Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Biomedical Document Summarization using Deep Neural Networks - A Systematic Review

Umme Habiba
SaralPixel-AI Labs, New Delhi, 110025, India
Mantsha Saleem
SaralPixel-AI Labs, New Delhi, 110025, India
Ajay Rastogi
ajayrastogijmi@gmail.com

Amity Institute of Information Technology, Amity University, Noida, 201313, UP, India
https://orcid.org/0000-0002-6677-3650
Nripendra Kumar Singh
nripendrakr123@gmail.com

Department of Computer Applications, GCET, Greater Noida, 201310, UP, India https://orcid.org/0000-0002-4531-9985
Mohammad Faisal
Faculty of Dentistry, Jamia Millia Islamia, New Delhi, 110025, India https://orcid.org/0000-0003-3140-8204
Syed Shafat Ali
University of Kashmir, Srinagar, Jammu and Kashmir, 190006, India https://orcid.org/0000-0002-2974-8026

Systematic Review

Keywords: biomedical document summarization, deep learning, medical reports, large language models, transformers

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9956236/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Biomedical Document Summarization using Deep Neural Networks - A Systematic Review

Umme Habiba[1], Mantsha Saleem[1], Nripendra Kumar Singh[1,2*], Ajay Rastogi[3*], Mohammad Faisal[4], Syed Shafat Ali[5]

[1]SaralPixel-AI Labs, New Delhi, 110025, India.
[2]Department of Computer Applications, GCET, Greater Noida, 201310, UP, India.
[3]Amity Institute of Information Technology, Amity University, Noida, 201313, UP, India.
[4]Faculty of Dentistry, Jamia Millia Islamia, New Delhi, 110025, India.
[5]University of Kashmir, Srinagar, Jammu and Kashmir, 190006, India.

*Corresponding author(s). E-mail(s): nripendrakr123@gmail.com; ajayrastogijmi@gmail.com;
Contributing authors: ummehabibajmi@gmail.com; mantshasaleem@gmail.com; mfaisal@jmi.ac.in;
shafat159074@gmail.com;

Abstract

Biomedical document summarization is a field with significant potential to improve healthcare outcomes and patient care. The rise of digital health care data, including electronic health records, clinical notes, biomedical literature, and doctor-patient dialogue, has created a need to develop efficient techniques in summarizing medical reports. In this paper, we provide a comprehensive and systematic review of various research articles published between 2017 and 2025 for biomedical document summarization. For data collection, we searched digital libraries including IEEE, ACM, ACL, ScienceDirect etc., initially retrieving 167 research articles, of which 58 were selected after applying several filtering criteria. This review spans a broad spectrum across four key dimensions: medical reports and their types, approaches to medical report summarization, diverse datasets, and evaluation frameworks. We highlight current trends and the effectiveness of deep learning models (CNNs, RNNs, transformers, and LLMs) for medical report summarization. In addition, we provide a summary of diverse datasets—including benchmark, multimodal, and dialogue-centric datasets—and evaluation methods, highlighting their limitations and the need for more effective evaluation approaches. Overall, this study provides future directions for researchers in the biomedical summarization field by highlighting progress, current trends, and existing research gaps.

Keywords: biomedical document summarization, deep learning, medical reports, large language models, transformers

1 Introduction

1.1 Background

The field of medical report summarization has gained ever-increasing significance due to the exponential growth of digital medical data, including electronic health records, clinical notes, and published literature. Large numbers and highly complex nature of documents create substantial challenges for health care professionals who need to efficiently extract relevant information for patient care and for supporting clinical decisions [1, 2]. Traditionally, clinicians devoted large quantities of time to the onerous task of manual perusal and consolidation of patient data, an activity potentially hindering direct patient interaction as well as increasing the risk of information overload [2, 3]. Automated summarization methods, including extractive, abstractive, and hybrid approaches, convert voluminous unstructured medical narratives into concise, coherent, and clinically relevant summaries [1, 4–6]. Due to the evolution of deep learning models, summarizing medical reports has recently gained significant attention from researchers and practitioners [1, 6–8]. As a result, medical report summarization has become an important tool for enhancing the efficiency, accuracy, and accessibility of healthcare data in the present-day clinical landscape [1, 2, 9].

1.2 Motivation and Challenges

The motivation behind summarizing medical reports lies in the acute need for lessening the requirements of documenting and relieving overload of information in the field of healthcare. Healthcare professionals and researchers are often confronted with the difficulty of comprehending voluminous and highly complex medical documents within strict time constraints, potentially hampering productivity, increasing cognitive load, and affecting the

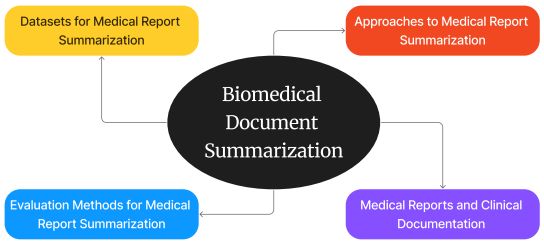


Fig. 1: Key dimensions covered in this review on biomedical document summarization

quality of care for patients [3, 4, 8, 10]. Automatic summarization emerges as an effective solution by providing faster access to key content, aiding clinical decision-making, and improving overall patient care as biomedical literature continues to expand.[1, 4, 8, 10].

Despite its promise, this area faces many lingering challenges. Ensuring accuracy and reliability of facts at the highest level is of the greatest priority as summarizing systems tend to unwittingly create inaccuracies, miss important details, or synthesize content potentially leading to clinical consequences [8, 11, 12]. Due to the heterogeneity of medical documents containing both structured radiology reports as well as unstructured patient queries, developing models extracts the relevant information succinctly but necessarily bearing brevity as well as clinical relevance [1, 3, 8]. In addition, the difficulty involved in summarizing long unstructured narratives, incorporating domain-specific medical knowledge, as well as data privacy and ethical guideline compliance, add further technical as well as practical challenges [3, 4, 10, 13]. Popular evaluation measures sometimes fail to assess clinical usefulness, highlighting the need for better domain-specific metrics [2, 4]. Overcoming these challenges plays a vital role in the construction of reliable, efficient, and clinically relevant summarization tools safely applicable in practical health-care applications.

1.3 Highlights of our Review

In this paper, we have provided an in-depth and systematic review of biomedical document summarization with a detailed analysis of research articles published between 2017 and 2025. For data collection, we searched digital libraries, including IEEE, ACM, ACL, and ScienceDirect, initially retrieving 167 research articles. Of these, 58 were selected after applying several filtering criteria based on factors such as publication outlet, indexing status (e.g., Scopus, Web of Science), and publisher reputation (e.g., IEEE, ACM, Springer, Elsevier). Subsequently, we categorize the surveyed literature along four key dimensions: medical reports and their types, approaches to medical report summarization, diverse datasets—including benchmark, multimodal, and dialogue-centric datasets—and evaluation frameworks, as shown in Figure 1. Furthermore, this survey highlights current trends and the effectiveness of deep learning models, ranging from early CNN- and LSTM-based approaches to recent transformer- and large language model (LLM)–based frameworks, for summarizing different types of medical reports, such as discharge summaries, radiology reports, clinical notes, and patient–clinician dialogues. In this study, we pose the following research questions:

Q1: What are the different types of medical reports, and how are they used in the existing biomedical literature?

Q2: What diverse datasets have been curated for report summarization?

Q3: What are the various deep learning approaches used for medical report summarization, and how have they progressed?

Q4: Which evaluation measures are critical for assessing the performance of report summarization?

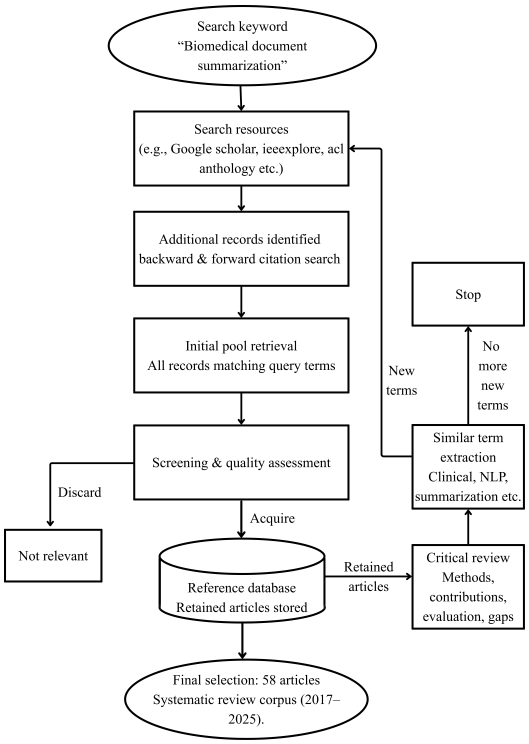
To address the above research questions, we make the following key contributions:

- This study provides a systematic review of 58 quality research articles published between 2017 and 2025 in biomedical document summarization.
- We provide an overview of medical reports and their types to automate clinical documentation for healthcare professionals (Section 3).
- We categorize diverse datasets along with current trends in deep learning approaches used for biomedical report summarization (Sections 3 and 4).
- This review also highlights different evaluation strategies employed in the biomedical literature summarization (Section 5).
- This review outlines future research directions in biomedical summarization by highlighting its progress, current trends and existing gaps (Section 6).

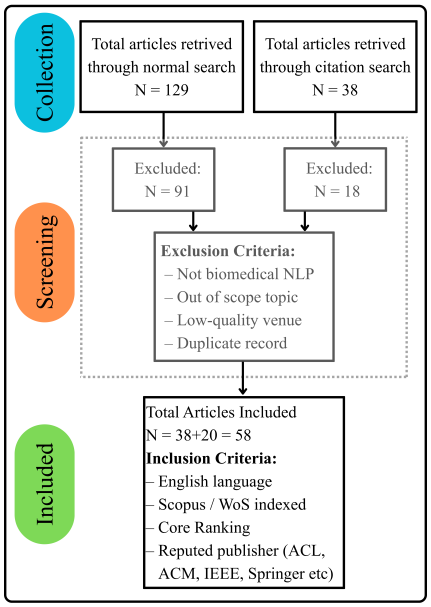
2 Literature Review Methodology

In this study, we adopt a systematic literature review methodology to analyze research articles published between 2017 and 2025 in biomedical document summarization, as shown in Figure 2. We start by searching for the query term "Biomedical Document Summarization" in major scholarly search engines and online libraries, including Google Scholar, IEEE Xplore, ACL Anthology, and other scientific archives. The articles identified during this initial search were used as base papers for the subsequent detailed literature review. After a careful scrutiny of these base papers, we identified several synonymous and closely related terms, including clinical text summarization, medical text summarization, healthcare report summarization, biomedical NLP summarization etc. Later, we searched for different permutations of new terms to build our literature repository. In addition, we also performed backward and forward citation searches on selected studies to identify additional relevant articles not captured by the keyword-based search (Figure 2a). In this way, we retrieve a total of 167 research articles, including 129 articles from keyword-based search, and 38 from citation-search, satisfying our query.

To make sure that the credibility and quality of the literature so chosen, we used several inclusion criteria depending on parameters like English language, publication outlet, status of being indexed (e.g., Scopus, Web of Science), and reputation of the publisher (e.g.,



(a) Literature Collection



(b) Literature Selection

Fig. 2: Graphical representation of literature review methodology

IEEE, ACM, ACL, Springer, Elsevier etc.). In addition, we remove duplicate records and other irrelevant articles that did not contribute to biomedical summarization. In this way, articles satisfying the quality and relevance criteria were kept and stored within a structured reference database. Following this, each retained article was critically reviewed to extract methodological details, key contributions, evaluation strategies, and identified limitations. Throughout the review process, other domain-specific terms occasionally emerged, prompting us to refine our search terms and repeat the search cycle. This iterative cycle consisting of keyword search, literature retrieval, quality filtering, and critical evaluation was carried out until no new terms or related articles were forthcoming, indicating saturation of the search process. By adopting this iterative and systematic approach, we gathered a rich and high-quality corpus consisting of a final set of 58 research articles on biomedical document summarization (Figure 2b).

3 Medical Reports and Clinical Documentation

This section aims to answer the research questions *Q1* and *Q2* raised in Section 1 to provide an overview of medical reports and related datasets in the biomedical domain. More specifically, we shed light on various types of medical report, including discharge summaries, radiology reports, progress notes, consultation notes, clinical QA sessions, and biomedical literature along with the datasets utilized in the biomedical document summarization domain. In addition, we also highlight standard documentation formats such as SOAP notes, pathology reports, etc. which are crucial to develop effective computational models for clinical documentation.

3.1 Types of Medical Reports

In the field of biomedical, clinical documentation plays a vital role for authentic and accurate patients' diagnoses and effective treatment planning including symptoms, test results, treatments, medications, and other relevant clinical data. These reports are generally created, maintained and accessed by healthcare professionals such as doctors, nurses, pharmacists and others. We categorize the medical reports used in the reviewed literature into six categories: *discharge summaries*, *radiology reports*, *progress note*, *consulation notes*, *clinical QA sessions*, and *biomedical literature*, as shown in Figure 3. Table 1 summarizes all such report types along with their dataset details. This table illustrates the variability of a total of 30 datasets in terms of size, modality (e.g., text, text + image, audio + text), and availability (e.g., open access, private, limited access), thus helping researchers select suitable resources for specific summarization tasks. Moreover, for reference, Figure 4 shows the frequency distribution of research articles across different types of reports in the reviewed literature. It is clear that *consultation notes* take the maximum count, whereas discharge summaries come at lowest. As *consultation notes* contain long clinical documentation that includes patient history, symptoms, diagnoses, and treatment plans, making them a comprehensive and practically important target for biomedical summarization research using LLMs. Moreover, this predominance of *consultation notes* can be attributed to the availability of large-scale public clinical datasets (Table 1). However, *discharge summaries* are already condensed clinical documents containing critical information, and further summarization may increase the risk of information loss or factual inaccuracies. This

can also be verified by the limited availability of public datasets for discharge summaries (Table 1).

3.1.1 Discharge Summaries

Discharge summaries are important clinical documents that summarize the entire process of a patient's hospital stay, from admission to diagnosis, interventions, medications ordered, and post-discharge care [14]. Ineffective or incomplete summaries can compromise patient safety and hinder communication between primary and secondary care. Since they play a central role, discharge summaries have been an important target in clinical summarization studies. From our literature, 2 well-known public datasets MIMIC-III [15] and MIMIC-IV [16] falls in this category (Table 1). These datasets are in-house biomedical datasets containing institution-specific or proprietary information that may include de-identified clinical reports, dialogues, or data obtained from individual healthcare providers or local studies. These datasets tend to be bounded by privacy and confidentiality guidelines, but are very helpful for institutional-specific NLP applications such as clinical summarization, report structuring, and dialogue-to-note generation [8, 17]. Several studies including Michalopoulos *et al.* [18], Sharma *et al.* [19] , and Kraljevic *et al.* [20] have utilized these datasets for large language modeling and demonstrated strong performance in automatic discharge summary generation.

3.1.2 Radiology Reports

Radiology reports describe imaging procedures such as *X-rays*, *CT scans*, and *MRIs* that detail clinical findings, diagnoses, and recommendations. Automated summarization of such reports assists radiologists and physicians in expediting diagnostic workflows and improving clinical communication. Some of the well-known datasets in this category are transitioned from exclusively textual data to multi-modal that consist of visual and textual content. Key examples include MIMIC-CXR [21] which associates more than 370,000 chest radiographs with structured radiology reports, and the Indiana University Chest X-ray (Open-i) dataset [23] comprising 7,470 X-ray images with 3,955 corresponding reports (Table 1). Another popular dataset, MEDIQA 2021 RAD [25], is a benchmark shared-task dataset designed for the evaluation and comparison of models in medical NLP research. Recent works [22, 24, 26, 29] state that multimodal integration of text and visual data considerably improves the quality and clinical applicability of radiology report summarization.

3.1.3 Progress Notes

Progress notes document the dynamic state of a patient's condition over a course of clinical observations, treatments, and care decisions over several visits. They require consistent language and high contextual fidelity, making them natural candidates for both self-supervised and fine-tuned summarization schemes. A well-known public repository, MTSamples, offers medical transcripts with accompanying SOAP-style notes, falls into this category. It includes doctor–patient dialogues, tele-consultation conversations, and synthetic or anonymized dialogues for clinical NLP tasks. Two additional benchmark datasets, ProbSum [32], which focuses on summarizing structured problem lists from progress notes, and ShARe/-CLEF [31], which supports clinical concept and assertion detection, are widely used progress note–oriented corpora (Table 1). Various frameworks have since been introduced to enhance progress note summarization, including the RAG framework [30], built on GPT-3.5-Turbo, and the MedCAT framework [20], which combines NER+L with BioClinicalBERT to improve summarization performance. Moreover, some recent studies [8, 20] fine-tuned LLMs to produce context-aware summaries from progress notes.

3.1.4 Consultation Notes

Consultation notes capture patient–clinician interactions and are usually recorded in a structured SOAP (Subjective, Objective, Assessment, Plan) format. Summarizing these notes enables effective clinical documentation and allows for the development of dialogue-to-note generation systems. In recent years, several studies have provided datasets for consultation note summarization tasks. In this direction, most datasets are dialogue-centric, such as PriMock57 [33] , NoteChat-1000 [35] , MTS-Dialog [37] , MEDIQA-Chat 2023 [38] etc., while some are in-house datasets, such as Stanford Hospital Course Summ. [32] and the Family Medicine Dataset [53]. Proprietary datasets such as Abridge SOAP [43] and Verilogue Inc. transcripts [52] are true consultation data, but not publicly available because of confidentiality limitations. Consultation notes are highly important for clinical document summarization, as most reviewed research articles use them for summarization tasks (Figure 4). To understand consultation note summarization, several studies [17, 34, 39, 40, 42] investigate a broad range of architectures and optimization techniques to significantly enhance its quality and reliability. For more details on consultation reports and the datasets, kindly refer to Table 1.

3.1.5 Clinical QA Sessions

Clinical Question–Answer (QA) sessions are interactive conversations between patients and clinicians that are intended to explain diagnoses, treatments, or medical problems. Summarizing such sessions using models that maintain factual correctness, contextual consistency, and readability while handling multimodal inputs such as textual and visual data is challenging. In this direction, benchmark datasets such as MeQSum [57] have provided the means to systematically analyze clinical QA summarization frameworks, while the newer multi-modal corpora MMQS [54] and MMCQS [56] have generalized such investigations to image-text domains (Table 1). These datasets are also code-mixed medical datasets made up of bilingual/multilingual statements (e.g., Hindi–English) interlaced with relevant clinical images or textual summaries. Current studies have focused on fusing LLMs with multimodal encoders [54, 55] to improve the clinical QA summarization process. Moreover, a recent study



Fig. 3: Comprehensive overview of biomedical document summarization: (a) taxonomy of approaches for medical report summarization, including CNN-based, RNN-based, transformer-based, generative, and LLM-based methods; (b) datasets used for medical report summarization, covering in-house, benchmark, multimodal, and dialogue-centric datasets; (c) evaluation methods, including domain-specific metrics, general-domain automated metrics, and human evaluation strategies; and (d) medical reports and clinical documentation, illustrating types of reports and standardized documentation formats considered in this survey.

Table 1: Summary of Medical Reports Utilized in Biomedical Field

Report Type	Dataset	Dataset Type	Modality	Size	Description	Availability	Referred In	Link
Discharge summary	MIMIC-III [15]	In-house dataset	Text	2M+ notes, 46K patients	ICU clinical database including discharge summaries, nursing notes, and radiology reports	Credential Access	[8, 17, 20]	MIMIC-III
	MIMIC-IV [16]	In-house dataset	Text	331K summaries, 2.3M reports	Clinical text corpus containing discharge summaries and radiology reports	Credential Access	[19]	MIMIC-IV
Radiology reports	MIMIC-CXR (v2.0) [21]	Multi-modal	Text + Image	377k images, 227k reports	Chest X-rays, paired with images	Credential Access	[8, 22]	MIMIC-CXR
	Indiana Chest X-ray (Open-i) [23]	Multi-modal	Text + Image	7,470 images, 3,955 reports	Chest Radiography reports (Indiana University Chest X-Ray Collection)	Open Access	[22, 24]	Open-i
	MEDIQA 2021 RAD [25]	Benchmark & shared-task	Text	3,600 reports	Radiology report summarization (Findings → Impression)	Open Access	[26]	MEDIQA2021
	Mayo Clinic Chest CT Reports	In-house dataset	Text	7k reports	De-identified CT reports from Mayo Clinic	Limited Access	[27]	Mayo
	CheXpert (Stanford PACS) [28]	In-house dataset	Text + Image	224K images, 65K patients	Large chest radiograph dataset with reports	Open (DUA)	[29]	CheXpert
Progress notes	MTSamples	Dialogue-centric	Text	5k documents	Multi-specialty clinical notes (web-based collection)	Open Access	[30]	MTSamples
	ShARe/CLEF 2014 [31]	Benchmark & shared-task	Text	300 notes	Annotated clinical notes for disorder extraction	Credential Access	[20]	ShAReCLEF
	ProbSum [32]	Benchmark & shared-task	Text	1K notes, 2.7k diagnoses	Task for summarizing diagnoses and problem lists from clinical progress notes	Credential Access	[8]	ProbSum
Consultation notes	PriMock57 [33]	Dialogue-centric	Dialogue + Text	57 consultations	Primary care consultation transcripts	Open access	[34]	PriMock57
	NoteChat-1000 [35]	Dialogue-centric	Dialogue + Text	1,000 synthetics	Conversational transcripts for clinical note generation	Open Access	[34]	NoteChat
	ACIBench [36]	Dialogue-centric	Dialogue + Text	1,000 encounters	Dataset for automatic clinical impression generation from dialogues	Open Access	[17]	ACIBench
	MTS-Dialog [37]	Dialogue-centric	Dialogue + text	1,726 dialogues	Conversations with SOAP summaries	Open Access	[37–40]	MTS-Dialog
	MEDIQA-Chat 2023 [38]	Dialogue-centric	Dialogue + Text	1.7K MTS Dialog + 207 ACI-Bench	Shared task dataset for dialogue summarization and note generation	Open Access	[41]	MEDIQA-Chat-2023
	MM-CliConSumm [2]	Dialogue-centric	Dialogue + Text + Image	1,200	Multimodal consultation summarization	Open Access	[42]	MM-CliConSumm
	Abridge SOAP [43]	Proprietary / In-house	Dialogue + Clinical Notes	146k utterance–SOAP pairs	Doctor–patient conversations paired with structured SOAP notes and evidence alignment	Private	[43, 44]	-
	Stanford Hospital Course Summ. [32]	In-house dataset	Text	20K	Hospital course summarization dataset	Limited Access	[45]	STARR
	Ambulatory Orthopedic Surgery Encounters	Dialogue-centric	Text	Few hundred	Transcripts from outpatient orthopaedic consultations with notes	Private	[46]	-
	FigShare + MTS-Dialog (combined)	Dialogue-centric	Text	2K	Combined datasets used in research (not official standalone dataset)	Open Access	[40]	FigShare MTS-Dialog
	Anonymized clinical dialogues [47]	Dialogue-centric	Text	27 dialogues	Primary care conversations for summarization	private/ controlled access	[48–50]	-
	MCDH dataset [51]	Multimodal and Code-mixed	Text + Image	1,668 dialogues	Multimodal Code-Mixed Dialogues in Healthcare	Open Access	[51]	MCDH

	Verilogue Inc. + YouTube [52]	Dialogue-centric	Audio + Text	1k dialogues + 117 videos	Real consultation recordings used in research	Verilogue Inc. (Private), YouTube (Open Access)	[52]	-
	Family Medicine Dataset [53]	In-house dataset	Text	150 audio transcribed consultations	Speaker-diarized primary-care transcripts	Private	[53]	-
Clinical QA sessions	MMQS [54]	Multi-modal & Code-mixed	Text + Image	3,000	Medical queries from online forums augmented with relevant images	Open Access	[51, 55]	MMQSD
	MMCQS [56]	Multi-modal & Code-mixed	Text + Image	2,500	Code-mixed medical Q&A (Hindi–English) + images	Open Access	[51, 55]	MMQSD
	MeQSum [57]	Benchmark & shared-task	Text	1k	Consumer health questions with expert summaries	Open Access	[8, 17]	MeQSum
Biomedical literature	MedMentions [58]	Biomedical corpora	Text	4,392 abstracts	PubMed abstracts fully annotated with UMLS concepts	Open Access	[20]	MedMentions
	PubMed	Biomedical corpora	Text	30M+	Biomedical abstracts and articles	Open Access	[41, 59, 60]	PubMed
	arXiv	Biomedical corpora	Text	1.7M	Preprint Scientific abstracts and articles	Open Access	[17, 45, 61–66]	arXiv

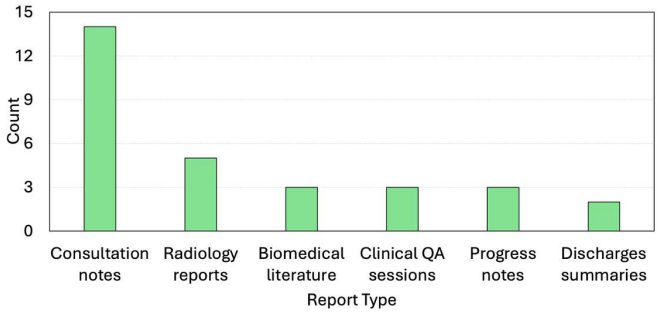


Fig. 4: Frequency distribution of research articles across different medical report types

by Van Veen *et al.* [8] explored parameter-efficient fine-tuning on these datasets and reported improved performance. In-depth details of deep learning models trained on these datasets can be found in Section 4.

3.1.6 Biomedical Literature

The corpora of biomedical literature include publicly available research articles such as pre-prints, journal articles, and conference papers intended mainly for scientific knowledge extraction, information retrieval, and text mining, but not for the direct analysis of clinical data. Their emphasis lies in extracting structured biomedical knowledge, such as terminologies, relationships, and contextual evidence embedded within scientific articles. Some examples of biomedical literature are large-scale corpora such as PubMed [5, 13] that store millions of abstracts as well as full-text research articles, MedMentions [58] consisting of 4,392 PubMed abstracts richly annotated with UMLS concepts and arXiv [17, 45, 61–66] that archives preprints on a wide range of topics from bioinformatics through computational biology to biomedical engineering. Several studies such as Kraljevic *et al.* [20], Dou *et al.* [60], Dalal *et al.* [59], Beltagy *et al.* [61], and Lee *et al.* [67] have utilized various biomedical and general-domain corpora, such as Reddit, XSum, CNN/DailyMail, WikiHow, NYT, and PubMed, for biomedical summarization tasks.

Figure 5 presents a taxonomy of medical datasets, highlighting each dataset type and the corresponding deep learning models employed in the literature.

3.2 Standardized Documentation Formats

In the previous section, we emphasize the discussion on variety found in medical reports, but in the field of medical science, foundational documentation formats are crucial for creating effective computational models. While the structure of medical documentation varies, many medical institutions and specialties depend on standardized formats to ensure consistency, enhance clarity, and support subsequent applications such as research, billing and automated summarization. Consequently, summarization models must be engineered to adapt not only to the unique linguistic characteristics of medical reports but also to their specific documentation style. In our reviewed literature, we categorize standardized formats into 3 broader types: *SOAP notes*, *pathology reports* and *medical evidence*.

3.2.1 SOAP Notes

The Subjective, Objective, Assessment, and Plan (SOAP) note is a widely used method of documentation among healthcare professionals. The SOAP structure reminds clinicians of essential tasks, provides a framework for analyzing information, and supports clinical reasoning [68]. It enables healthcare professionals to evaluate, diagnose, and treat patients based on reported information, while also serving as a checklist and cognitive aid. The basic structure of a SOAP note is shown in Figure 6. A SOAP note is divided into four sections: Subjective (S). Objective (O), Assessment (A) and Plan (P).

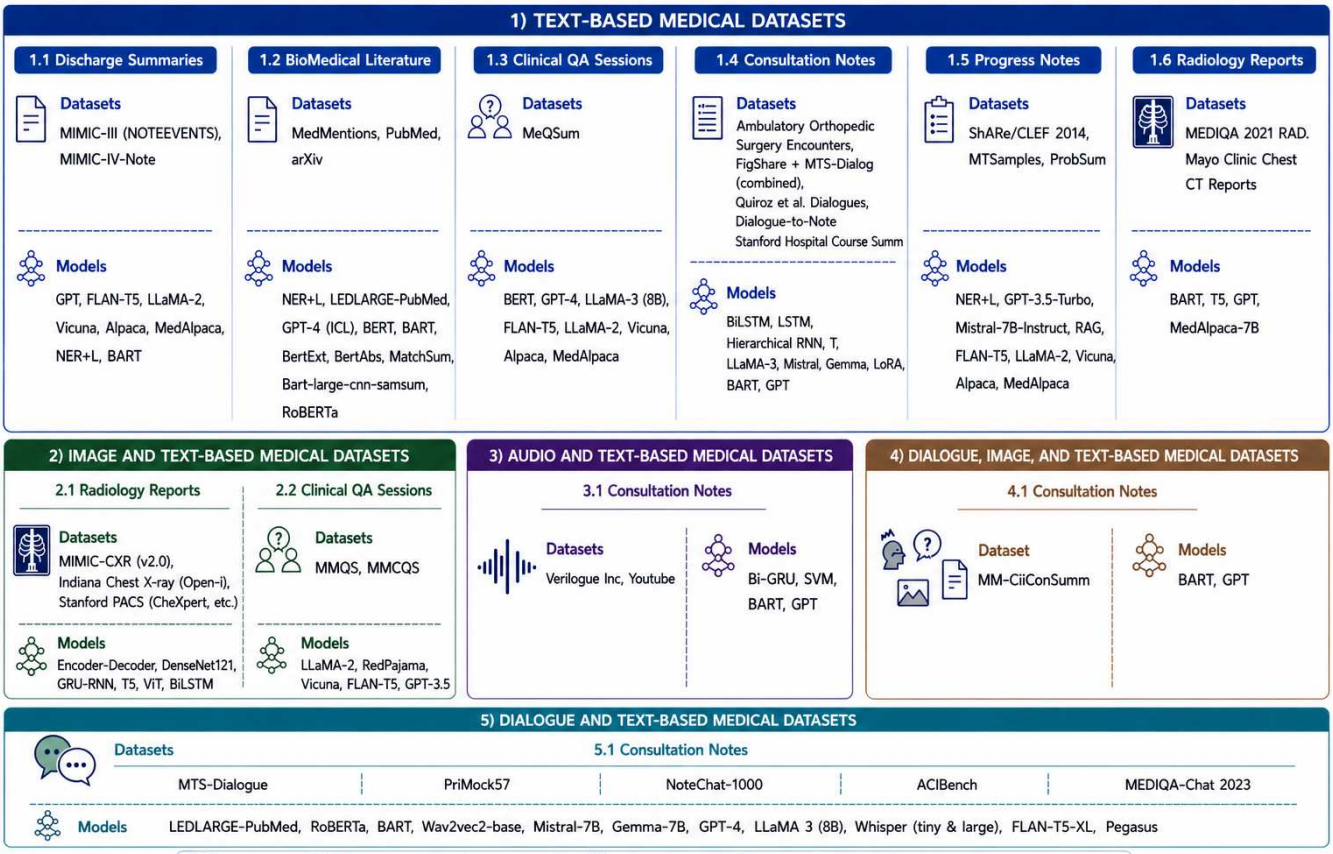


Fig. 5: Taxonomy of medical summarization datasets and models. Here, the models are associated with the datasets on which they are trained across four modalities: text, image, audio, and dialogue

Subjective (S): This section records details of the patient's health concerns and medical history.

Objective (O): This section captures vital signs, diagnosis, lab reports, and medications for patient care.

Assessment (A): This section provides a concise overview of both past and active diagnoses, explaining the reason for hospital admission, and highlighting current issues along with comorbidities.

Plan (P): This section outlines the need for further testing and consultations with other healthcare providers to address the patient's condition. It also describes additional measures being implemented for patient care, helping subsequent physicians understand the next steps.

This structured approach mirrors the natural workflow of a consultation and has therefore become the gold standard for transforming raw doctor–patient dialogues into structured clinical notes.

3.2.2 Pathology Reports

Pathology reports contribute largely to the medical corpus and are essential for guiding clinical decisions, but are often complex and lengthy, containing a mix of crucial and superfluous data. Large Language Models (LLMs) tackle this complexity by extracting key information and producing accurate and concise summaries to support rapid decision-making. One primary application of LLMs is to generate readable summaries of long and medical domain reports for both patients and other medical professionals [50]. *Synoptic reporting* is another key area of pathological summaries, which involves documenting clinical information in a structured, standardized format rather than a free-form narrative. One recent study by Rajaganapathy *et al.* [69] proposed QLoRA-based framework to fine-tune LLMs such as LLaMA-2, to convert unstructured pathology reports into a structured synoptic format with predefined data elements (e.g., "Tumor Size") and their corresponding responses (e.g., "1.3×0.8×0.2 cm"). As pathological reports sometimes contain visual interpretation of results, it leads to another key application of LLMs in multi-modal summarization. In this direction, Ahmed *et al.* [70] introduced POLYPATH, a model that leverages the long-context capabilities of Gemini 1.5 Flash to generate reports directly from image patches of multiple whole-slide images (WSIs). *The Challenge of Lay Summarization* is another important research area where studies simplify complex scientific information for non-expert audiences like patients. To this end, Al-Hussaini *et al.* [71] proposed a two-stage approach for lay summarization, generating summaries with BART and simplifying them using MUSS. While this study improved readability, it reduced relevance and factual accuracy, highlighting the challenge of simplifying medical text without losing critical information, which remains a key challenge in patient-facing summarization tools.

3.2.3 Medical Evidence

Medical evidence, particularly systematic reviews of randomized controlled trials (RCTs), serves as the gold

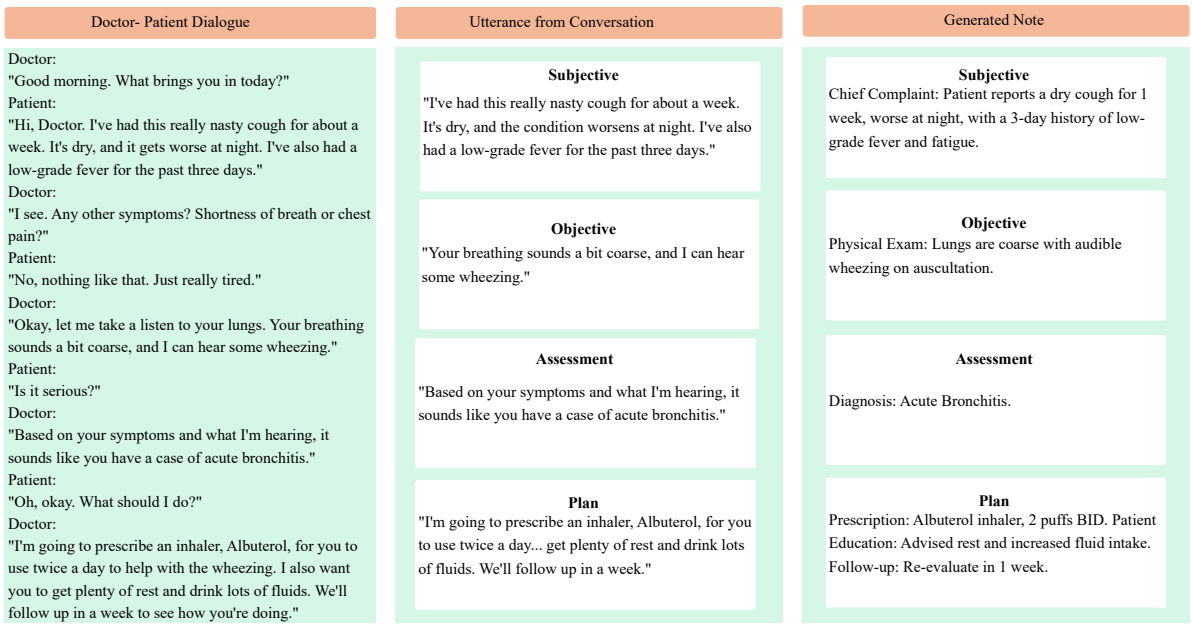


Fig. 6: Basic structure of a SOAP note

standard for guiding healthcare decisions. The time-consuming and complex process of creating these reviews using much faster and effective automated text summarization with Large Language Models (LLMs) presents a promising solution to streamline and accelerate the synthesis of medical evidence [72]. Recent studies [24, 39, 48, 50, 73] on medical summarization largely rely on powerful but proprietary (closed source) LLMs such as the GPT series. While effective, these models raise concerns about transparency, vendor dependence, and data privacy. Later, researchers offer a compelling alternative as open source LLMs for better transparency, control, and the ability to customize [20, 45, 69]. A recent study [74] on medical summarization utilized MedReview dataset and fine tuned 3 open source LLMs including PRIMERA, LongT5, and Llama-2, and demonstrated a significant improvement in summarization performance.

4 Approaches to Medical Report Summarization

This section aims to answer the research question *Q3* raised in Section 1 and provides a concise summary of deep learning architectures for the summarization of medical reports. In this review, we cover not only the basic deep learning architectures, but also the advanced and domain-adapted models that have shaped recent progress in the field. From the foundational CNN and LSTM-based methods to transformer-based architecture, generative models, and large language models, each approach reflects a step forward in addressing the complexity and sensitivity of medical texts. To get a summary of the deep learning models used in recent work for biomedical document summarization, readers can refer to Table 2.

4.1 Convolution Neural Networks (CNN)

Convolution Neural Networks (CNNs) [75] demonstrate strong effectiveness in medical text summarization. In text summarization task, CNNs use Convolution, pooling, and fully connected layers to learn local patterns inherent in the text, as shown in Figure 7. Recent developments have combined CNNs and domain-specific methods for extractive summarization of biomedical transcripts. As an example, Bedi *et al.* [76] designed a Deep-Dense LSTM + CNN (DDCNN) pipeline involving ELMo embeddings to summarize biomedical transcripts. When tested on MT Samples dataset, DDCNN obtained an average ROUGE score of 93.5%. Similarly, Dalal *et al.* [59] proposed MedicoVerse, a multi-stage system for pharmaceutical text summarization that combines SapBERT embeddings with the BART-large-CNN-SAMSum model, achieving ROUGE-L: 68%, and BERTScore: 0.76, demonstrating strong performance in pharmaceutical and regulatory domains. These works show that CNNs perform well in medical summarization with advanced embeddings or hybrid models, however, they may struggle with long-range dependencies compared to transformer-based approaches.

4.2 Recurrent Neural Networks (RNN)

Recurrent neural networks (RNNs) are a family of deep learning architectures specially designed to process sequential input data [77]. Through their internal memory or state, RNNs possess the unique ability to retain information from prior inputs, distinguishing them from feedforward neural networks. [90] A basic RNN has input, hidden, and output layers with recurrent connections, as shown in Figure 7. At each time step t, the hidden state and output are computed as:

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = \sigma_y(W_{hy}h_t + b_y) \quad (2)$$

Table 2: Summary of Deep Learning Models utilized for Biomedical Document Summarization

Model	Strength	Weakness	Variant	Results	Dataset	Referred In
CNN [75]	Efficient for local feature detection, scalable for multimodal data	Limited in handling sequential dependencies and long-range context	BART-large-cnn-samsum [59]	ROUGE-1: 69%, ROUGE-2: 56%, ROUGE-L: 68%, BERTScore: 0.76	PubMed	[59, 76]
RNN [77] *1. LSTM* [78] *2. BiLSTM* [79] *3. GRU* [80] *4. BiGRU* [80]	Effective for modelling long-term dependencies with gating mechanisms	Vanishing gradient problem limits long-range dependency capture	Encoder-decoder based LSTM [81]	ROUGE-1: 38.9%, ROUGE-2: 17.1%, ROUGE-L: 35.7%	CNN/DailyMail	[81, 82]
	Captures context from both directions, improves comprehension	Higher computational cost due to bidirectional processing	Encoder-decoder based BiLSTM [29]	ROUGE-1: 48.6%, ROUGE-2: 35.3%, ROUGE-L: 47.1%	Stanford Hospital PACS radiology report dataset	[29, 44, 83]
	Computationally efficient, fewer parameters than LSTM	Simpler than LSTM but may underperform on very long sequences	DenseNet121 (GRU-based encoder-decoder) [22]	ROUGE-2 F_1: 54.3%, BLEU-score: 28.3%	Indiana University Chest X-ray	[22]
	Utilizes full sequence context, enhances accuracy in temporal tasks	Increases complexity with bidirectional flows	Two-layer bidirectional GRU with SOAP format [52]	F_1 (Diagnosis): 0.78, F_1 (SOAP): 0.60	Proprietary dataset from Verilogue Inc. and YouTube	[52]
Transformer [84] *1. Encoder-Decoder (BART, FLAN-T5, Pegasus etc.)* *2. Domain-specific Models (BERT, BioBERT, MedCAT etc.)* *3. Hybrid Models (BigBird, LongFormer, BART-GPT2 etc.)*	Encoder-decoder structure, strong contextual and factual understanding	Requires fine-tuning, may hallucinate facts	FLAN-T5-Large Medalyz [45]	ROUGE-L: 37%, BERTScore: 0.65, SpaCy Sim: 0.84	Curated medical dataset from Stanford University	[37, 51, 61] [19, 42, 45] [85]
			BART [55]	ROUGE-L: 53.4%, METEOR: 61.1%	MM_CliCon Summation	
			BART and Pegasus [37]	Factual F_1: 0.76	MTS-DIALOG	
	Domain-adapted embeddings, superior on biomedical datasets	Limited to domain of pretraining, not general-purpose	Bio_ClinicalBERT [20]	F_1: 0.95	MIMIC-III	[20, 38, 66] [18, 67]
			UMLS-BERT [18]	NER F_1: 88.6, MedNLI Acc: 83.0	MIMIC-III v1.4	
			MedCAT [20]	ROUGE-1: 58%, BLEURT: 78%	MEDIQA-Chat 2023	
	Handles long inputs, incorporates structured knowledge, improves factual faithfulness	Complex integration, high resource demand	Hybrid Transformer LongFormer [61]	ROUGE-1: 46.6%, ROUGE-L: 41.9%	arXiv summarization	[39, 42, 61, 86]
			Hybrid BART–GPT2 [42]	ROUGE-L: 51.8%, BLEU: 34.7%, BERTScore: 0.92	MM_CliCon Summation	
Generative Models [87]	Learns joint data distribution, high linguistic fluency and adaptability	Can hallucinate or generate incoherent text	Hierarchical attentional model [49]	F_1: 72.8, ROUGE-L F_1: 55.4%	Anonymized clinical dialogues	[49, 53, 88]
Large Language Models (LLMs) [89]	Strong zero-shot capabilities, adaptable via prompting	May lack domain precision without fine-tuning	Instruction-tuned LLMs-GPT-3.5, GPT-4, LLaMA-3, MEDSAGE [34]	ROUGE-L: 13.3%, F_1 Score: 24.5%, BERTScore: 0.531	PriMock57	[34, 64]
	Reduces training cost, efficient adaptation to new domains	Sensitive to low-resource training data	PEFT-based LLMs– LoRA, QLoRA, Adapters [40]	ROUGE-L F_1: 55%, BERTScore F_1: 0.84	FigShare + MTS-Dialog	[40, 74]
	Enhances factual precision, supports multimodal input (text + image)	Integration complexity, dependent on retrieval quality	RAG & Multimodal LLMs - MedSumm [56]	ROUGE-L: 38.4%, BERTScore: 0.80	Code-mixed medical Q&A (Hindi–English)	[30, 42, 56]
	High factual accuracy, low hallucination, domain-aware generation	Requires domain-specific corpora	Domain-Specific LLMs - BioGPT, GPT-4, FLAN-T5 [8]	Lower error rates (13%) than medical experts (25%)	Open-i, MIMIC-CXR, MIMIC-III, MeQSum, ProbSum	[8, 27]

where σ_h is activation function, W are weights between different layers, and b is bias. RNNs have been utilized by a variety of architectures, including unidirectional LSTMs & GRUs and bidirectional variants Bi-LSTM & Bi-GRU.

4.2.1 Long Short-Term Memory (LSTM)

To address the vanishing gradient problem found in basic RNNs, a study by [78] developed LSTM networks, which are distinguished by their use of gating mechanisms that regulate the flow of information. The internal structure of each LSTM cell, shown in Figure 7, consists of input, forget, and output gates that regulate how new information is added, how past information is discarded, and how the cell state is produced [91]. For biomedical text summarization, several studies have employed LSTMs as a benchmark approach for evaluation. Li *et al.* [81] proposed a Key Information Guiding

Network (KIGN) using an LSTM-based encoder–decoder model to improve abstractive summarization, reporting ROUGE-1: 38.95%, ROUGE-2: 17.12%, and ROUGE-L: 35.68% on CNN/DailyMail corpus. Similarly, another work by Krishna et al. [82] developed a modular summarization pipeline for the end-to-end generation of SOAP notes from doctor–patient conversations using hierarchical LSTM and BERT–LSTM models, achieving ROUGE-1 scores of 57.5% on the medical dataset and 57.4$ on the AMI corpus.

4.2.2 Bidirectional LSTM (BiLSTM)

As Illustrated in Figure 7, Bidirectional LSTM (Bi-LSTM) is an extension of LSTM and is characterized by an external recurrence between layers as they work through the input sequence in both directions, much like BiRNNs [79], resulting in a more context-aware model. Bi-LSTMs are also very common in NLP-based text summarization tasks for clinical documents. Zhang *et al.* [29] proposed a Bi-LSTM–LSTM encoder–decoder with a pointer-generator network for radiology report summarization, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 48.56%, 35.25%, and 47.06%, respectively, on the Stanford Hospital PACS dataset. Similarly, See *et al.* [83] proposed a Bi-LSTM–based encoder-decoder architecture augmented with attention, and reported ROUGE-L = 36.38%, establishing a strong baseline for hybrid extractive–abstractive summarization. Another work by Schloss and Konam [44] utilized a stacked Bi-LSTM with ELMo embeddings to label utterances as SOAP categories, forming a foundation for structured clinical note generation.

4.2.3 Gated Recurrent Unit (GRU)

GRU reduces LSTM architecture by fusing input and forget gates into one update gate [80], making the model simpler and computationally more efficient. GRU structure is made up of two gates: the reset gate and the update gate [92], as shown in Figure 7. As far as biomedical report summarization is concerned, several studies have employed GRU architectures for improvement over LSTMs. In [22], a multimodal radiology report summarization model combining a GRU-based encoder–decoder with DenseNet121 was proposed, using both report findings and X-ray images to generate impressions. The model achieved ROUGE-2 F_1 scores of 54.38% and 48.98%, and BLEU/METEOR scores of 28.32% and 26.10% on the Indiana University Chest X-ray and MIMIC-CXR v2.0 datasets, respectively.

4.2.4 Bidirectional GRU (BiGRU)

As illustrated in Figure 7, BiGRU is an extension of GRU that utilizes two GRU layers (forward and backward) to capture both past and future context, making it highly effective for tasks such as text processing where bidirectional context is important. In biomedical clinical context, Jeblee *et al.* [52] introduced a multi-component pipeline using Bi-GRU for utterance classification in automated doctor-patient dialogue-based clinical note-taking. Assessed on 800 patient–clinician conversations from Verilogue Inc. and simulated YouTube interactions, their model achieved F_1 scores of 0.78 (Diagnosis), 0.64 (Entity Extraction), 0.77 (Modality), and 0.60 (SOAP), demonstrating Bi-GRU's ability in capturing conversational context for clinical note generation.

4.3 Transformer Architecture

The Transformer model proposed by Vaswani *et al.* [84], unlike earlier CNN- and RNN-based architectures, eliminates recurrence and convolutions, leading to significant advancements in clinical tasks. It solely depends on a self-attention mechanism for assigning importance to various words in a sequence, letting the model concurrently process all the input elements, greatly enhances training efficiency and captures long-range dependencies. Based on Transformer architecture, numerous pre-trained models have been established and fine-tuned for different tasks and domains.

4.3.1 Encoder-Decoder Models

Most of the Transformer models use an encoder-decoder architecture and are thus highly suitable for sequence-to-sequence problems such as summarization. Some well-known models include BART [93], T5 [94], Pegasus [85] and Longformer-Encoder-Decoder (LED) [61] (Figure 7). Recent work highlights their effectiveness in biomedical and clinical domains. For example, one recent study [45] adapted FLAN-T5-Large in Medalyze for medical conversation summarization (SpaCy similarity: 0.88). Another work by Michalopoulos *et al.* [26] presented MedicalSum, a dual-encoder Transformer model for abstractive summarization of doctor–patient conversations (ROUGE-1 F_1: 68.2%).

Later, multimodal encoder–decoder architectures became the backbone of biomedical summarization. Examples include HealthAlignSumm [51], which combines T5 and BART encoders (ROUGE-L: 46.8%, MedAlign F_1: 72.4%), and Extended BART with cross-modal attention [55], which achieves ROUGE-L: 53.37% and METEOR: 61.09% on the MM_CliConSummation dataset. Moreover, fine-tuning the base models in specific domains has proven to be very effective. Sharma *et al.* [19] fine-tuned BART-Large, achieving ROUGE-1 and ROUGE-2 scores of 59.11% and 23.69%, respectively, while Abacha *et al.* [37] fine-tuned BART and Pegasus on the MTS-DIALOG dataset, reporting a factual F_1 score of 0.77 with low hallucination (0.03) and omission (0.33) rates.

4.3.2 Domain-Specific Models

To address the domain-specific requirements of clinical documents, numerous studies have explored domain-specific models based on the general BERT architecture, as shown in Figure 7. BERT utilizes a transformer-based architecture with bidirectional self-attention that effectively captures contextual relations. Examples of BERT-based models employed in the clinical domain include BioBERT [67] and Bio_ClinicalBERT [66], both trained on clinical and biomedical corpora, as well as UMLS-BERT [18], a BERT-based model enhanced with Unified Medical Language System (UMLS) embeddings

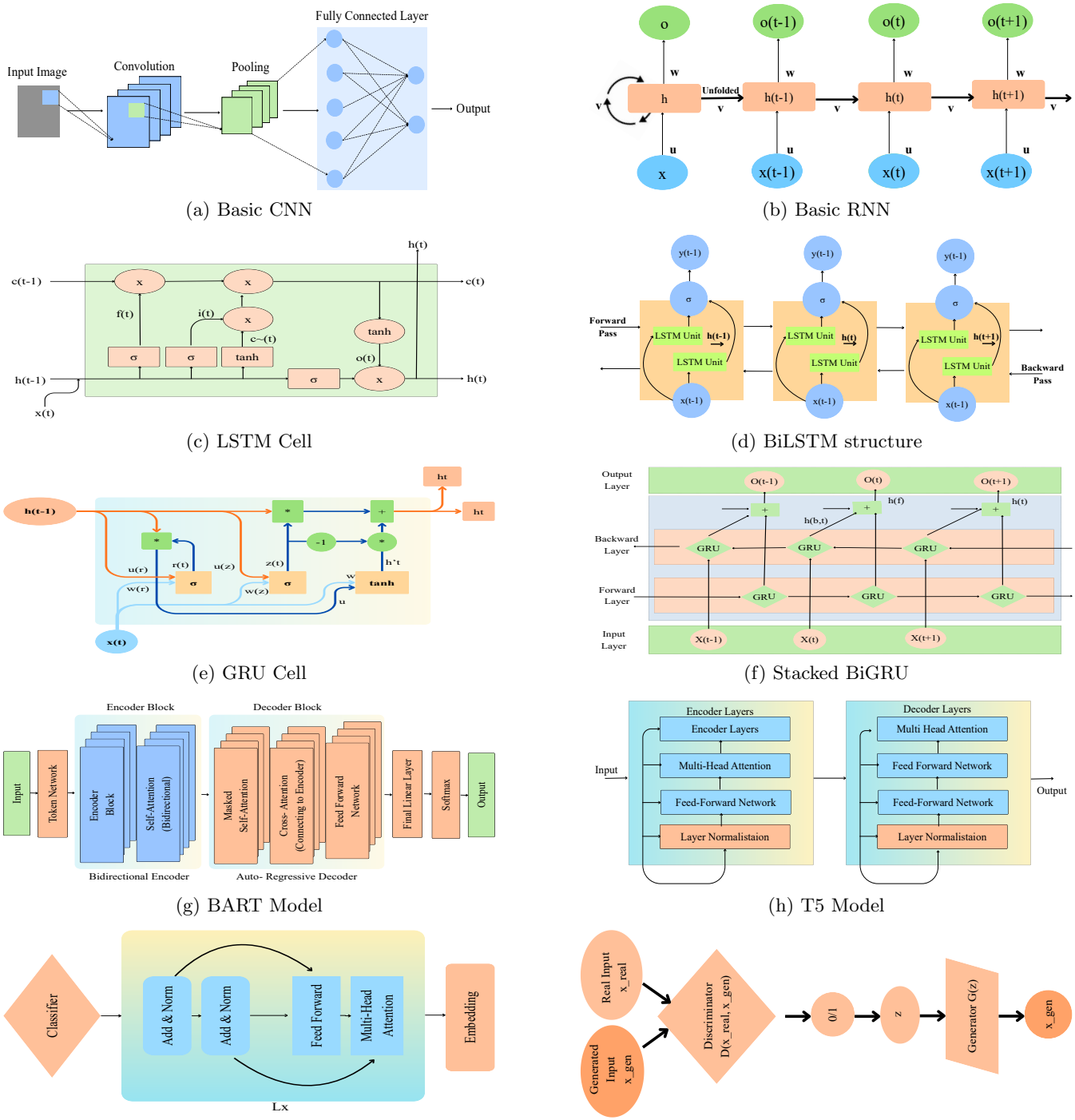


Fig. 7: Architecture of various deep leaning models employed for biomedical document summarization

that achieved an NER F_1 score of 88.6 and a MedNLI accuracy of 83.0%, and MedCAT [20], which integrates Bio_ClinicalBERT-based embeddings with clinician feedback and achieved an F_1 score of 0.95.

4.3.3 Hybrid Models

In addition to pure encoder–decoder or domain-specialized architectures, hybrid frameworks integrate architectural advancements or combine multiple modalities to extend the capacity of Transformer architectures. Some well-known hybrid architectures include BigBird [86] and Longformer [61], which overcome the quadratic complexity of vanilla attention to facilitate processing of very long sequences. Longformer-Encoder-Decoder (LED) [61] is specifically suited for long-document summarization tasks, achieved ROUGE-1: 46.63% and ROUGE-L: 41.83% on arXiv Summarization dataset [95] Moreover, multimodal and knowledge-infused architectures also extend this field of transformers. Tiwari *et al.* [42] proposed a hybrid adapter-based BART–GPT-2 model, MM-CliConSummation, which achieved a ROUGE-L score of 51.7%, a BLEU score of 34.7%, and a BERTScore of 0.91. Similarly, SummQA [39], an in-context learning paradigm based on an ensemble of GPT-4 and BioBERT, reported ROUGE-1, ROUGE-2, and

ROUGE-L scores of 52.76%, 37.82%, and 43.60%, respectively. Overall, these advances reflect a shift toward scalable hybrid and multimodal transformers that handle long sequences, integrate medical knowledge, and improve factual accuracy in clinical narratives.

4.4 Generative Models

A generative model learns the underlying patterns and distribution of training data to generate new, similar content [87]. Figure 7 shows the general structure of a generative model. Such models have shown great importance in summarization tasks where models can create new text sequences. Numerous research works have utilized these models for summarizing text over specialized domains. For example, a hierarchical attentional model [88] for multimodal meeting summarization achieved ROUGE-1: 53.3%, ROUGE-2: 13.5%, and ROUGE-L: 26.9%, while a similar model by Joshi *et al.* [49] fine-tuned on telemedicine dialogues, captures local conversational structures and achieves concept F_1: 72.8 and ROUGE-L F_1: 55.4%, highlighting improvements in factual consistency and concept preservation in medical contexts. Moreover, Molenaar *et al.* [53] propose a speech-to-text pipeline using Frog, FRED, and Ollie analyzers to enable automated clinical report generation, reporting an F_1 score of 0.24.

4.5 Large Language Models (LLMs)

Recent advancements in medical text summarization include the emergence of large language models (LLMs), which scale transformer architectures to billions of parameters and leverage massive pretraining corpora along with sophisticated instruction-tuning methods [89]. Compared to standard encoder–decoder transformers, LLMs achieve flexibility via large-scale pretraining and prompt-based adaptation, enabling high-quality performance on a wide range of tasks with minimal supervision, including biomedical text summarization, medical dialogue generation, and clinical decision support. The architecture of an LLM is similar to that of encoder–decoder transformers such as T5 and BART, as shown in Figure 7. Based on the reviewed literature, we categorize LLMs into several categories, including *instruction-tuned*, *parameter-efficient fine-tuning (PEFT)*, *retrieval-augmented*, and *domain-specific* models.

4.5.1 Instruction-Tuned and Foundation Models

Instruction-tuned LLM models such as GPT-3.5, GPT-4, and LLaMA-3 are excellent zero- and few-shot models in biomedical summarization [64]. Some recent works include *MEDSAGE* [34], which combines ASR models (Whisper and Wav2Vec2) with LLaMA-3-8B and Mistral-7B, achieving a ROUGE-L score of 13.3% and a BERTScore of 0.53, and *uMedSum* [17], which integrates GPT-4 and LLaMA-3 using in-context learning for hybrid abstractive–extractive summarization.

4.5.2 Parameter-Efficient Fine-Tuning (PEFT)

Despite their strong performance, fully fine-tuning billion-parameter LLMs remains computationally expensive. Parameter-efficient fine-tuning (PEFT) methods such as LoRA, QLoRA, and adapter-based approaches have therefore emerged as practical alternatives [96, 97]. For instance, Neupane *et al.* [40] introduced *CLIN-ICSUM*, which fine-tunes LLaMA-3 with LoRA on patient–doctor conversations and results in a ROUGE-L F_1 of 55% and a BERTScore of 0.84. Similarly, another study by Zhang *et al.* [74] fine-tuned LoRA with Long-T5 and LLaMA-2 on the MedReview dataset, reporting a ROUGE-L score of 24.6%, a METEOR score of 28.3%, and a PICO-F_1 score of 51.4%. Effective domain adaptation is made possible by these PEFT techniques, which also significantly lower computing costs while maintaining base model parameters.

4.5.3 Retrieval-Augmented and Domain-Specific LLMs

In recent years, to overcome the hallucination problem of LLMs, they have been coupled with retrieval-augmented generation (RAG) [102]. Several recent studies adopted RAG as a framework for summarization of medical text. In this direction, Neupane *et al.* [30] proposed *MedInsight*, which integrates GPT-3.5-Turbo and Mistral-7B with RAG pipelines and acheives RAGAS similarity of 0.93 and correctness of 0.84. Parallel to this, two multimodal architectures were proposed: *MedSumm* [56], which uses LLaMA-2, Mistral-7B, and ViT-based visual embeddings for code-mixed Hindi–English clinical query summarization, and *CLIPSyntel* [54], which unites CLIP with LLMs such as LLaMA-2, Vicuna, and FLAN-T5 (ROUGE-L: 38.4% and BERTScore: 0.80).

Furthermore, to improve performance in biomedical-based NLP tasks, some recent works have also explored domain-specific models such as BioGPT, MedAlpaca, and MedAlpaca-7B, which are fine-tuned in biomedical and clinical datasets [65, 103]. Tariq *et al.* [27] and Van Veen *et al.* [8] showed that conditioning models such as GPT-4, FLAN-T5, and BioGPT combined with in-context learning (ICL) or parameter-efficient fine-tuning (QLoRA) improves clinical summarization by reducing error and hallucination rates compared to ChatGPT and even medical experts.

In summary, Table 3 highlights the significant work in which fine-tuning the base model improves performance on biomedical and clinical summarization tasks. Some significant enhancements achieved through fine-tuning include *SumGPT* over GPT-2 [24] (with ROUGE-L increasing from 36.19% to 85.14%, an improvement of more than 100%), *HealthAlignSumm* over BART [51] (with ROUGE-L increasing from 29.88% to 50.46%, approximately 69%), and Transformer + Pointer-Generator over RNN [46] (with ROUGE-L improving from 50.1% to 65.4%, a 30% increase, and F_1 improving from 64.9% to 74.6%, a 15% increase).

Table 3: Comparison of Base and Fine-tuned Models for Biomedical Document Summarization

Ref.	Base Model	Fine-tuned Model	Dataset	Base Model			Fine-tuned		
				R-L	F_1	BERT	R-L	F_1	BERT
[17]	GPT-4 (Standard Prompting + ICL)	uMedSum Framework on GPT-4	MIMIC-III	20	-	86	25	-	89
[20]	MetaMap	MedCAT (Self-Supervised)	MIMIC-III	-	49.5	-	-	73.8	-
[22]	Monomodal (Text-only) Seq2Seq	Multimodal Seq2Seq (Text + Image)	Indiana University Chest X-Rays	-	52.18	-	-	54.38	-
[8]	FLAN-T5 (ICL)	FLAN-T5 (QLoRA Fine-tuning)	Open-i, MIMIC-CXR, MIMIC-III, MeQSum, ProbSum	58	-	93	65	-	94
[43]	BART-base	BART + section-specific Cross-Attention	proprietary, de-identified dataset from Abridge AI	42.27	61.72	-	43.28	62.83	-
[48]	BART-LSG-conv (Fine-tuned)	ChatGPT (Zero-shot)	anonymized clinical dialogues	18.46	-	-	20.06	-	-
[18]	Bio_ClinicalBERT	UmlsBERT (UMLS Knowledge Augmented)	MIMIC-III dataset	-	93.6	-	-	94.4	-
[51]	HealthSumm (Multimodal BART)	HealthAlignSumm (DPO Alignment)	MCDH (Multimodal Codemixed Dialogue Summarization in Healthcare)	29.88	-	-	50.46	-	-
[54]	GPT-3.5 (Text-only)	CLIPSyntel Framework (CLIP + LLM)	MMQS (Multimodal Medical Question Summarization)	39.6	-	90.1	40.9	-	92.1
[98]	LED-large (Vanilla)	LED-large (In-Domain Pre-trained)	doctor-patient conversations	25.43	-	-	27.19	-	-
[39]	GPT-3 (Zero-shot)	GPT-4 (In-Context Learning w/ Prompt Selection)	MEDIQA-Chat 2023 Task B (ACI-BENCH)	34.80	-	-	43.60	-	-
[41]	Fine-tuned LED	GPT-4 (In-Context Learning)	MEDIQA-Chat 2023 Task B (ACI-BENCH)	61.7	70.7	-	62.7	71.2	-
[37]	BART-large	BART (Pre-finetuned + Guided + Augmented)	MTS-DIALOG	26.91	-	-	34.9	-	-
[61]	PEGASUS	LED-large (Longformer Encoder-Decoder)	arXiv summarization dataset	38.83	-	-	41.83	-	-
[46]	Hierarchical RNN + Pointer-Generator	Transformer + Pointer-Generator	800,000 orthopedic surgery encounters	50.1	64.9	-	65.4	74.6	-
[99]	BART-large (Single-stage)	BART-large (Multistage "Chunking")	History of Present Illness (HPI) summaries	-	38.47	-	-	40.93	-
[74]	LongT5-xl (Zero-shot)	LongT5-base (Fine-tuned)	MedReview	14.72	-	-	24.61	-	-
[26]	Enarvi-PG (Transformer + Pointer-Gen)	MedicalSum (UMLS-Guided Transformer)	MEDIQA 2021 radiology report (RAD)	34.21	-	-	35.22	-	-
[45]	GPT-4	Flan-T5-Large (Fine-tuned "M-Passage")	Stanford University study	64.4	-	54.53	37.28	-	65.53
[34]	Llama-3-8B (Zero-shot)	Llama-3-8B (Fine-tuned on Augmented Data)	NoteChat-1000 dataset	11.8	19.9	-	13.1	22.8	-
[88]	Pointer-Generator Network (PGN)	Multi-modal Hierarchical Attention Model	AMI Meeting Corpus	23.81	-	-	26.9	-	-
[55]	KM-CliConSummation	EDI-Summ (Image-Guided Encoder-Decoder)	MM-CliConSummation dataset	50.87	-	-	53.37	-	-
[82]	Oracle Extractive	Cluster2Sent (Modular BART)	doctor-patient conversations with corresponding SOAP notes	32.87	-	-	38.38	-	-
[100]	PEGASUS	FROST (Entity Chain Planning + Pretraining)	XSum dataset	39.64	-	-	39.76	-	-
[83]	Seq2Seq + Attention	Pointer-Generator + Coverage	CNN / Daily Mail dataset	28.83	-	-	36.38	-	-
[101]	BART	CTRLSUM (Keyword-conditioned BART)	CNN/DailyMail	22.50	-	10.7	24.24	-	16.5
[42]	GPT-2, BART, T5, and MAF	BART with Multi-modality Fusion of text, visual, and knowledge	MM-CliConSummation	20.37	22	66	51.77	56.31	91.8
[60]	BERT and BART	BERT and BART (multi-attention mechanism)	Reddit, XSum, CNN/DM, WikiHow, NYT, PubMed	38.76	-	-	40.66	-	-
[81]	LSTM, KIGN	Attentional encoder-decoder + pointer	CNN/Daily Mail	33.42	-	-	35.68	-	-
[56]	Mistral-7B (Unimodal)	MedSumm (LLAMA-2, Multimodal)	MMCQS	33.96	-	74	38.41	-	80
[11]	BART, GPT-3, GPT-4, RoBERTa, SciBERT	Fine-tuning (CONFIT, Davinci), RoBERTa classification, few-shot in-context learning (GPT-4)	MTS-Dialog (MEDIQA 2023)	32.14	-	-	33.22	-	-
[24]	VisualGPT (GPT-2 + ViT)	SumGPT (T5 + Vision Transformer)	Radiology images/reports	36.19	-	-	85.14	-	-
[71]	BART	BART + MUSS	BioLaySumm 2023 (PLOS and eLife articles)	43.7	-	86.4	46.0	-	87.1

5 Evaluation Methods for Medical Report Summarization

Assessing medical report summarization is particularly challenging due to the complexity of the domain, its safety-critical nature, and the diversity of documentation styles (e.g., discharge summaries, radiology reports, SOAP notes, and conversational transcripts). In this section, our objective is to answer the research question *Q4*, raised in Section 1, by providing a categorization of the evaluation methods used in the existing literature into three main categories: automated general-domain metrics, domain- and task-specific measures, and human evaluation techniques. Table 4 summarizes quantitative results from recent biomedical summarization literature, highlighting the evaluation metrics and identifying the best-performing models for each metric across different datasets.

5.1 Automated General-domain Metrics

5.1.1 Lexical overlap measures

Lexical overlap measures like ROUGE, BLEU, and METEOR compare the machine-generated summary with human-created references by estimating word- or phrase-level overlap. Here, ROUGE [104] is the most widely used measure which comes in 3 forms: *ROUGE-1:* It estimates unigram overlap, *ROUGE-2:* It estimates bigram overlap, and *ROUGE-L:* It captures the longest common subsequence (LCS) as defined in Eq. 3.

$$\text{ROUGE-L} = \frac{(1+\beta^2)\cdot \text{Recall}_{LCS}\cdot \text{Precision}_{LCS}}{\text{Recall}_{LCS}+\beta^2\cdot \text{Precision}_{LCS}} \tag{3}$$

where $\text{Recall}_{LCS} = \frac{\text{LCS}(X,Y)}{|Y|}$ and $\text{Precision}_{LCS} = \frac{\text{LCS}(X,Y)}{|X|}$, X is the generated summary, Y is the reference, and β is a constant to balance precision and recall.

Moreover, another measure ROUGE-LSum, a variant of ROUGE using sentence-tokenization, is especially conducive to multi-document summarization. BLEU [105], originally developed for machine translation, emphasizes n-gram precision and, although less prevalent than ROUGE in summarization, is still employed to estimate adequacy and fluency. METEOR [106] broadens such methods by incorporating stemming, synonym matching, and paraphrasing, enabling a closer semantic match, as defined in Eq. 4.

$$\text{METEOR} = (1-\text{Penalty})\cdot \frac{10\cdot \text{Precision}\cdot \text{Recall}}{\text{Recall}+9\cdot \text{Precision}} \tag{4}$$

5.1.2 Semantic and factual consistency measures

These measures depend on contextual embeddings and natural language inference (NLI). One well-known such measure is BERTScore F_1 [63] which uses contextual embedding to estimate semantic similarity as given in Eq. 5.

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}}\cdot R_{\text{BERT}}}{P_{\text{BERT}}+R_{\text{BERT}}} \tag{5}$$

where $P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j\in\hat{x}} \max_{x_i\in x} x_i^\top \hat{x}_j$ is BERT-precision, $R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i\in x} \max_{\hat{x}_j\in\hat{x}} x_i^\top \hat{x}_j$ is BERT-recall, x represents the reference text, and $\hat{x}$ represents the candidate text.

Another metric SummaC [107] computes the factual consistency between the source and the generated summary as given in Eq. 6

$$\text{SummaC} = \frac{1}{N}\sum_{i=1}^{N}\text{NLI}(s_i, D) \tag{6}$$

where s_i denotes the summary sentence, D represents the source document, and NLI denotes the entailment score obtained from a Natural Language Inference (NLI) model.

Other recent factuality-centric measures, such as QuestEval [108] and RAGAS [109], ensure that generated summaries do not contain information unsupported by the source document.

5.2 Domain-specific and Task-specific Metrics

In clinical and biomedical summarization, evaluation requires more than general language quality, it must also account for clinical accuracy and task relevance. Metrics such as Concept F_1 and Medical Concept Coverage (also known as quickUMLS and UMLSScorer) measure how precisely and completely medical entities such as diagnoses, medications, or symptoms are captured in a summary, ensuring that no vital information is lost. Negation F_1 addresses another important aspect by checking whether models correctly identify and represent negated concepts (for example, distinguishing "no evidence of pneumonia" from "evidence of pneumonia"). Similarly, the Domain-Specific Entity (DSE) F_1-score targets the correct identification of entities tied to clinical tasks, such as conditions, treatments, or patient attributes. For clinical trials, PICO-F_1 [72] is widely used, which evaluates the ability of a model to extract and summarize the four essential components - Patient, Intervention, Comparison, and Outcome, that define trial evidence. Together, these domain- and task-specific metrics provide a more faithful and safety-focused evaluation than general-purpose approaches.

5.3 Human Evaluation Methods

With the spread of computer-based measures, human assessment by subject-area experts is still the gold standard. Measures such as Likert scales [110] and Fleiss Kappa [111] are utilized to measure human judgments on numerous quality aspects, i.e., fluency, relevance, coherence, and clinical utility. Clinical evaluation tools, such as the PDQI-9 [112] and Revised-IDEA [113], can be used to inform expert judgments. In addition, new methods such as LLM-as-Judge (LLM-EVAL) [114] are being developed to take advantage of the strength of LLMs to automate human-like evaluation.

6 Discussions and Findings

This section summarizes the findings of our comprehensive review, including the shift in trends from CNN-/RNN-based models to transformer- and LLM-based architectures, an analysis of the reviewed literature, and the identification of research gaps for future directions.

6.1 Summary of Existing Studies

This review covers quality research articles published in the last decade for biomedical document summarization. Table 5 summarizes these articles, highlighting how research in this domain has progressed and evolved from traditional CNN/RNN approaches to transformers, LLMs, and their advanced fine-tuned architectures, representing a paradigm shift in biomedical summarization. To ensure the quality of the articles reported in Table 5, we also provide the rankings and indexing of the corresponding conferences and journals, including most CORE A*/A[1] and SCI-indexed[2] venues. Early research (2017–2019) has mainly employed CNN- or RNN-based variants such as LSTMs, BiLSTMs, and BiGRUs [52,

[1] https://portal.core.edu.au/conf-ranks/
[2] https://mjl.clarivate.com/search-results

Table 4: Common Evaluation Metrics for Biomedical Summarization, highlighting the best score for each metric

Metric	Description	Model	Best Score	Dataset
ROUGE-1	Measures overlap of individual words (unigrams)	SumGPT [24]	85.14%	Indiana Chest X-ray (Link)
ROUGE-1 F_1	Harmonic mean of ROUGE-1 precision and recall	Transformer encoder-decoder [26]	68.22%	MEDIQA 2021 RAD (Link)
ROUGE-2	Measures overlap of word pairs (bigrams)	SumGPT [24]	84.71%	Indiana Chest X-ray (Link
ROUGE-2 F_1	Harmonic mean of ROUGE-2 precision and recall	Attention-based encoder-decoder + GRU [22]	54.38%	MIMIC-CXR (Link)
ROUGE-L F_1	Harmonic mean of ROUGE-L precision and recall	GPT-3 + PEGASUS [73]	58.44%	Private medical conversation dialogues
ROUGE-L	Measures longest matching sequence of words	SumGPT [24]	85.14%	Indiana Chest X-ray (Link)
F_1-Score	Harmonic mean of precision and recall	MedCAT (Self-Supervised) [20]	94.17%	MIMIC-II (Link)
BERTScore	Measures semantic similarity using BERT embeddings	CLIPSyntel Framework (CLIP + LLM) [54]	0.92	MMQS dataset (Link)
BLEU	Measures n-gram precision between generated and reference texts	SumGPT [24]	84.7%	Indiana Chest X-ray (Link

83] for medical text summarization. Even traditional machine learning architectures have helped researchers generate extractive summaries [52]. Following this, several studies used Pointer-Generator Networks (PGNs) to generate both abstractive and extractive summaries from medical text with the aim of improving ROUGE scores [49, 88]. After the emergence of transformer architectures and attention mechanisms, biomedical summarization has undergone a revolutionary transformation. From 2020 onward, several transformer-based architectures including BERT [115], BART [38, 51, 99, 116], PEGASUS [38], FLAN-T5 [41], and LED [41] are employed using pre-trained embeddings, enabling more nuanced contextual representations of medical text. Notable works in this direction include Zhang et al. [99], reporting BART-Large & BigBird with an improved ROUGE-L of 34.12%, and Abacha et al. [38] utilizing BART-LARGE & PEGASUS with ROUGE-L of 34.9%. In addition, several studies have fine-tuned base transformer architectures and proposed domain-specific biomedical models such as BioMedBERT [115], fine-tuned BART [99], and MedicalSum [26].

Recent breakthroughs from 2023 to 2025 have been propelled by cutting-edge large language models (LLMs) such as GPT-4 [17, 41], LLaMA-2 [51], LLaMA-3 [17, 34], Mistral-7B [34], Gemini 1.5 Flash [70], and Gemma-7B [34]. Moreover, most recent studies focus on extensive fine-tuning of these LLMs using in-context learning (ICL), multimodal extensions, and augmentation in specific domain, demonstrating improved generalization and adaptability across various biomedical summarization tasks [34, 41, 70]. Some of the most notable works in this direction include HealthAlignSumm [51], which reported a ROUGE-L score of 50.46%, uMedSum [17], achieved a BERTScore of 0.87, and SumGPT, reported a ROUGE-L score of 85.14%.

6.2 Analysis of Literature

This section presents an analysis of the reviewed research articles and highlights current research trends to better understand advancements in biomedical document summarization. The manner in which research in this domain is disseminated is characterized by a high degree of dependence on "timely" academic conferences. As shown in Figure 8(a). the bulk of research, approximately 65%, is published in conferences, followed by 28% in journals, and 7% in arXiv indicating most recent "cutting-edge" pre-prints breakthroughs in this domain. In addition, the majority conference papers are concentrated in prestigious NLP and AI venues, including ACL, EMNLP, AAAI, ACM CIKM, and NAACL, as well as specialized biomedical ACL events such as BioNLP and ClinicalNLP (Figure 8(b)). Furthermore, all selected research papers meet strict quality criteria, as they are published in SCIE Q1/Q2 journals or CORE A*/A-ranked conferences (Figure 8(e)). The year-wise trend, as shown in Figure 8(d), indicates that research output remained relatively modest until 2019, followed by a steady increase in 2020–2021 and a sharp acceleration from 2023 onward, reflecting the widespread adoption and advancement of large language models in medical research. As for journal publications (Figure 8(c)), these venues include specialized journals in bio-informatics and digital medicine (e.g., NPJ Digital Medicine and NPJ Health Systems), as well as multidisciplinary outlets such as Scientific Reports and IEEE Access.

Figure 8 also highlights the tools driving this research through a model architecture heatmap, where research papers are grouped by their architectural lineage and model implementations. As shown, this field is currently in a transitional phase. Despite the "Transformer Revolution," there remains a notable legacy density for LSTMs, with 5 unique papers utilizing this sequential approach. This is because they are the "de facto" baseline against which all other models are evaluated. In the middle section of this heatmap, Encoder-Decoder architectures such as BART and T5 dominate, with 7 total papers. These are the go-to tools for any task requiring high-precision output, such as summarization of radiology reports, where strict mappings are required without any chance for "hallucination." At the most modern section of this heatmap, the GPT and LLaMA architectures represent a massive shift. These Decoder-Only models are increasingly being preferred for noisy, unstructured data like doctor-patient dialogues, where

Table 5: Summary of Biomedical Document Summarization Approaches, highlighting year-wise evolution from CNNs and RNNs to Transformers, LLMs, and Fine-Tuned architectures

Reference	Year	Base model	Proposed model	Results	Publisher (Ranking / Indexing)
See et al. [83]	2017	BiLSTM encoder + LSTM decoder	Pointer-Generator + Coverage	ROUGE-1: 39.53%, ROUGE-2: 17.28%, ROUGE-L: 36.38%	ACL (Core A*)
Li et al. [81]	2018	Attention + encoder–decoder	KIGN + Prediction-Guide (Extractive + Abstractive)	ROUGE-1: 38.95%, ROUGE-2: 17.12%, ROUGE-L: 35.68%	NAACL (Core A)
Jeblee et al. [52]	2019	Rule/lexicon-based clinical models (BiGRU, SVM, LR)	Context-aware clinical note-taking pipeline	Utterance F_1: 0.71, Entity F_1: 0.64, Modality F_1: 0.77, SOAP F_1: 0.60	ACL LOUHI Workshop
Li et al. [88]	2019	Pointer-Generator Network (PGN)	TopicSeq + VFOA Hierarchical Attention	ROUGE-1: 53.3%, ROUGE-2: 13.5%, ROUGE-L: 26.9%, BLEU: 40.98%	ACL (Core A*)
Joshi et al. [49]	2020	Pointer-Generator (PGN) + Coverage	Dr. Summarize	ROUGE-L: 55.4%, ± 0.9, Concept F_1: 72.8 ± 0.8, Negation F_1: 67.3 ± 3.3	EMNLP (Core A*)
Chakraborty et al. [115]	2020	BERT-Large	BioMedBERT	Precision: 78.04, Recall: 82.01, F_1: 92.46	COLING (Core B)
Zhang et al. [99]	2021	BART-Large, BigBird	Multistage Fine-tuned BART	ROUGE-1: 33.3%, ROUGE-2: 11.88%, ROUGE-L: 34.12%;	EMNLP (Core A*)
Delbrouck et al. [22]	2021	Text-only BiGRU encoder–GRU decoder + attention	Multimodal encoder–decoder (Multi E5 allv)	BLEU: 28.32%, METEOR: 25.30%, ROUGE-2 F_1: 54.38%;	ACL BioNLP Workshop
Michalopoulos et al. [26]	2022	Transformer based Seq2Seq model	MedicalSum	ROUGE-L (HPI: 35.22%, PE: 64.48%, AP: 37.34%, RAD: 26.06%)	EMNLP (Core A*)
Knoll et al. [116]	2022	Fine-tuned BART Seq2Seq model	Real-time medical note generation system	Generated-note usage: 13–97%, ROUGE-L recall overlap: 13–50%	NAACL (Core A*)
Abacha et al. [37]	2023	BART-Large, Pegasus-Large	BART-PFT-GS-DA (Fine-tuned BART)	ROUGE-1: 42.52%, ROUGE-2: 17.5%, ROUGE-L: 34.9%, BERTScore-M2: 0.41, BLEURT: 51.23%	EACL (Core A*)
Giorgi et al. [41]	2023	GPT-4, LED, FLAN-T5	Fine-tuned GPT-4 few-shot ICL with Instructor retrieval	ROUGE-1: 44.66%, ROUGE-2: 22.82%, ROUGE-L: 38.37%	ACL ClinicalNLP Workshop
Ghosh et al. [51]	2024	BART-Base, T5-Base, LLaMA-2	HealthAlignSumm	ROUGE-1: 60.2%, ROUGE-2: 38.54%, ROUGE-L: 50.46%	EMNLP (Core A*)
Nagar et al. [17]	2024	GPT-4, LLaMA-3-8B	uMedSum	ROUGE-LSum: 25%, BERTScore: 0.87	ACL (Core A*)
Sultan et al. [24]	2025	GPT, BERT, ViT multimodal baselines	SumGPT (T5 + ViT)	ROUGE-1: 85.14%, ROUGE-2: 84.71%, ROUGE-L: 85.14%, BLEU: 84.70%	IEEE Access (SCI, Q1)
Ahmed et al. [70]	2025	Gemini 1.5 Flash	PolyPath (Fine-tuned Gemini)	ROUGE-L: 49.8%, METEOR: 50.3%	Elsevier Modern Pathology (SCI, Q1)
Binici et al. [34]	2025	LLaMA-3-8B, Mistral-7B, Gemma-7B	MEDSAGE (Fine-tuned and augmented)	F_1: 24.2%, ROUGE-L: 13.3%, BERTScore: 0.53	AAAI (Core A*)

the inherent reasoning power of a model becomes more important compared to pattern matching.

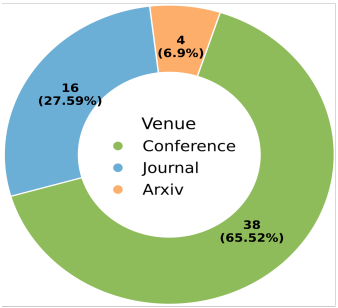
6.3 Research Gaps

Despite significant advances in biomedical document summarization using deep learning architectures, several critical challenges persist that hinder the field's progress toward reliable and clinically deployable systems. Some identified research gaps are listed in this section.

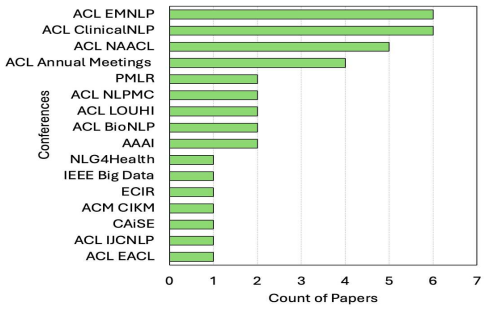
Limited Domain-Specific and Large-Scale Datasets: One of the most significant barriers in biomedical document summarization is the lack of large-scale and domain-specific publicly available datasets that capture the complexity of medical text. Most of the available work has relied on a small number of publicly available datasets, which are specific for some regions and accessible only for the English language. In addition, many existing datasets used in biomedical summarization do not capture the full complexity of real-world healthcare, such as multimodal data (combination with text, image etc.) and longitudinal patient histories. This limitation hinders the ability of deep learning models to generalize effectively and be fine-tuned for domain-specific healthcare settings.

Factual Accuracy and Hallucination Control: Factual inaccuracies and hallucinations, where models produce made-up or missing information, are a major challenge of abstractive summarization, which tends to maximize fluency at the expense of source text fidelity. In biomedical scenarios, such faults have serious impacts, such as mischaracterizing symptom progression within longitudinal electronic health records (EHRs) or creating information in radiology reports. Although Retrieval-augmented generation (RAG) and chain-of-thought prompters hold out hope for hallucinations reduction based on anchoring outputs on evidence from the external world, but biomedical application is still compromised due to retrieval faults along with incompleteness of time reasoning.

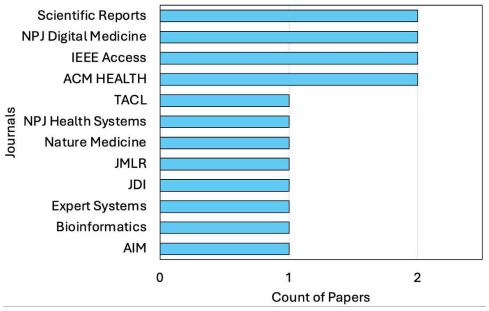
Inadequate Evaluation Frameworks: The evaluation of biomedical document summarization continues inadequate, with the majority of work still employing general-domain metrics such as ROUGE, BLEU, or METEOR, which evaluate lexical overlap rather than clinical fidelity or usefulness. While domain-specific metrics such as Concept F_1, PICO-based measures have been proposed, adoption remains sporadic, with no shared criterion for identifying clinical fidelity for a task at large.



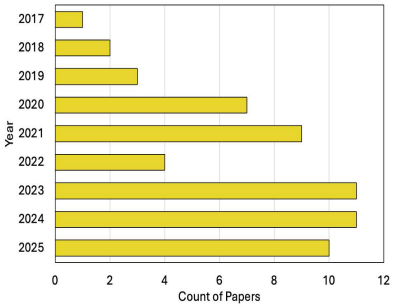
(a) Publication venue distribution



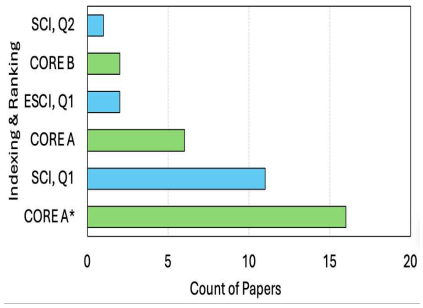
(b) Conference-wise distribution



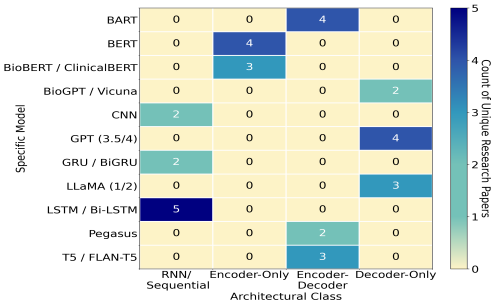
(c) Journal-wise distribution



(d) Year-wise publication count



(e) Indexing and ranking of publication venues



(f) Architectural classes vs. model usage heatmap

Fig. 8: Consolidated research trends in medical summarization. In (f), color intensity scale is 0–5.

Integration of Structured Medical Knowledge: State-of-the-art biomedical summarizers have mostly tackled unstructured text, without much integration of structured medical knowledge such as ontologies, laboratory findings, or treatment plans. This lack reduces model effectiveness for reasoning across deep relationships within the clinic, storing longitudinal patient information, and compliance with adopted medical plans.

Privacy, Ethics, and Legal Barriers: Biomedical data contain a high degree of sensitivity, privacy, ethical, and regulatory limitations that impose restrictions on data sharing and model building. Datasets are highly localized or proprietary, creating a challenge for training and benchmarking in diversely populated groups. Ethical risk is at issue when propagating biases, disclosing sensitive patient information, or creating inaccurate summaries. Geographically varying legal frameworks further impose limitations on collaborative study, as well as on AI deployment within medical contexts, presenting a challenge. In future, privacy-protected approaches, federated learning, and cautious governance must be adopted for a safe and equitable model deployment.

Efficiency and Deployment Constraints: State-of-the-art biomedical summarization models, particularly large LLMs, incur high computational costs and memory usage, limiting their usage in real-time medical workflow practice. Their high inference latency, energy, and infrastructure requirements constitute serious impediments to integration into hospital infrastructure or resource-poor locations. Lightweight, resource-efficient models and optimization methods, including knowledge distillation, pruning, and on-device inference, have not been extensively explored in this domain.

7 Conclusion

The automatic summarization of biomedical reports has become an important method for tackling the exponentially growing volume of medical information, enabling clinicians, researchers, and decision-makers to rapidly access essential information, reduce cognitive burden, and improve patient care. This paper provides a scholarly account of work undertaken between 2017 and 2025, examining medical report types, datasets, summarization strategies, model architectures, and evaluation schemes. We have highlighted the trend from the early sequential, extractive, and early transformer-based models toward generative ones and the subsequent use of large language models, which have elevated fluency, coherence, and clinical relevance for summaries significantly. Despite such progress, some challenges persist. Limited large-scale, domain-specific datasets, hallucinations, and factual inaccuracies, inadequate evaluation protocols, and lack of integration with structured medical knowledge remain discouraging factors for the generalizability and reliability of summarization systems. Ethically, privacy, and legal considerations, deployment, and computational limitations, make deployment of such technologies in realistic scenarios more difficult. Closing such gaps continues to require multidisciplinary solutions, such as the development of diverse, multilingual datasets, advanced fact-checking protocols, privacy-friendly strategies, and clinically mindful, efficient models.

Acknowledgment

The authors would like to acknowledge SaralPixel-AI Labs for its support, resources, and facilities that contributed to the completion of this research work.

References

1. Jain, R., Jangra, A., Saha, S., Jatowt, A.: A survey on medical document summarization. J. ACM **37**(111), 29 (2022) https://doi.org/10.1145/1122445.1122456
2. Vinod, P., Safar, S., Mathew, D., Venugopal, P., Joly, L.M., George, J.: Fine-tuning the bertsumext model for clinical report summarization. In: 2020 IEEE International Conference for Emerging Technology (INCET), pp. 1–7 (2020). https://doi.org/10.1109/INCET49848.2020.9154087
3. Pivovarov, R., Elhadad, N.: Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association **22**(5), 938–947 (2015) https://doi.org/10.1093/jamia/ocv032
4. Afantenos, S., Karkaletsis, V., Stamatopoulos, P.: Summarization from medical documents: a survey. Artificial Intelligence in Medicine **33**(2), 157–177 (2005) https://doi.org/10.1016/j.artmed.2004.07.017
5. Keszthelyi, D., Gaudet-Blavignac, C., Bjelogrlic, M., Lovis, C.: Patient information summarization in clinical settings: Scoping review. JMIR Med Inform 2023;11:e44639 https://medinform.jmir.org/2023/1/e44639 **11**(1), 44639 (2023) https://doi.org/10.2196/44639
6. Luo, Z., Ji, Y., Gupta, A., Li, Z., Frisch, A., He, D.: Towards accurate and clinically meaningful summarization of electronic health record notes: A guided approach. BHI 2023 - IEEE-EMBS International Conference on Biomedical and Health Informatics, Proceedings (2023) https://doi.org/10.1109/BHI58575.2023.10313411
7. Zhu, Y., Yang, X., Wu, Y., Zhang, W.: Leveraging summary guidance on medical report summarization. IEEE Journal of Biomedical and Health Informatics **27**(10), 5066–5075 (2023) https://doi.org/10.1109/JBHI.2023.3304376
8. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C.P., Hom, J., Gatidis, S., Pauly, J., Chaudhari, A.S.: Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine 2024 30:4 **30**(4), 1134–1142 (2024) https://doi.org/10.1038/s41591-024-02855-5
9. Sarker, A., Yang, Y.C., Al-Garadi, M.A., Abbas, A.: A light-weight text summarization system for fast access to medical evidence. Frontiers in Digital Health **2**, 585559 (2020) https://doi.org/10.3389/fdgth.2020.585559
10. Lee, C., Vogt, K.A., Kumar, S.: Prospects for ai clinical summarization to reduce the burden of patient chart review. Frontiers in Digital Health **6**, 1475092 (2024) https://doi.org/10.3389/fdgth.2024.1475092
11. Tang, X., Tran, A., Tan, J., Gerstein, M.: Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 546–554 (2023) https://doi.org/10.18653/v1/2023.clinicalnlp-1.58
12. Williams, C.Y.K., Subramanian, C.R., Ali, S.S., Apolinario, M., Askin, E., Barish, P., Cheng, M., Deardorff, W.J., Donthi, N., Ganeshan, S., Huang, O., Kantor, M.A., Lai, A.R., Manchanda, A., Moore, K.A., Muniyappa, A.N., Nair, G., Patel, P.P., Santhosh, L., Schneider, S., Torres, S., Yukawa, M., Hubbard, C.C., Rosner, B.I.: Physician- and large language model–generated hospital discharge summaries. JAMA Internal Medicine **185**(7), 818–825 (2025) https://doi.org/10.1001/jamainternmed.2025.0821
13. Bednarczyk, L., Reichenpfader, D., Gaudet-Blavignac, C., Ette, A.K., Zaghir, J., Zheng, Y., Bensahla, A., Bjelogrlic, M., Lovis, C.: Scientific evidence for clinical text summarization using large language models: Scoping review. Journal of Medical Internet Research **27**(1), 68998 (2025) https://doi.org/10.2196/68998
14. Scarfield, P., Shepherd, T.D., Stapleton, C., Starks, A., Benn, E., Khalid, S., Dayment, B., Moate, A., Mohamed, S., Lee, J.: Improving the quality and content of discharge summaries on acute medicine wards: a quality improvement project. BMJ Open Quality **11**(2), 1780 (2022) https://doi.org/10.1136/bmjoq-2021-001780
15. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Data descriptor: Mimic-iii, a freely accessible critical care database. Scientific Data **3**(1), 160035 (2016) https://doi.org/10.1038/sdata.2016.35
16. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Moody, B., Gow, B., Lehman, L.w.H., Celi, L.A., Mark, R.G.: Mimic-iv, a freely accessible electronic health record dataset. Scientific Data 2023 10:1 **10**(1), 1 (2023) https://doi.org/10.1038/s41597-022-01899-x
17. Nagar, A., Liu, Y., Liu, A.T., Schlegel, V., Dwivedi, V.P., Kaliya-Perumal, A.-K., Kalanchiam, G.P., Tang, Y., Tan, R.T.: umedsum: a unified framework for clinical abstractive summarization. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2653–2672 (2025)
18. Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., Wong, A.: Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 1744–1753 (2021) https://doi.org/10.18653/v1/2021.naacl-main.139
19. Sharma, A., Feldman, D.I., Jain, A.: Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 228–235 (2023) https://doi.org/10.18653/v1/2023.clinicalnlp-1.28
20. Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., Bendayan, R., Richardson, M.P., Stewart, R., Shah, A.D., Wong, W.K., Ibrahim, Z., Teo, J.T., Dobson, R.J.B.: Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. Artificial Intelligence in Medicine **117**, 102083 (2021) https://doi.org/10.1016/j.artmed.2021.102083
21. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data 2019 6:1 **6**(1), 317 (2019) https://doi.org/10.1038/s41597-019-0322-0
22. Delbrouck, J.B., Zhang, H., Rubin, D.L.: Qiai at mediqa

2021: Multimodal radiology report summarization. Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP 2021, 285–290 (2021) https://doi.org/10.18653/v1/2021.bionlp-1.33
23. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
24. Sultan, T., Rony, M.A.T., Islam, M.S., Alshathri, S., El-Shafai, W.: Sumgpt: A multimodal framework for radiology report summarization to improve clinical performance. IEEE Access **13**, 15929–15945 (2025) https://doi.org/10.1109/ACCESS.2025.3528335
25. Abacha, A.B., M'rabet, Y., Zhang, Y., Shivade, C., Langlotz, C., Demner-Fushman, D.: Overview of the mediqa 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 74–85 (2021)
26. Michalopoulos, G., Williams, K., Singh, G., Lin, T.: Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4741–4749. Association for Computational Linguistics, Stroudsburg, PA, USA (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.349
27. Tariq, A., Trivedi, S., Urooj, A., Ramasamy, G., Fathizadeh, S., Stib, M., Tan, N., Patel, B., Banerjee, I.: Patient-centric summarization of radiology findings using two-step training of large language models. ACM Transactions on Computing for Healthcare **6**(2) (2025) https://doi.org/10.1145/3709154
28. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 590–597 (2019) https://doi.org/10.1609/aaai.v33i01.3301590
29. Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D., Langlotz, C.P.: Learning to summarize radiology findings. EMNLP 2018 - 9th International Workshop on Health Text Mining and Information Analysis, LOUHI 2018 - Proceedings of the Workshop, 204–213 (2018) https://doi.org/10.18653/v1/w18-5623
30. Neupane, S., Mitra, S., Mittal, S., Gaur, M., Golilarz, N.A., Rahimi, S., Amirlatifi, A.: Medinsight: A multi-source context augmentation framework for generating patient-centric medical responses using large language models. ACM Transactions on Computing for Healthcare **6**(2), 1–19 (2025) https://doi.org/10.1145/3709365
31. Mowery, D.L., Velupillai, S., South, B.R., Christensen, L., Martinez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.: Task 2: Share/-clef ehealth evaluation lab 2014. Focus on Microscopy (1), 255 (2014) https://doi.org/10.34894/VQ1DJA
32. Gao, Y., Dligach, D., Miller, T., Churpek, M.M., Afshar, M.: Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 461–467 (2023) https://doi.org/10.18653/v1/2023.bionlp-1.43
33. Papadopoulos, A., Babylon, K., Babylon, F.M., Moramarco, F., Savkov, A., Savkov, B.S., Uk, b.C.: Primock57: A dataset of primary care mock consultations. Proceedings of the Annual Meeting of the Association for Computational Linguistics **2**, 588–598 (2022) https://doi.org/10.18653/v1/2022.acl-short.65
34. Binici, K., Ramesh Kashyap, A., Schlegel, V., Liu, A.T., Dwivedi, V.P., Nguyen, T.-T., Gao, X., Chen, N.F., Winkler, S.: Medsage: Enhancing robustness of medical dialogue summarization to asr errors with llm-generated synthetic dialogues. Proceedings of the AAAI Conference on Artificial Intelligence **39**(22), 23496–23504 (2025) https://doi.org/10.1609/aaai.v39i22.34518
35. Guo, J., Wang, M., Qiao, X., Wei, D., Shang, H., Li, Z., Yu, Z., Li, Y., Su, C., Zhang, M., Tao, S., Yang, H.: Ucorrect: An unsupervised framework for automatic speech recognition error correction, 1–5 (2024) https://doi.org/10.1109/ICASSP49357.2023.10096194
36. Yim, W.w., Fu, Y., Ben Abacha, A., Snider, N., Lin, T., Yetisgen, M.: Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Scientific Data 2023 10:1 **10**(1), 586 (2023) https://doi.org/10.1038/s41597-023-02487-3
37. Ben Abacha, A., Fan, Y., Yim, W.W., Lin, T.: An empirical study of clinical note generation from doctor-patient encounters. EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 2291–2302 (2023) https://doi.org/10.18653/v1/2023.eacl-main.168
38. Abacha, A.B., Yim, W.W., Adams, G., Snider, N., Yetisgen, M.: Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 503–513 (2023) https://doi.org/10.18653/v1/2023.clinicalnlp-1.52
39. Mathur, Y., Rangreji, S., Kapoor, R., Palavalli, M., Bertsch, A., Gormley, M.R.: Summqa at mediqa-chat 2023: In-context learning with gpt-4 for medical summarization. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 490–502 (2023) https://doi.org/10.18653/v1/2023.clinicalnlp-1.51
40. Neupane, S., Tripathi, H., Mitra, S., Bozorgzad, S., Mittal, S., Rahimi, S., Amirlatifi, A.: Clinicsum: Utilizing language models for generating clinical summaries from patient-doctor conversations. In: 2024 IEEE International Conference on Big Data (BigData), pp. 5050–5059 (2024). https://doi.org/10.1109/BigData62323.2024.10825266
41. Giorgi, J., Toma, A., Xie, R., Chen, S., An, K., Zheng, G., Wang, B.: Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In: Proceedings of the 5th Clinical Natural Language Processing Workshop, pp. 323–334. Association for Computational Linguistics, Stroudsburg, PA, USA (2023). https://doi.org/10.18653/v1/2023.clinicalnlp-1.36
42. Tiwari, A., Saha, A., Saha, S., Bhattacharyya, P., Dhar, M.: Experience and evidence are the eyes of an excellent summarizer! towards knowledge infused multi-modal clinical conversation summarization. International Conference on Information and Knowledge Management, Proceedings, 2452–2461 (2023) https://doi.org/10.1145/3583780.3614870
43. Ramprasad, S., Ferracane, E., Selvaraj, S.P.: Generating

more faithful and consistent soap notes using attribute-specific parameters. Proceedings of Machine Learning Research **219**, 631–649 (2023)
44. Schloss, B., Konam, S.: Towards an automated soap note: Classifying utterances from medical conversations. Proceedings of Machine Learning Research, 610–631 (2020)
45. Nguyen, V.-T., Pham, H.-D., To, T.-H., Do, C.-T.H., Dong, T.-T.-T., Le, V.-T.D., Hoang, V.-P.: Medalyze: Lightweight medical report summarization application using flan-t5-large (2025)
46. Enarvi, S., Amoia, M., Del-Agua Teba, M., Delaney, B., Diehl, F., Gallopyn, G., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., Rubini, L., Ruiz, M., Singh, G., Stemmer, F., Sun, W., Vozila, P., Lin, T., Ramamurthy, R.: Generating medical reports from patient-doctor conversations using sequence-to-sequence models. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 22–30 (2020) https://doi.org/10.18653/v1/2020.nlpmc-1.4
47. Quiroz, J.C., Laranjo, L., Kocaballi, A.B., Briatore, A., Berkovsky, S., Rezazadegan, D., Coiera, E.: Identifying relevant information in medical conversations to summarize a clinician-patient encounter. Health Informatics Journal **26**(4), 2906–2914 (2020) https://doi.org/10.1177/1460458220951719
48. Fraile Navarro, D., Coiera, E., Hambly, T.W., Triplett, Z., Asif, N., Susanto, A., Chowdhury, A., Azcoaga Lorenzo, A., Dras, M., Berkovsky, S.: Expert evaluation of large language models for clinical dialogue summarization. Scientific Reports 2025 15:1 **15**(1), 1195 (2025) https://doi.org/10.1038/s41598-024-84850-x
49. Joshi, A., Katariya, N., Amatriain, X., Kannan, A.: Dr. summarize: Global summarization of medical dialogue by exploiting local structures. Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, 3755–3763 (2020) https://doi.org/10.18653/v1/2020.findings-emnlp.335
50. Shasha, G.Z., Balint-Lahat, N., Schiby, G., Debby, A., Barshack, I., Mayer, C., Shasha, G.Z., Balint-Lahat, N., Schiby, G., Debby, A., Barshack, I., Mayer, C.: Pathology made simple: Chatgpt's summarization of pathology reports. Archives of Pathology and Laboratory Medicine **150**(3), 250–252 (2026) https://doi.org/10.5858/arpa.2025-0072-oa
51. Ghosh, A., Acharya, A., Saha, S., Pandey, G., Raghu, D., Sinha, S.: Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024, 11546–11560 (2024) https://doi.org/10.18653/v1/2024.findings-emnlp.675
52. Jeblee, S., Khattak, F.K., Crampton, N., Mamdani, M., Rudzicz, F.: Extracting relevant information from physician-patient dialogues for automated clinical note taking. LOUHI@EMNLP 2019 - 10th International Workshop on Health Text Mining and Information Analysis, Proceedings, 65–74 (2019) https://doi.org/10.18653/v1/d19-6209
53. Molenaar, S., Maas, L., Burriel, V., Dalpiaz, F., Brinkkemper, S.: Medical dialogue summarization for automated reporting in healthcare. Lecture Notes in Business Information Processing **382 LNBIP**, 76–88 (2020) https://doi.org/10.1007/978-3-030-49165-9_7
54. Ghosh, A., Acharya, A., Jain, R., Saha, S., Chadha, A., Sinha, S.: Clipsyntel: Clip and llm synergy for multimodal question summarization in healthcare. Proceedings of the AAAI Conference on Artificial Intelligence **38**(20), 22031–22039 (2024) https://doi.org/10.1609/aaai.v38i20.30206
55. Ghosh, A., Tomar, M.S., Tiwari, A., Saha, S., Salve, J.A., Sinha, S.: From sights to insights: Towards summarization of multimodal clinical documents. Proceedings of the Annual Meeting of the Association for Computational Linguistics **1**, 13117–13129 (2024) https://doi.org/10.18653/v1/2024.acl-long.708
56. Ghosh, A., Acharya, A., Jha, P., Saha, S., Gaudgaul, A., Majumdar, R., Chadha, A., Jain, R., Sinha, S., Agarwal, S.: Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **14612 LNCS**, 106–120 (2024) https://doi.org/10.1007/978-3-031-56069-9_8
57. Abacha, A.B., Demner-Fushman, D.: On the summarization of consumer health questions. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2228–2234 (2019) https://doi.org/10.18653/v1/p19-1215
58. Mohan, S., Li, D.: Medmentions: A large biomedical corpus annotated with umls concepts (2019)
59. Dalal, A., Ranjan, S., Bopaiah, Y., Chembachere, D., Steiger, N., Burns, C., Daswani, V.: Text summarization for pharmaceutical sciences using hierarchical clustering with a weighted evaluation methodology. Scientific Reports 2024 14:1 **14**(1), 20149 (2024) https://doi.org/10.1038/s41598-024-70618-w
60. Dou, Z.Y., Liu, P., Hayashi, H., Jiang, Z., Neubig, G.: Gsum: A general framework for guided neural abstractive summarization. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 4830–4842 (2021) https://doi.org/10.18653/v1/2021.naacl-main.384
61. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer (2020)
62. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
63. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. 8th International Conference on Learning Representations, ICLR 2020 (2020)
64. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2025)
65. Han, T., Adams, L.C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Figueroa, A., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)
66. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission (2020)
67. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language

representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020) https://doi.org/10.1093/bioinformatics/btz682
68. Podder, V., Lew, V., Ghassemzadeh, S.: Soap notes. Ploughshares **48**(4), 145–148 (2023) https://doi.org/10.1353/plo.2022.0155
69. Rajaganapathy, S., Chowdhury, S., Li, X., Buchner, V., He, Z., Zhang, R., Jiang, X., Yang, P., Cerhan, J.R., Zong, N.: Synoptic reporting by summarizing cancer pathology reports using large language models. npj Health Systems 2025 2:1 **2**(1), 11 (2025) https://doi.org/10.1038/s44401-025-00013-8
70. Ahmed, F., Yang, L., Jaroensri, T., Sellergren, A., Matias, Y., Hassidim, A., Corrado, G.S., Webster, D.R., Shetty, S., Prabhakara, S., Liu, Y., Golden, D., Wulczyn, E., Steiner, D.F.: Polypath: Adapting a large multimodal model for multislide pathology report generation. Modern Pathology **38**(11), 100886 (2025) https://doi.org/10.1016/j.modpat.2025.100886
71. Al-Hussaini, I., Wu, A.X., Mitchell, C.S.: Pathology dynamics at biolaysumm: the trade-off between readability, relevance, and factuality in lay summarization. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 592–601 (2023) https://doi.org/10.18653/v1/2023.bionlp-1.63
72. Zhang, G., Zhou, Y., Hu, Y., Xu, H., Weng, C., Peng, Y.: A span-based model for extracting overlapping pico entities from randomized controlled trial publications. Journal of the American Medical Informatics Association **31**(5), 1163–1171 (2024) https://doi.org/10.1093/jamia/ocae065
73. Chintagunta, B., Katariya, N., Amatriain, X., Kannan, A.: Medically aware gpt-3 as a data generator for medical dialogue summarization. Proceedings of Machine Learning Research **126**, 354–372 (2021)
74. Zhang, G., Jin, Q., Zhou, Y., Wang, S., Idnay, B., Luo, Y., Park, E., Nestor, J.G., Spotnitz, M.E., Soroush, A., Campion, T.R., Lu, Z., Weng, C., Peng, Y.: Closing the gap between open source and commercial large language models for medical evidence summarization. npj Digital Medicine 2024 7:1 **7**(1), 239 (2024) https://doi.org/10.1038/s41746-024-01239-w
75. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. International Journal for Research in Applied Science and Engineering Technology **10**(12), 943–947 (2015) https://doi.org/10.22214/ijraset.2022.47789
76. Bedi, P.P.S., Bala, M., Sharma, K.: Extractive text summarization for biomedical transcripts using deep dense lstm-cnn framework. Expert Systems **41**(7), 13490 (2024) https://doi.org/10.1111/exsy.13490
77. Alhajeri, M.S., Ren, Y.M., Ou, F., Abdullah, F., Christofides, P.D.: Model predictive control of nonlinear processes using transfer learning-based recurrent neural networks. Chemical Engineering Research and Design **205**, 1–12 (2024) https://doi.org/10.1016/j.cherd.2024.03.019
78. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735
79. Kiperwasser, E., Goldberg, Y.: Simple and accurate dependency parsing using bidirectional lstm feature representations. Transactions of the Association for Computational Linguistics **4**, 313–327 (2016) https://doi.org/10.1162/tacl_a_00101
80. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1724–1734 (2014) https://doi.org/10.3115/v1/d14-1179
81. Li, C., Xu, W., Li, S., Gao, S.: Guiding generation for abstractive text summarization based on key information guide network. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **2**, 55–60 (2018) https://doi.org/10.18653/v1/n18-2009
82. Krishna, K., Khosla, S., Bigham, J., Lipton, Z.C.: Generating soap notes from doctor-patient conversations using modular summarization techniques. ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference **1**, 4958–4972 (2021) https://doi.org/10.18653/v1/2021.acl-long.384
83. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) **1**, 1073–1083 (2017) https://doi.org/10.18653/v1/P17-1099
84. Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
85. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 11328–11339 (2020)
86. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Research, A.G.: Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems **33**, 17283–17297 (2020)
87. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020) https://doi.org/10.1145/3422622
88. Li, M., Zhang, L., Ji, H., Radke, R.J.: Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2190–2196 (2019) https://doi.org/10.18653/v1/p19-1210
89. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020)
90. Mienye, I.D., Swart, T.G., Obaido, G.: Recurrent neural networks: A comprehensive review of architectures, variants, and applications. Information 2024, Vol. 15, Page 517 **15**(9), 517 (2024) https://doi.org/10.3390/info15090517
91. Mienye, I.D., Sun, Y.: A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access **11**, 30628–30638 (2023) https://doi.org/10.1109/ACCESS.2023.3262020
92. Zhang, W., Li, H., Tang, L., Gu, X., Wang, L., Wang, L.: Displacement prediction of jiuxianping landslide using gated recurrent unit (gru) networks. Acta Geotechnica 2022 17:4 **17**(4), 1367–1382 (2022) https://doi.org/10.1007/s11440-022-01495-8

93. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 7871–7880 (2020) https://doi.org/10.18653/v1/2020.acl-main.703
94. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
95. Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N.: A discourse-aware attention model for abstractive summarization of long documents. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **2**, 615–621 (2018) https://doi.org/10.18653/v1/n18-2097
96. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
97. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)
98. Grambow, C.A., Zhang, L., Schaaf, T.: In-domain pre-training improves clinical note generation from doctor-patient conversations, 9–22 (2022)
99. Zhang, L., Negrinho, R., Ghosh, A., Jagannathan, V., Hassanzadeh, H.R., Schaaf, T., Gormley, M.R.: Leveraging pretrained models for automatic summarization of doctor-patient conversations. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, 3693–3712 (2021) https://doi.org/10.18653/v1/2021.findings-emnlp.313
100. Narayan, S., Zhao, Y., Maynez, J., Simões, G., Nikolaev, V., McDonald, R.: Planning with learned entity prompts for abstractive summarization. Transactions of the Association for Computational Linguistics **9**, 1475–1492 (2021) https://doi.org/10.1162/tacl_a_00438
101. He, J., Kryściński, W., McCann, B., Rajani, N., Xiong, C.: Ctrlsum: Towards generic controllable text summarization. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 5879–5915 (2022) https://doi.org/10.18653/v1/2022.emnlp-main.396
102. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
103. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), 409 (2022)
104. Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries, 74–81 (2004)
105. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 311–318 (2002) https://doi.org/10.3115/1073083.1073135
106. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, 65–72 (2005)
107. Laban, P., Schnabel, T., Bennett, P.N., Hearst, M.A.: Summac: Re-visiting nli-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics **10**, 163–177 (2022)
108. Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., Gallinari, P.: Questeval: Summarization asks for fact-based evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6594–6604 (2021)
109. Es, S., James, J., Anke, L.E., Schockaert, S.: Ragas: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 150–158 (2024)
110. Allen, I.E., Seaman, C.A.: Likert Scales and Data Analyses. www.sloan-c.org,
111. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia Medica **22**(3), 276–282 (2012)
112. Stetson, P.D., Bakken, S., Wrenn, J.O., Siegler, E.L.: Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). Applied Clinical Informatics **3**(2), 164–174 (2012) https://doi.org/10.4338/ACI-2011-11-RA-0070
113. Schaye, V., Miller, L., Kudlowitz, D., Chun, J., Burk-Rafel, J., Cocks, P., Guzman, B., Aphinyanaphongs, Y., Marin, M.: Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. Journal of General Internal Medicine 2021 37:3 **37**(3), 507–512 (2021) https://doi.org/10.1007/s11606-021-06805-6
114. Lin, Y.T., Chen, Y.N.: Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 47–58 (2023) https://doi.org/10.18653/v1/2023.nlp4convai-1.5
115. Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., Mosconi, F.: Biomedbert: A pre-trained biomedical language model for qa and ir. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 669–679 (2020)
116. Knoll, T., Moramarco, F., Korfiatis, A.P., Young, R., Ruffini, C., Perera, M., Perstl, C., Reiter, E., Belz, A., Savkov, A.: User-driven research of medical note generation software. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 385–394 (2022)