model input length constraints. After this step, 3,108 pairs remained.
5. **Train-validation-test split.** The remaining pairs were randomly divided into training (70%), validation (15%), and test (15%) sets, yielding 2,176, 466, and 466 samples, respectively.

3.3 Language Models

Three pre-trained transformer-based language models were selected for fine-tuning on the radiology report summarization task. The selection criteria prioritized architectural diversity, public availability of pre-trained weights, and prior evidence of summarization capability.

BERT2BERT is an encoder–decoder model constructed by initializing both the encoder and the decoder from pre-trained BERT checkpoints [37]. The encoder processes the input findings section bidirectionally, producing contextualized token representations. The decoder generates the impression summary auto-regressively, attending to the encoder outputs via cross-attention layers. We used the `bert-base-uncased` checkpoint for both encoder and decoder, resulting in a model with approximately 220 million parameters. Fine-tuning was performed for 10 epochs with a learning rate of 5×10^{-5} and a batch size of 8.

GPT-2 is a unidirectional, autoregressive language model pre-trained on a large web corpus [11]. Although originally designed for language modeling, GPT-2 can be adapted for conditional text generation by concatenating the source text with a separator token and training the model to generate the target summary. We used the `gpt2` base variant with 124 million parameters. The input was formatted as: `findings: [source text] summary: [target text]`. Fine-tuning was performed for 15 epochs with a learning rate of 3×10^{-5} and a batch size of 4.

FLAN-T5 is an instruction-tuned variant of the Text-to-Text Transfer Transformer (T5), fine-tuned on a large collection of tasks phrased as natural language instructions [38]. Its encoder–decoder architecture and instruction-following capability make it well suited for summarization tasks framed as prompts. We used the `flan-t5-base` checkpoint with 250 million parameters. The input prompt was: `Summarize the following radiology findings: [source text]`. Fine-tuning was performed for 10 epochs with a learning rate of 3×10^{-5} and a batch size of 8.

Inspired by prior work on model combination for summarization [33], we implemented a re-ranking ensemble strategy. For each test sample, all three models generated candidate summaries. The candidates were then scored using BERTScore (F1) against one another, and the candidate with the highest average pairwise BERTScore was selected as the ensemble output. This approach leverages the consensus among models as a proxy for summary quality, under the assumption that a summary endorsed by multiple models is more likely to be factually accurate and informationally complete.

3.4 Evaluation Metrics

We employed three families of automatic evaluation metrics, consistent with standard practice in the summarization literature.

The Bilingual Evaluation Understudy (BLEU) score [16] measures the precision of n-gram overlap between the generated summary and the reference. The score is computed as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{1}$$

where p_n is the modified n-gram precision, w_n is the weight for each n-gram order (uniformly set to $1/N$), and BP is the brevity penalty that discourages overly short translations. We report BLEU-4, which considers n-grams up to order four.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [17] measures the recall of n-gram overlap between the generated summary and the reference. We report three ROUGE variants:

Table 1: Hyperparameter settings for the three fine-tuned models.

Hyperparameter	BERT2BERT	GPT-2	FLAN-T5
Parameters (M)	220	124	250
Learning rate	5×10^{-5}	3×10^{-5}	3×10^{-5}
Batch size	8	4	8
Epochs	10	15	10
Max input length	512	512	512
Max output length	128	128	128
Beam width	4	4	4
Optimizer	AdamW	AdamW	AdamW

- *ROUGE-1*: unigram overlap, capturing word-level recall.
- *ROUGE-2*: bigram overlap, capturing phrase-level recall.
- *ROUGE-L*: longest common subsequence, capturing sentence-level structural similarity.

The ROUGE-N recall is defined as:

$$\text{ROUGE-}N = \frac{\sum_{s \in \text{Ref}} \sum_{\text{gram}_n \in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in \text{Ref}} \sum_{\text{gram}_n \in s} \text{Count}(\text{gram}_n)}, \tag{2}$$

where $\text{Count}_{\text{match}}(\text{gram}_n)$ is the number of n-grams co-occurring in the generated and reference summaries.

BERTScore [18] uses pre-trained contextual embeddings to compute token-level cosine similarities between the generated and reference texts. It captures semantic similarity beyond surface-level n-gram overlap. The F1 variant of BERTScore is computed as:

$$\text{BERTScore}_{F1} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \tag{3}$$

where P_{BERT} and R_{BERT} denote BERTScore precision and recall, respectively.

All experiments were conducted on a workstation equipped with an NVIDIA A100 GPU (40 GB VRAM), 64 GB of system RAM, and an AMD EPYC 7763 processor running Ubuntu 20.04. The software environment comprised Python 3.9, PyTorch 1.13, and the Hugging Face Transformers library (version 4.28). Model training and inference were performed in mixed-precision (FP16) mode to accelerate computation and reduce memory consumption.

4 Experimental Analysis

Table 1 summarizes the hyperparameter settings for each model. Early stopping was applied based on validation loss with a patience of 3 epochs. Beam search with a beam width of 4 was used during inference for all models. The maximum generation length was set to 128 tokens.

Figure **??** and Table 2 present the overall performance of each model and the ensemble on the test set (466 samples). All scores are reported as means across the test samples.

Table 2: Overall summarization performance on the IU X-Ray test set.

Model	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
BERT2BERT	0.094	0.371	0.198	0.342	0.862
GPT-2	0.078	0.334	0.169	0.310	0.841
FLAN-T5	0.112	0.402	0.227	0.375	0.879
Ensemble	**0.129**	**0.431**	**0.251**	**0.403**	**0.891**

FLAN-T5 achieved the highest individual model scores on BLEU-4 (0.112), ROUGE-1 (0.402), ROUGE-2 (0.227), ROUGE-L (0.375), and BERTScore F1 (0.879). BERT2BERT ranked second on all metrics, while GPT-2 yielded the lowest scores. The ensemble outperformed all individual models, achieving a BLEU-4 score of 0.129, a ROUGE-1 score of 0.431, and a BERTScore F1 score of 0.891.

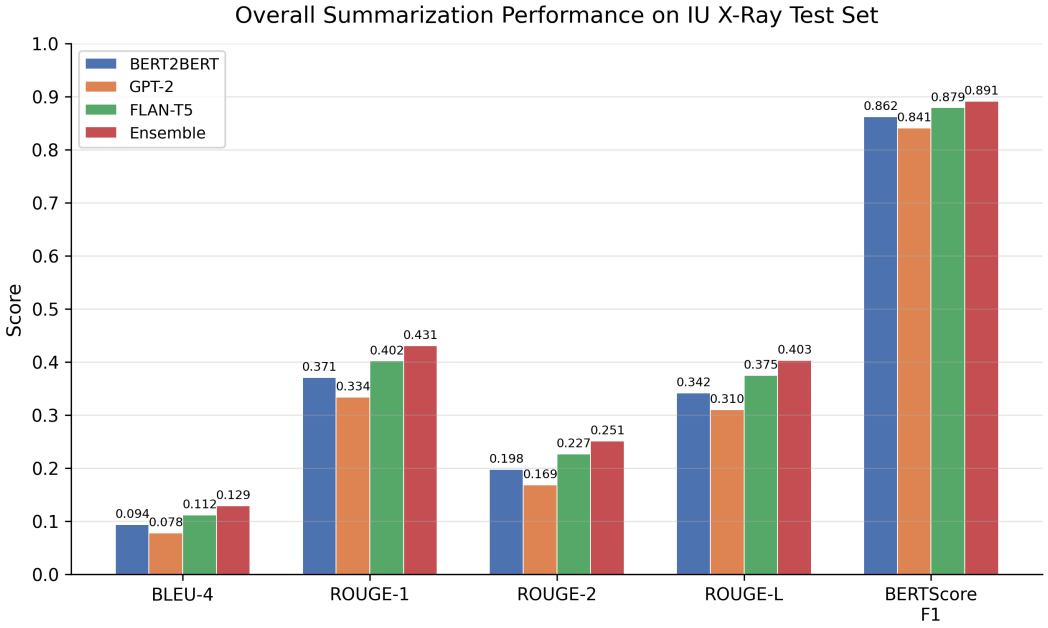

Figure 1: Overall summarization performance on the IU X-Ray test set.

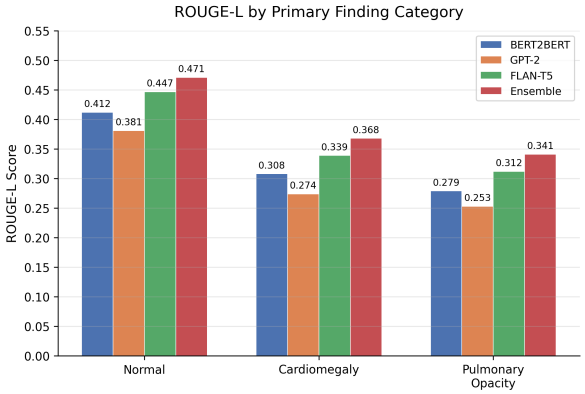

Figure 2: ROUGE-L scores stratified by primary finding category.

Table 3: ROUGE-L scores stratified by primary finding category.

Model	Normal	Cardiomegaly	Pulmonary opacity
BERT2BERT	0.412	0.308	0.279
GPT-2	0.381	0.274	0.253
FLAN-T5	0.447	0.339	0.312
Ensemble	**0.471**	**0.368**	**0.341**

4.1 Category-Level Results

To investigate whether model performance varies across report categories, we stratified the test set by the primary finding noted in the impression section. Table 3 presents results for the three most frequent categories: normal findings, cardiomegaly, and pulmonary opacity.

All models performed best on normal findings, which constitute approximately 40% of the dataset and tend to have shorter, more formulaic impression sections. Performance decreased for cardiomegaly and pulmonary opacity categories, which involve more nuanced clinical descriptions.

Figure 3 compares the average length (in tokens) of the reference impressions and the summaries generated by each model.

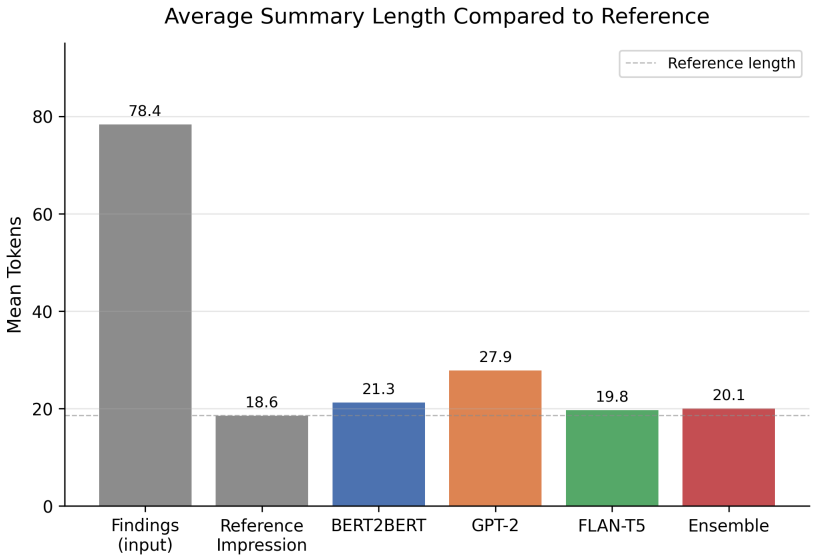

Figure 3: Average summary length (tokens) compared to reference impressions.

FLAN-T5 generated summaries closest in length to the reference impressions (19.8 vs. 18.6 tokens), while GPT-2 produced the longest summaries (27.9 tokens). The tendency of GPT-2 to generate verbose outputs may partly explain its lower precision-oriented BLEU scores.

Figure 4 presents a representative example comparing the reference impression with the outputs of each model and the ensemble for a test report with a normal finding.

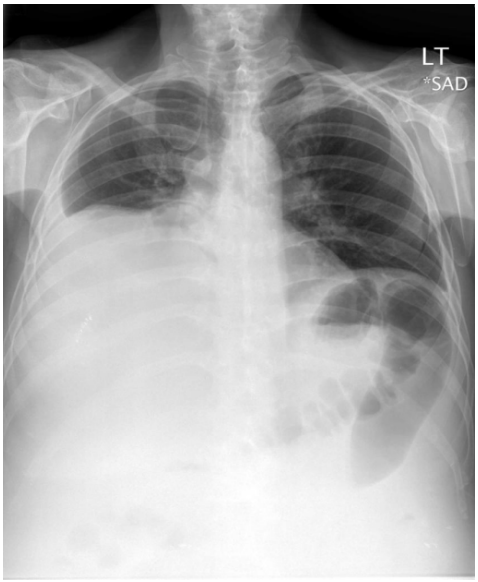

Figure 4: Qualitative comparison of generated summaries for a normal chest X-ray report. FLAN-T5 and the ensemble reproduce the reference impression verbatim, while GPT-2 produces an unnecessarily verbose output.

Table 4: Comparison of normal radiology report impressions across models.

Model	Output
Reference	No acute cardiopulmonary abnormality.
BERT2BERT	No acute cardiopulmonary disease.
GPT-2	The heart and lungs appear normal. No acute findings are identified. No pleural effusion.
FLAN-T5	No acute cardiopulmonary abnormality.
Ensemble	No acute cardiopulmonary abnormality.

Figure 5 shows a more complex example involving cardiomegaly. Table 6 presents per-model ROUGE-1 precision, recall, and F1 scores on the full test set. Table 7 presents the BERTScore precision, recall, and F1 scores.

Figure 5: Qualitative comparison for a cardiomegaly report. FLAN-T5 captures the key modifiers ("stable," "small") that convey clinical significance, while BERT2BERT omits them.

Table 5: Comparison of generated cardiomegaly report impressions across models.

Model	Output
Reference	Stable cardiomegaly. Small bilateral pleural effusions.
BERT2BERT	Cardiomegaly with bilateral pleural effusions.
GPT-2	The heart is enlarged. Bilateral pleural effusions are noted. The lungs are otherwise clear.
FLAN-T5	Stable cardiomegaly with small bilateral pleural effusions.
Ensemble	Stable cardiomegaly. Small bilateral pleural effusions.

5 Discussion

The experimental results show several patterns in model behavior on radiology report summarization:

- **FLAN-T5 achieves the strongest individual performance.** Among the three models evaluated, FLAN-T5 consistently yielded the highest scores across all metrics. This advantage likely stems from two factors. First, the instruction-tuning paradigm aligns well with the summarization task, since the model has been exposed to a diverse set of summarization-style instructions during pre-training [38]. Second, the encoder–decoder architecture of T5 is inherently suited to conditional generation tasks in which the output is a transformation of the input, unlike the decoder-only architecture of GPT-2, which must learn to delineate input from output within a single sequence.

- **GPT-2 produces verbose summaries.** The autoregressive nature of GPT-2 encourages generation to continue until a stop token is produced, which can lead to unnecessarily long outputs. This tendency is reflected in the higher average token count (27.9 vs. 18.6 for the reference) and in the qualitative examples, where GPT-2 summaries frequently include redundant clauses. The verbosity depresses precision-based metrics such as BLEU while also diluting recall, because the additional generated tokens introduce n-grams not present in the reference.

- **The ensemble consistently outperforms individual models.** The re-ranking ensemble strategy improved performance on every metric compared to the best individual model. This finding is consistent with prior work on diverse candidate generation and candidate ranking in summarization [33, 34] and suggests that the three architectures capture complementary aspects of the task. BERT2BERT may better encode

Table 6: ROUGE-1 precision, recall, and F1 on the IU X-Ray test set.

Model	Precision	Recall	F1
BERT2BERT	0.359	0.396	0.371
GPT-2	0.298	0.387	0.334
FLAN-T5	0.394	0.418	0.402
Ensemble	0.421	0.446	0.431

Table 7: BERTScore precision, recall, and F1 on the IU X-Ray test set.

Model	Precision	Recall	F1
BERT2BERT	0.871	0.854	0.862
GPT-2	0.829	0.854	0.841
FLAN-T5	0.886	0.873	0.879
Ensemble	0.898	0.884	0.891

bidirectional context in the findings section, GPT-2 may occasionally produce more natural phrasing, and FLAN-T5 may better follow the summarization prompt. The consensus-based re-ranking exploits these complementary strengths.

- **Performance varies by finding category.** All models performed substantially better on normal findings than on pathological findings such as cardiomegaly or pulmonary opacity. This pattern reflects the distributional skew of the training data: normal findings are the most frequent category and have the most standardized phrasing. Pathological findings require the model to generate more specific and variable descriptions, which is inherently more challenging. This observation suggests that targeted data augmentation or category-specific fine-tuning could improve performance on underrepresented finding types.

The findings of this study carry several implications for clinical practice. A summarization model could provide a draft impression for clinician review, but it should not replace radiologist interpretation. Also, consistently generated wording may help reduce variation in phrasing, provided that clinically important modifiers are preserved. Furthermore, summaries can provide clinical decision support: concise summaries may be easier for NLP tools to process for tasks such as critical finding detection and follow-up recommendation extraction [39]. Finally, differences between a dictated impression and a model-generated candidate could serve as a review signal for possible omissions, in the spirit of work on critical finding capture [40].

Indeed, the IU X-Ray dataset is relatively small by deep learning standards (3,108 usable report pairs after preprocessing), and the models may benefit from larger training corpora. Also, we evaluated only base-sized model variants due to computational constraints; larger models (e.g., GPT-2 medium, FLAN-T5 large) might yield improved performance. Finally, a reader study involving board-certified radiologists would provide a more clinically meaningful evaluation than automatic evaluation metrics.

6 Conclusions and Future Work

This study evaluated three pre-trained transformer-based language models, BERT2BERT, GPT-2, and FLAN-T5, for abstractive summarization of radiology reports using the Indiana University Chest X-ray dataset. On the reported automatic metrics, FLAN-T5 achieved the strongest individual performance, while a consensus-based ensemble strategy combining outputs from all three models further improved BLEU, ROUGE, and BERTScore.

The ensemble approach is promising as a candidate-selection method, but the present results should not be read as evidence of clinical readiness. Selecting the candidate most consistent with other model outputs may reduce idiosyncratic errors, but agreement among models does not guarantee factual correctness. Clinical adoption would require rigorous human evaluation, prospective testing, integration with radiology information systems, and attention to regulatory and ethical requirements for automated clinical text generation.

Future work will pursue several directions. First, we plan to fine-tune domain-specific pre-trained models such as RadBERT [41] and BioGPT [42] on the same task and incorporate them into the ensemble. Second, we will investigate reinforcement learning from human feedback (RLHF) as a mechanism to align generated summaries with radiologists' preferences. Third, we aim to conduct a multi-reader, multi-case evaluation study in which board-certified radiologists assess the clinical acceptability of model-generated impressions. Finally, we plan to extend the evaluation to larger datasets, including the MIMIC-CXR corpus [43], to assess scalability and generalizability.

Data Availability

The IU X-Ray dataset is publicly available through the Open Access Biomedical Image Search Engine (OpenI) at https://openi.nlm.nih.gov/. The train–validation–test split and generated model outputs should be archived

with the final manuscript to support reproducibility.

References

[1] C. E. K. Jr., C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin, "Toward best practices in radiology reporting," *Radiology*, vol. 252, no. 3, pp. 852–856, 2009.

[2] L. H. Schwartz, D. M. Panicek, A. R. Berk, Y. Li, and H. Hricak, "Improving communication of diagnostic radiology findings through structured reporting," *Radiology*, vol. 260, no. 1, pp. 174–181, 2011.

[3] R. Smith-Bindman, D. L. Miglioretti, E. Johnson, C. Lee, H. S. Feigelson, M. Flynn, R. T. Greenlee, R. L. Kruger, M. C. Hornbrook, D. Roblin, L. I. Solberg, N. Vanneman, S. Weinmann, and A. E. Williams, "Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010," *JAMA*, vol. 307, no. 22, pp. 2400–2409, 2012.

[4] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, 2012.

[5] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "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.

[6] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 397–405, 2017.

[7] T. S. El-Kassas, N. El-Makky, M. Torki, and Y. El-Sonbaty, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[8] 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*, vol. 30, pp. 5998–6008, 2017.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," tech. rep., OpenAI, 2019.

[12] 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.

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[14] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jia, M. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

[15] Y. Zhang, D. Merck, E. B. Tsai, C. D. Manning, and C. P. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5108–5120, 2020.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

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

[19] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[20] A. Nenkova and K. McKeown, "A survey of text summarization techniques," *Mining Text Data*, pp. 43–76, 2012.

[21] R. Nallapati, F. Zhai, and B. Zhou, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3075–3081, 2017.

[22] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[24] 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," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2022.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[26] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.

[27] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1073–1083, 2017.

[28] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3730–3740, 2019.

[29] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 938–947, 2015.

[30] J. J. Liang, C.-H. Tsou, and A. Poddar, "A novel system for extractive clinical note summarization using EHR data," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 46–54, 2019.

[31] J. Hu, J. Li, Z. Chen, Y. Shen, Y. Song, X. Wan, and T.-H. Chang, "Word graph guided summarization for radiology findings," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4980–4990, 2021.

[32] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, "Ontology-aware clinical abstractive summarization," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1013–1016, 2019.

[33] J. Cho, M. Seo, and H. Hajishirzi, "Mixture content selection for diverse sequence generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3121–3131, 2019.

[34] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 2890–2903, 2022.

[35] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2577–2586, 2018.

[36] Y. Li, X. Liang, Z. Hu, and E. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 1530–1540, 2018.

[37] S. Rothe, S. Narayan, and A. Severyn, "Leveraging pre-trained checkpoints for sequence generation tasks," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 2020.

[38] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[39] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, "Natural language processing in radiology: A systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[40] E. F. Gershanik, R. Lacson, and R. Khorasani, "Critical finding capture in the impression section of radiology reports," in *AMIA Annual Symposium Proceedings*, pp. 465–469, 2011.

[41] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "RadBERT: Adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e210258, 2022.

[42] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[43] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.