HSPI The Open Access Publisher

J C C M JOURNAL OF CARDIOLOGY AND CARDIOVASCULAR MEDICINE
ISSN 2575-0143

Review Article

Transformer-based Clinical Report Generation from Stroke Electronic Health Records (EHRs): A Comprehensive Review

Rabia Amin[1]* and Ali Raza Aslam[2]

[1]University of Engineering and Technology, Lahore, Pakistan
[2]University of Superior, Lahore, Pakistan

More Information

***Corresponding author:** Rabia Amin, University of Engineering and Technology Lahore, Pakistan, Email: rabiaamin947@gmail.com

Submitted: October 16, 2026
Accepted: October 30, 2026
Published: October 31, 2026

Citation: Amin R, Aslam AR. Transformer-based Clinical Report Generation from Stroke Electronic Health Records (EHRs): A Comprehensive Review. J Cardiol Cardiovasc Med. 2026; 11(5): 71-81. Available from: https://dx.doi.org/10.29328/journal.jccm.1001230

Copyright license: © 2026 Amin R, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Check for updates

OPEN ACCESS

Abstract

Structured Electronic Health Records (EHRs) provide a valuable source for generating clinical narrative reports that are increasingly automated in this rapidly evolving field of medical AI, especially in stroke care where timely documentation is vital. Complex multimodal data comes with stroke patients in the form of lab results, neurological scores (e.g., NIHSS), imaging results, medication histories, and structured diagnostic codes. Converting structured information to coherent clinical narratives like radiology reports or discharge summaries is still a difficult natural language generation (NLG) problem. In recent years, various Transformer-based architectures have emerged, such as encoder–decoder models like T5 and BART, and decoder-only large language models (LLMs) like GPT, LLaMA, and Phi, which have greatly enhanced the generation of medical texts. However, the generation of a stroke-specific EHR-to-report is not widely discussed, as datasets are limited, there is a lack of standardized benchmarks, and there is little clinical validation. This is a systematic review of literature from 2015–2026 related to Transformer-based clinical report generation for stroke patients. Datasets, architectures, evaluation metrics, baseline models, and clinical applications are discussed. We also list some limitations such as risk of hallucination, poor generalization and lack of paired datasets for strokes. Lastly, we suggest future directions such as multimodal learning, retrieval-augmented generation, and clinically-based evaluation frameworks.

Introduction

Stroke is a major cause of death and chronic disability in the world, and it is essential to diagnose and correctly document the stroke at every level of care. Structured Electronic Health Record (EHR) data, such as vital signs, laboratory tests, medications, imaging interpretations, neurological assessment (e.g., National Institutes of Health Stroke Scale (NIHSS)), and others, are generated at high volume in clinical workflows. Clinicians have to manually process this structured information and make it available as narrative reports such as discharge documentations and radiology interpretations, but this is time-consuming and prone to errors. With the advent of Transformer architectures [1], the field of natural language processing got revolutionized with the ability to model in parallel, in an attention-based way. These capabilities were extended to medical domains in subsequent models like BERT, GPT, T5, and BART, thus enhancing clinical summarization and text generation tasks. This development is still insufficient for the generation of structured text in stroke care. The current systems are mainly based on imaging rather than structured data transformation in EHRs. This leaves a big gap in clinical documentation automation systems.

Figure 1 shows an architecture of multimodal Transformer for automatic report generation, with medical imaging features and clinical symptom information combined. The architecture consists of visual feature extraction, Transformer encoding and an attention-driven decoder to produce clinical narratives. Clinical symptom information is also added in the process of report generation, providing additional context to the model, and enabling it to account for imaging and clinical data. While this architecture provides a promising example of the capabilities of multimodal clinical reporting, more research is needed to explore the applicability of the architecture for comprehensive stroke EHR-to-report generation, especially in terms of longitudinal laboratory results, neurological evaluations, treatment data, and numerical consistency.

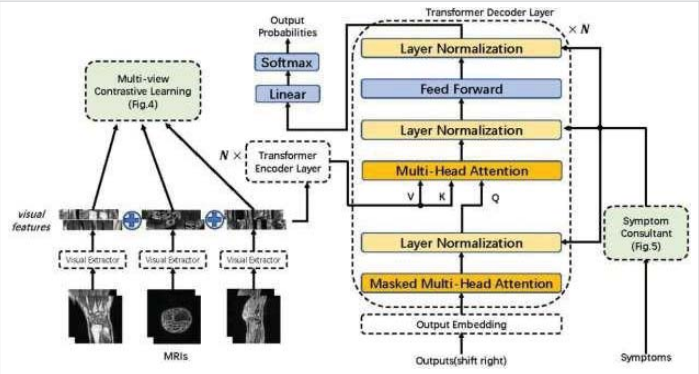


Figure 1: Multimodal Transformer architecture integrating imaging features and clinical symptoms for automated report generation.

The left side of the architecture shows the Transformer decoder layer that is used to generate the clinical narrative. The output sequence is shifted to the decoder input in the right direction and embedded in a masked multi-head attention layer to forbid attending to future tokens when training. Importantly, the architecture uses a standard cross-attention mechanism in which keys (K) and values (V) are responsibility of the Transformer output (encoder), and the query (Q) is responsible for the decoder. This permits the decoder to concentrate on the pertinent visual characteristics as it creates each word of the report.

One of the distinctive features of this architecture is the module called a "Symptom Consultant," which is listed in Fig. 5 in the source. This module uses the clinical symptoms as input for this module which directly passes the symptoms into the decoder's multi-head attention mechanism. Its integration ensures that the report is generated not only based on the imaging findings, but mainly on the patient's clinical presentation, which is very important for stroke care given that the symptoms that bring the patient to the hospital, such as sudden weakness, aphasia or altered consciousness affect the radiological interpretation. The final decoder output goes through feed-forward networks (layer normalization), another feed-forward network (layer normalization), a linear projection, and a softmax layer to get to the final output probabilities for the generated clinical report. This architecture is a strong base for multi-modal clinical documentation that will support the future development of multi-modal learning (EHR + MRI + CT).

Figure 2 shows a common clinical practice process that is focused on the Electronic Health Record (EHR) system. Patient registration is the first step, whereby the patient is registered in the front office modules for scheduling and patient chart management. The patient portal is connected with the front office for scheduling and with the physician for taking the patient's history. The doctor is then able to create clinical notes and retrieve medical history, the essential clinical documentation in the EHR. From this will flow out to external partners, such as prescribing physicians delivering prescription orders to the pharmacist; the laboratory

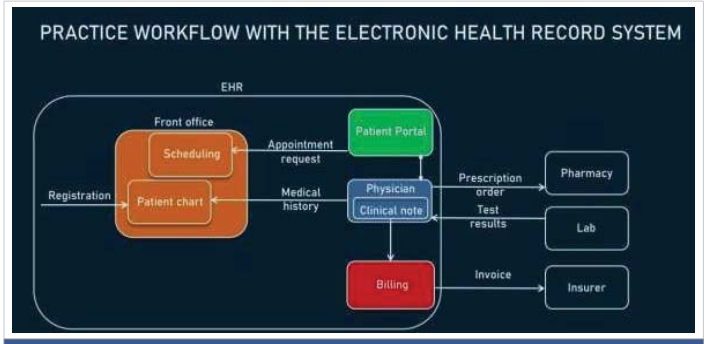


Figure 2: Practice workflow with the Electronic Health Record (EHR) system.

receiving test results; and the insurance company receiving billing data. The billing component of the EHR also handles billing details of the encounter. This is particularly relevant to stroke EHR-to-report generation, as it represents the actual sources feeding into the unstructured and structured data that are consumed by Transformer models. In the case of stroke care, the inputs are:

IIn stroke care, the inputs X for a report generation model $Y=f\theta(X)$ include demographics, vital signs, laboratory results, ICD diagnostic codes

Figure 2 shows a typical Electronic Health Record (EHR) workflow, including patient registration, clinical record data entry, laboratory tests, medication management, and communication between healthcare providers. The interrelated processes generate structured and unstructured clinical data that can be used to generate clinical reports automatically. For stroke care, data would consist of patient information, neurologic data, laboratory data, medication history, imaging data, and time data. Data linkage, temporal alignment, privacy protection and clinical verification are needed to integrate these data sources into a Transformer-based reporting system.

Figure 3 shows a conceptual model for generating reports from the stroke EHRs using a Transformer. Structured clinical data, such as demographics, Lab data, vital signs, NIHSS scores, medications, treatment time, etc., is gathered

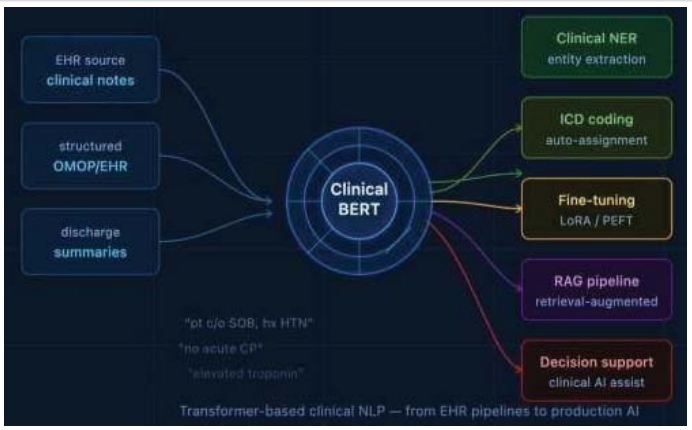


Figure 3: Proposed Transformer-based framework for generating and verifying stroke clinical reports from structured EHR data.

and processed using data cleaning, data normalization, and temporal alignment. This processed information is then fed into a generation model based on a Transformer to generate a draft clinical report. A next verification phase verifies the consistency between clinical entities, numerical measurements, and treatment events in the source EHR. Lastly, a clinician review is used prior to approval for clinical use. This framework is suggested as a possible research pathway and not as a clinically validated pathway.

Literature review (Tables 1 & 2)

Table 1. Critical synthesis of foundational models, clinical language resources, and stroke-related studies relevant to Transformer-based EHR-to-report generation.

Table 2. EHR data resources cited in the review and their implications for constructing a stroke report-generation benchmark.

While some of the reviewed literature offers complementary support for stroke EHR-to-report generation, evidence is significant in varying strengths and degrees of linkage. Architectures proposed for conditional text generation models, such as T5, BART and Clinical-T5, are more general and can be implemented for other tasks, including representation learning and information extraction, which are primarily supported by models derived from BERT. While imaging studies of stroke illustrate automated extraction and reporting of stroke-related information, they have not been used to build a comprehensive narrative from longitudinal EHRs. The work by Rouhizadeh et al. (2026) does show high-quality ML report generation under controlled conditions; however, a multispecialty evaluation using synthetic EHR inputs doesn't support that the reports are reliable under regular conditions for stroke. All of these serve as a justification to propose a stroke-specific benchmark for rating improvements in the time-limited clinical inputs, and for assessing factual correctness, clinically relevant omissions, numerical consistency, and institution-specific generalizations, in addition to linguistic quality, in clinician-authored reports.

The evolution of clinical NLP stroke reports (2015-2026)

Rule based and early machine learning era (2015-2018)

The initial clinical NLP solutions were based on rule-based templates and statistical models, including Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). These systems were only for small-scale use and didn't work well if you tried to expand them to other institutions or patients.

Deep learning transition (2018–2020)

Neural networks, specifically LSTMs and GRUs are introduced, which allowed modeling the clinical text more effectively as a series of sentences. Medical terms were better represented semantically by Word embeddings like Word2Vec and GloVe. The representation of medical terms was improved at the semantic level using Word embeddings like Word2Vec and GloVe. These models were, however, less successful in dealing with long-range dependencies and factual consistency.

Transformer revolution (2019–2026)

Recurrent architectures were replaced by self-attention mechanisms for scalable and context-aware text generation in transformers. Key models include:

- BERT (Devlin et al., 2019) [2]
- GPT series (OpenAI)
- T5 (Raffel et al., 2020) [3]
- BART (Lewis et al., 2020) [4]
- ClinicalBERT (Alsentzer et al., 2019) [5]

These models did a remarkable job of summarization, prediction of diagnosis, and report generation in clinical NLP tasks.

Figure 4 shows a historical panoramic overview of the advancements of artificial intelligence and natural language processing (NLP) in medicine from 1956 to today, to show the technological background of today's clinical report generation using Transformers. In 1956 the notion of Artificial Intelligence as a research area was formalized. The early NLP programs from the 1960s and 1970s included ELIZA, an early NLP program, and rule-based knowledge-based systems such as MYCIN (1972) for the diagnosis and treatment of infectious diseases, and CASNET for the diagnosis and treatment of glaucoma. These were primitive systems using manually written symbolic rules, and were not easily scalable and adaptable.

In the 1980's and 1990's, there was a trend to making things more standard and introducing early neural networks. In 1985, the DICOM standard was introduced, which allowed for the first time a unified communication of medical images and became the groundwork for applications using data. Other systems followed, such as Decision support systems (DxPlain), and in 1998, the first Convolutional neural networks (LeNet-5) were designed, bringing deep learning to medical data analysis. The 'backward' innovation in the 2000s and 2010s also spanned new breakthroughs, such as IBM Watson for RNA binding proteins identification in 2000 and significantly, the launch of the Transformer model in 2017. This architectural change marked the dawn of fundamental changes in NLP, motivating parallel processing and context-aware text generation.

Table 1: Critical synthesis of foundational models, clinical language resources, and stroke-related studies relevant to Transformer-based EHR-to-report generation.

Study and evidence category	Research objective	Methodology / architecture	Dataset / evidence base	Principal findings or contribution	Limitations and relevance to stroke EHR-to-report generation
Devlin et al. (2019) —[2] Foundational language representation	Develop contextual language representations for diverse NLP tasks.	Bidirectional Transformer encoder pretrained using masked language modelling and next-sentence prediction, followed by task-specific fine-tuning.	General-domain BookCorpus and English Wikipedia; language-nderstanding benchmarks.	Established a reusable representation-learning framework with strong performance across language-understanding tasks.	BERT is an encoder rather than a standalone autoregressive report generator. Its relevance lies in clinical entity extraction, input encoding, and factual verification; this study provides no direct evidence of stroke report-generation performance.
Alsentzer et al. (2019) — [11] Clinical representation learning	Adapt contextual embeddings to clinical language and discharge summaries.	Further pretraining of BERT-based models on clinical text, followed by evaluation on clinical NLP tasks.	MIMIC-III clinical notes and discharge summaries; five downstream clinical NLP tasks.	Clinical adaptation improved performance on three of five tasks and established a new state of the art on MedNLI; improvements did not extend to two de-identification tasks.	Benefits are task-dependent. The models support clinical understanding rather than independently generating narrative reports. Stroke- specific generation, numerical fidelity, and temporal consistency require separate evaluation.
Lee et al. (2020) —[12] Biomedical representation learning	Improve extraction and understanding of biomedical information.	BioBERT: continued pretraining of a BERT encoder on biomedical literature, with fine-tuning for biomedical text-mining tasks.	PubMed abstracts, PMC full-text articles, and biomedical entity-recognition, relation-extraction, and question-answering benchmarks.	Biomedical pretraining improved performance over general BERT on several biomedical text-mining tasks.	Biomedical publications differ from routine EHR language, abbreviations, and documentation practices. BioBERT can support terminology extraction and verification, but its results do not establish clinical narrative-generation accuracy.
Raffel et al. (2020) — [3] Foundational generative architecture	Unify NLP tasks within a common text-to-text framework.	T5 encoder–decoder Transformer; systematic comparison of transfer-learning configurations and pretraining objectives.	Colossal Clean Crawled Corpus (C4) and multiple general-domain NLP benchmarks.	Introduced a unified framework in which inputs and outputs are represented as text, supporting transfer across diverse tasks.	Provides an architectural basis for EHR-to-report generation, but general-domain pretraining does not establish clinical reliability. Stroke applications require explicit input serialization, paired clinical supervision, and evaluation of omissions and unsupported statements.
Lewis et al. (2020) —[4] Foundational generation and summarization	Improve sequence generation through denoising pretraining.	BART combines a bidirectional encoder with an autoregressive decoder and learns to reconstruct corrupted text, including masked spans and shuffled sentences.	General-domain text and generation, summarization, translation, and comprehension benchmarks.	Demonstrated strong generation performance, including state-of-the-art results on several abstractive tasks.	General summarization performance cannot be interpreted as stroke-specific clinical accuracy. Application to structured EHRs requires clinical fine-tuning and verification of diagnoses, medications, values, and event chronology.
Lehman and Johnson (2023) —[17] Clinical generative model resource	Provide T5-based models pretrained on clinical language.	Clinical-T5 variants using general-domain or biomedical initialization and training from scratch, with clinical masked-span pretraining.	MIMIC-III and MIMIC-IV clinical text.	Released clinically pretrained encoder–decoder models that can be fine- tuned for downstream NLP tasks.	The released models are not already fine-tuned for a specific task; the resource advises against treating them as ready for zero-or few-shot use. It provides a candidategeneration backbone, not a validated stroke EHR-to-report system.
Liu et al. (2023a) [13]—Stroke imaging assessment	Automate ASPECTS assessment from acute stroke MRI.	Infarct segmentation and regional injury quantification, followed by interpretable machine-learning prediction; an ensemble combines random forest, multilayer perceptron, and support vector machine models.	Clinical diffusion-weighted MRI; a reported 300-case training partition and 100-case testing partition.	Automated ASPECTS predictions were comparable to inter-reader agreement on several measures; outputs included regional scores, probabilities, and explanatory features.	This is imaging-based scoring rather than Transformer-based EHR-to-text generation. Atlas Alignment, anatomical variation, and class imbalance affected performance. Its contribution is a potential source of structured imaging evidence for a broader reporting system.
Liu et al. (2023b) [14]—Stroke radiological reporting	Generate comprehensive descriptions of acute ischemic injury from MRI.	Automated detection, segmentation, anatomical quantification, and machine-learning analysis producing radiological report outputs.	Clinical acute stroke diffusion-weighted MRI; an additional independent STIR testing cohort of 100 cases was used to assess generalization.	Demonstrated automated reporting of imaging-derived lesion characteristics alongside infarct masks, volumes, and regional injury information.	Directly relevant to stroke reporting, but its input is MRI rather than a complete longitudinal EHR. It does not establish Transformer generation from medications, laboratory results, neurological scores, and treatment histories.
Rouhizadeh et al. (2026) [15] — Clinical narrative generation	Assess bilingual medical report generation and whether experts can distinguish generated reports from human-written reports.	Benchmarking of multilingual open-weight LLMs using automatic similarity metrics, blinded expert ratings, and human-versus-machine detection.	BiMedReport-4K: 4,212 English and French reports across specialties; generation used synthetic EHR inputs derived from source narratives. Nine clinical experts participated in evaluation.	Phi-4 achieved ROUGE-1 of **0.70** and BERTScore of **0.83**. Overall expert ratings were **4.6/5**; human detection accuracy was **0.60**, compared with **0.98** for automatic classifiers.	The controlled, multispecialty setting is not a prospective stroke EHR deployment. Synthetic inputs and report provenance limit direct transfer to routine hospital documentation. Fluency and difficulty detecting AI authorship do not establish factual safety; stroke-specific source verification remains necessary.

Duarte et al. (2026) — [16] Stroke-focused secondary evidence	Synthesize applications, challenges, and future directions of LLMs and NLP in stroke care.	Systematic literature review across six databases.	**2,991 screened records**, yielding **65 eligible studies**.	Provides a stroke-focused overview of the research landscape and a basis for identifying relevant application categories.	Broad stroke NLP evidence must be separated by task. Classification, information extraction, question answering, and report generation should not be pooled as equivalent evidence of structured EHR-to-narrative performance.

Table 2: EHR data resources cited in the review and their implications for constructing a stroke report-generation benchmark.

Resource study	Data contribution	Potential role in the proposed review	Critical limitation / benchmark requirement
Johnson et al. (2016): MIMIC-III [5]	De-identified critical-care records containing structured clinical variables and narrative documentation.	Supports investigation of relationships between patient measurements, interventions, diagnoses, and clinical narratives.	A general EHR database is not automatically a stroke input–report benchmark. Cohort selection, encounter linkage, report targets, missingness, and patient-level splits must be explicitly defined.
Johnson et al. (2023): MIMIC-IV [6]	Research-accessible EHR data from Beth Israel Deaconess Medical Center.	Provides a basis for constructing structured clinical inputs and linking appropriate documentation resources.	The exact database version, linked notes resource, stroke cohort, and input availability at report-generation time must be specified. Single-centre data cannot independently establish cross-hospital generalization.
Pollard et al. (2018): eICU-CRD [7]	Multicentre critical-care data covering more than 200,000 ICU admissions.	Supports investigation of institutional variation and portability of structured clinical features.	Availability of suitable narrative targets must be established before presenting it as an EHR-to-report dataset. Multicentre structured data alone does not guarantee paired report-generation supervision.

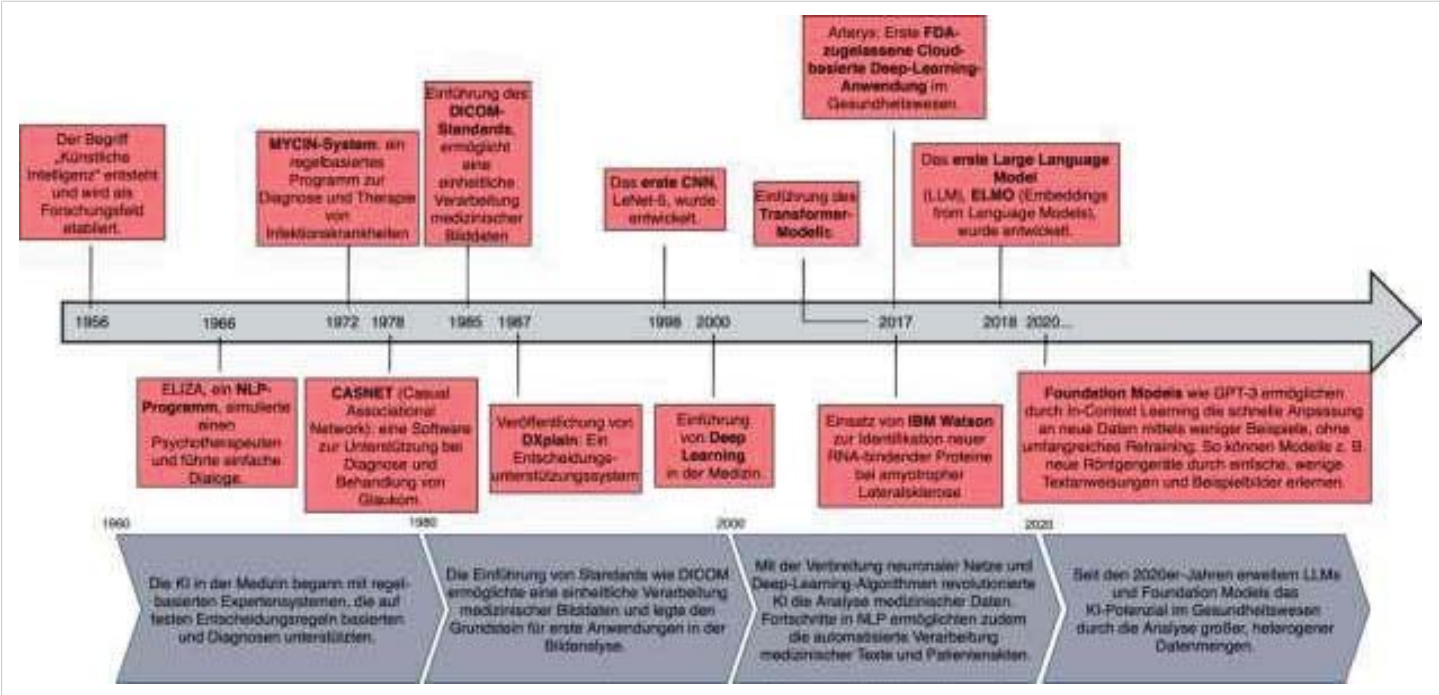


Figure 4: Historical evolution of AI and NLP in medicine: from rule-based expert systems to foundation models (1956–2020+).

In the late 2010s and 2020s, there were the advent of Large Language Models (LLMs) and Foundation Models. In 2018, ELMO, which was one of the earliest LLM models, was released and has been followed by many GPT-like models, such as the Foundation Models that enable the model to rapidly adapt to new tasks through in-context learning without a need for extensive retraining. Arterys is the first and only cloud-based deep learning application approved by the FDA for the healthcare space in 2020. The bottom portion of the timeline shows the broader impact: The use of AI in medicine started with rule-based expert systems, led to the advent of neural networks and deep learning for breaking down medical data in unprecedented ways, and is now extended into the use of LLMs and foundation models, which can process huge amounts of information from large and diversified datasets. Namely, this historical trajectory directly assists the current review by providing context: We can see that a progression occurred from early symbolic systems to more recent models like the Transformer-based models (T5, BART, GPT, LLaMA, Phi), which are well suited for the tasks of report generation in the stroke EHR domain (see Sections 2 and 4 of this paper).

Figure 5 shows the evolution of the architecture of a sequence-to-sequence model from an RNN to an attention model based on Transformers. Panel A illustrates the classical encoder-decoder RNN configuration where the encoder takes in the input sequence and squeezes it down to a last hidden state, and the decoder takes the last hidden state and produces one output token at a time. This design has sequential processing and loss of information, particularly for long

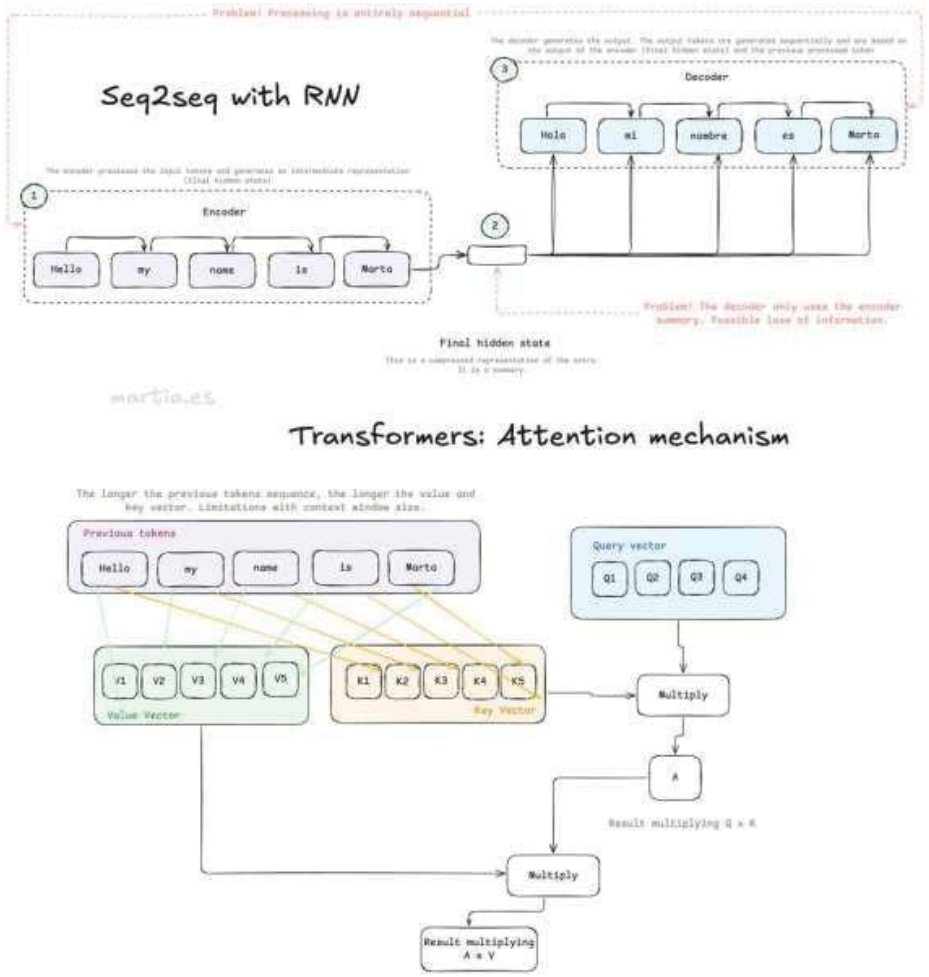


Figure 5: Evolution of Medical NLP Models: RNN Seq2Seq to Transformer Attention.

clinical documents. The Transformer attention mechanism is used in Panel B, where the value vectors and key vectors are generated from the previous tokens, and query vector generates attention weights. This enables parallel processing, improved long-range dependency modeling and more clinical text generation that is "context-aware".

The key point of interest is that the three most popular paradigms in the current natural language processing era are identified: Text-to-Text (T5), Bi-Encoder (BERT), and a new paradigm (GPT-2). The important thing to take note of is the three major paradigm architectures that are recognized in the field of modern natural language processing (NLP): Text-to-Text (T5), Bi-Encode (BERT), and a new paradigm (GPT-2). Because of its encoder-decoder design, T5 is well suited to tackle tasks like sequence-to-sequence, like generating a clinical narrative from structured EHR inputs. BERT has been specifically adapted to only encode information for better understanding and feature extraction, which enables the model to work well on clinical tasks such as named entity recognition (NER) and classification tasks, but not on tasks that require generating text, like report writing. Unlike BERT, GPT-2 is a decoder-only model that was pre-trained using causal language modeling, which allows for fluent and coherent text generation, with a less strong bidirectional capacity for understanding context.

This Table 3 outlines the key differences between these models with respect to their pre-training objectives, position embeddings, the number of parameters, and their respective strengths and weaknesses. In the context of stroke EHR-to-report generation, the sequence-to-sequence nature of T5 affords the most direct architectural match, as the generation sequence is equivalent to the entire generation process in EHRs. Its weaknesses are, however, that it needs both the encoder and the decoder during inference, such that its resource usage is higher. The pre-trained BERT model is useful for extracting structured features from unstructured clinical notes that can be inputted into a generative model. The framework of GPT-series models guarantees much higher fluency and longer coherence, but they can also incur the risk of factual errors and hallucination. This comparison validates the choice of particular architectures undertaken in the present work, and briefly delineates the advantages/disadvantages of text interpretation/generation and computational efficiency in clinical NLP.

Comparing Transformer-based Models

Aspect	T5	BERT	GPT-2
Architecture	Encoder-Decoder	Encoder-Only	Decoder-Only
Primary Use Case	Sequence-to-sequence tasks	Text understanding tasks	Text generation tasks
Pre-training Objective	Denoising span corruption	Masked Language Modeling	Causal Language Modeling
Position Embedding	Relative position embeddings	Absolute position embeddings	Absolute position embeddings
Parameter Count	Variable (e.g., T5-base: 220M, T5-Large: 770M)	Variable (e.g., BERT-Base: 110M, BERT-Large: 340M)	Variable (e.g., GPT-2 Medium: 345M, GPT-2 Large: 774M)
Strengths	Adaptability to diverse tasks with a single architecture.	Strong text understanding from bidirectional context	Fluent text generation with long-term coherence
Weakness	Requires both encoder and decoder during inference, making it more resource-intensive	Is limited to text understanding, cannot generate coherent text	Has limited understanding of bidirectional context
Example Models	T5-Small, T5-Large, T5-11B	BERT-Base, BERT-Large, RoBERTa	GPT-2, GPT-3, ChatGPT

Table 3: Comparative analysis of Transformer-based architectures (T5, BERT, and GPT-2) for clinical NLP and report generation tasks.

Figure 6 shows the evolutionary tree for the development of text generation AI models, stepping from rule-based models to more advanced and multilingual models. The bottom steps are baseline systems: Rule-based systems use hand-coded symbolic rules; Statistical Models use probabilistic patterns for recognising patterns. The next step was Neural Networks, which brought in the concepts of deep learning and representation learning. Transformer Models are the central step and a major breakthrough in the field by applying self-attention, allowing the parallel processing of text and generating context-appropriate text. This is the architectural paradigm that is the focus of this review, especially for stroke EHR to report generation.

The top steps depict new and upcoming horizons: Multimodal Generation (Incorporating visuals and structured data), Entity-Aware Models (Focused on specific clinical entities, factual consistency), Graph Neural Networks (Modelling complex relational data), and Multilingual Models (Cross-lingual approach for clinical documentation). This evolution reflects that, although Transformer-based models are highly prevalent in the current stroke NLP research, the research community has shifted to increasingly complex, multimodal, entity-aware models. The upper tiers are also directly connected with the potential to overcome the limitations on the factual accuracy, clinical validity, and generalizability of care currently available for stroke, such as multimodal learning (EHR + MRI + CT) and retrieval-augmented generation.

Transitions of Care (ToC) – transfer of SEHR data for clinical report generation

Stroke Electronic Health Records (EHRs) are complex clinical records that include heterogeneous information

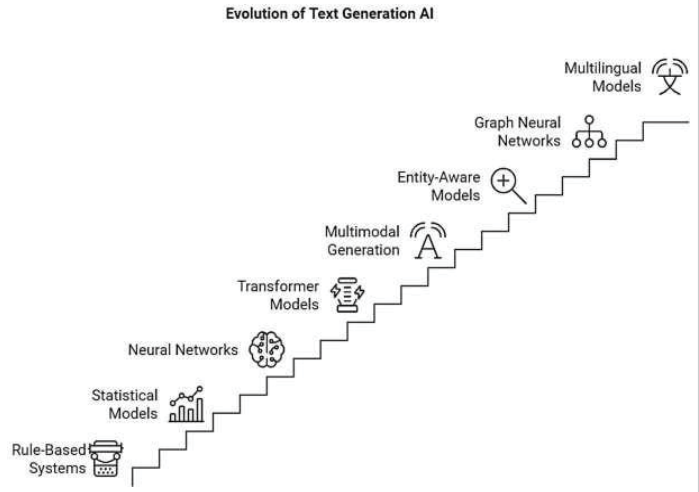


Figure 6: Evolution of Text Generation AI: A capability staircase from rule-based systems to multilingual models.

needed for diagnosis, treatment, monitoring, and patient documentation. The information that is recorded includes vital signs, laboratory values, neurologic evaluations, medications, imaging interpretations, events associated with treatment, and demographic information, among other data.

Documenting the stroke is an important part, and the National Institutes of Health Stroke Scale (NIHSS) is a standardized assessment of neurological impairment. An NIHSS score can be added to the longitudinal clinical data, which can provide a good representation of changes in neurologic status during hospitalization.

In acute stroke care, there is a critical time sensitivity to the information. These may be relevant times, such as when the symptoms began or the last time the person was known to be well, hospital arrival time, time of imaging, time of treatment, or, if intravenous thrombolysis was administered, the "door-to-needle" time. These events need to be represented as accurately as possible with the purpose of maintaining the clinical timeline.

Automated report-generating systems should also cover patient-safety issues with incorrect numerical values, missing findings, unsupported diagnoses, and inaccurate treatment descriptions. Thus, the making of a stroke report is not just about fluent writing; it's also about factual consistency, the accuracy of the time, and clinical verification.

- Laboratory measurements
- Neurological scores (NIHSS)

The clinical notes and discharge summaries will be reviewed. The clinical notes and discharge summaries will be checked.

Major public datasets

MIMIC-III and MIMIC-IV: structured and narrative clinical data in large datasets stored in the ICU. They are widely used

for clinical NLP and filtered using ICD codes to include stroke-relevant cases. eICU-CRD dataset contains structured clinical variables from multiple centers across hospitals. UK Biobank: large cohort of longitudinal design with incidence data for stroke. PhysioNet Stroke Annotations includes NIHSS-related structured annotations that are connected to clinical text.

The basic process of acquiring and storing hospital and ICU data to create a research-ready EHR database like MIMIC-II/III/IV is depicted in Figure 7. This is the top part, which features bedside monitoring, chart data, tests, orders, billing, demographics, and clinical notes, plus ICU types like MICU, SICU, CCU, CSRU, and NICU. These data are fed through a data archive, de-identification, date shifting, and formal conversion prior to being placed in MIMIC. Also included in the figure are the external data—Social Security Death Index, user feedback/corrections. "Major Public Datasets" highlights the significance of example major public datasets for stroke NLP research, such as MIMIC, eICU-CRD, UK Biobank, and PhysioNet.

The 4-layer clinical AI architecture is shown in Figure 8. The data input layer includes medical data like EHR and DICOM data, clinical data like reports and laboratory data, and external data like literature and guidelines (Layer 1). Layer 2: Data Processing, which includes data preprocessing, data validation, data cleaning, feature extraction, and embedding of text and images. The AI processing core (Layer 3) comprises a knowledge base, a vector database, semantic search, RAG system, query processing, context retrieval, response generation, and a medical AI decision support LLM. Layer 4 – Output and User Interaction generates clinical reports, automated analysis, risk assessment, treatment recommendations, decision support and clinical insights. The architecture also encompasses standards including HL7 FHIR, SNOMED CT, ICD-10; AI models like BioBERT, clinical Transformers, and RAG; privacy features like HIPAA compliance and differential privacy; and a system feedback loop for updates.

Text Generation AI has evolved over time, and Figure 9 provides a summary of its historical evolution. Panel A looks at the evolution from early rule-based and statistical systems, into neural networks, Transformers, large language models, to the foundation models. It emphasizes that it is moving from a symbolic, handcrafted decision rule to a data driven generative model like GPT-3 and newer LLMs. Panel B shows a capability staircase ranging from rule-based systems through to statistical models, neural networks, Transformer models, multimodal generation, entity-aware models, graph neural networks, and multilingual models.

Methods for generation of clinical reports based on a transformer

Problem formulation

If the EHR input comes in a structured form:

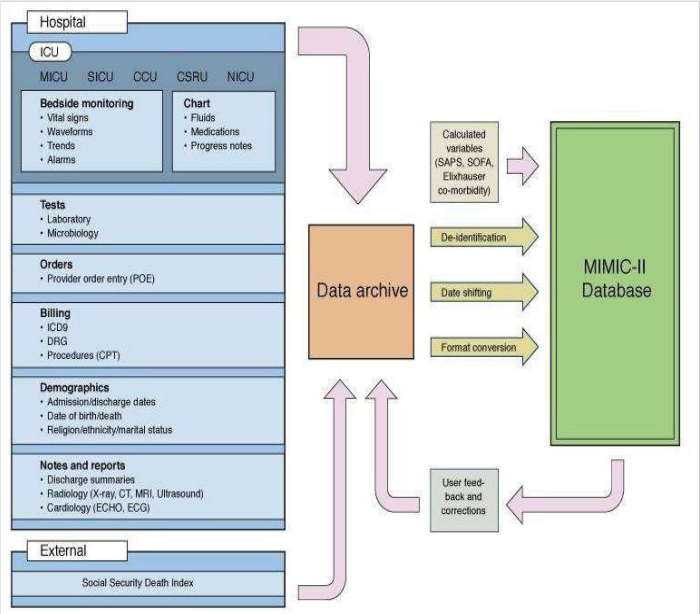


Figure 7: Major EHR Datasets Used in Stroke NLP.

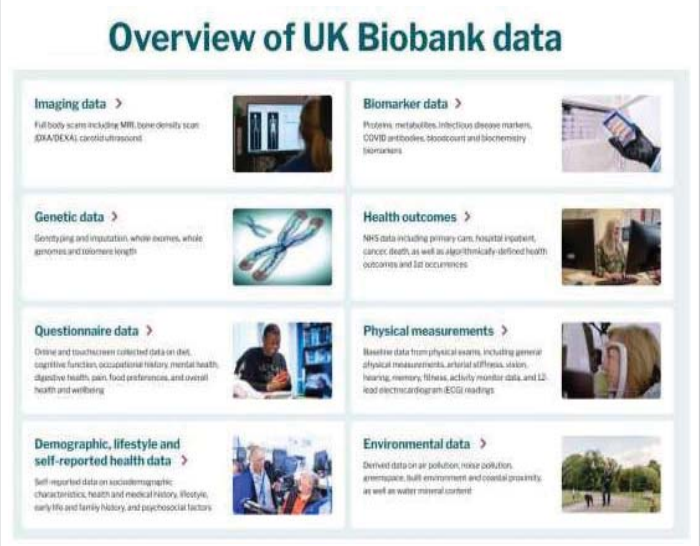


Figure 8: End-to-End Clinical AI System Architecture for Stroke Report Generation.

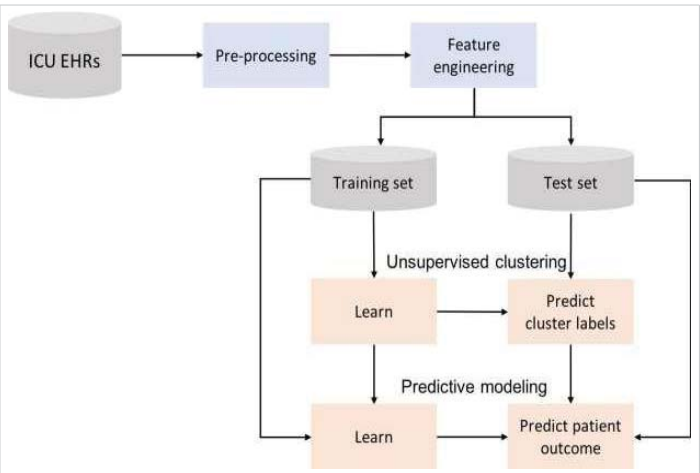


Figure 9: Evolution of Text Generation AI.

X contains the following information: demographics, vitals, labs, ICD codes, NIHSS scores. The objective is to generate:

Y = clinical narrative report.

such that:

$P_{\theta}(Y\mid X)=\prod_{t=1}^{T}P_{\theta}(y_t\mid y_{<t},X)$

where fθ is a model built by transformer.

Model architectures

- Encoder-Decoder Models: T5, BART
- Decoder-only and decoder-encoder LLM: BLOOM, BLOOM-7B, BLOOMZ

A new model called Clinical-T5 was developed to improve clinical QA. To improve Clinical QA, a new model, Clinical-T5, was developed.

Training strategies

- Clinical Corpus Fine-tuning with supervision
- Oracle's ability to learn about new domains using domain-adaptive pretraining (MIMIC, PubMed); instruction tuning of LLMs
- Retrieval-augmented generation (RAG)

The transformer-based encoder-decoder architecture, which is commonly adopted for EHR-to-report generation, is shown in Figure 10. The input embedding and positional encoding go through the encoder stack, consisting of multi-head attention, feed-forward layers, and add-and-normalize residual connections. The decoder stack has masked multi-head attention, multi-head attention over the encoder output, feed-forward layers and add-and-normalize blocks. The final layer in both linear and softmax is a layer with output probabilities corresponding to the clinical narrative generated. This architecture is the same as the problem formulation Y=fθ(X), where X is structured stroke EHR data, and Y is the generated clinical report.

Figure 11 shows the major data domains available in UK Biobank. These comprise imaging data (MRI, DXA, carotid ultrasound), biomarker data (proteins, metabolites, cancer registry, biochemistry markers), genetic data (genotyping, whole-exome sequencing), health outcome data (NHS primary care data, hospital inpatient data, cancer registry, NH death registry), questionnaire data, physical measurements, demographic, lifestyle, self-reported health, and environmental data. UK Biobank is valuable for stroke NLP, as it offers a wealth of longitudinal and multimodal information.

Figure 12 shows an example of the machine learning pipeline for an ICU EHR. The first step in the raw ICU EHRs

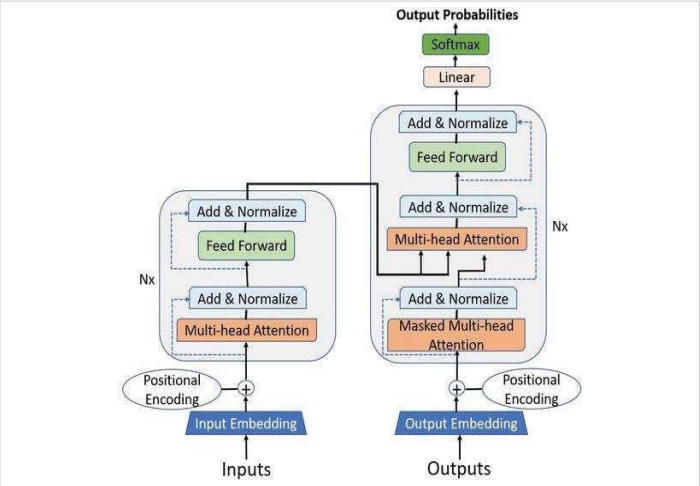


Figure 10: Transformer-Based Report Generation Architecture.

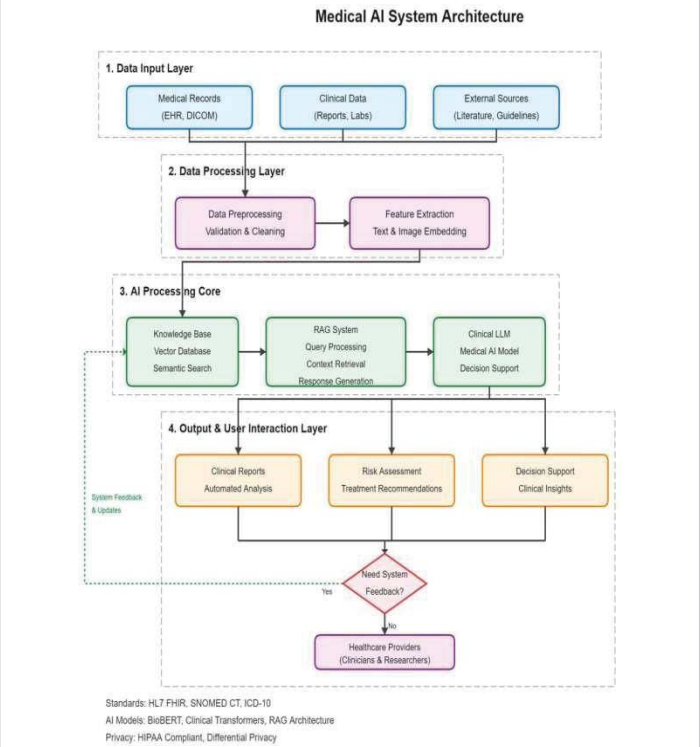


Figure 11: Overview of UK Biobank Data Domains Relevant to Stroke NLP.

is preprocessing and feature engineering. The outcomes are divided into training set and a test set. The training set is used for unsupervised clustering, and labels for clusters are predicted on the test set. At the same time, the predictive modeling is carried out using the training set, and predicting the outcomes are predicted on the test set. This number can be useful for demonstrating the structured nature of the data in the ICU and its potential for use in subsequent applications such as stroke subtype clustering, prediction of outcomes, and generation of clinical reports.

Evaluation metrics

Automatic metrics

- BLEU (n-gram overlap)

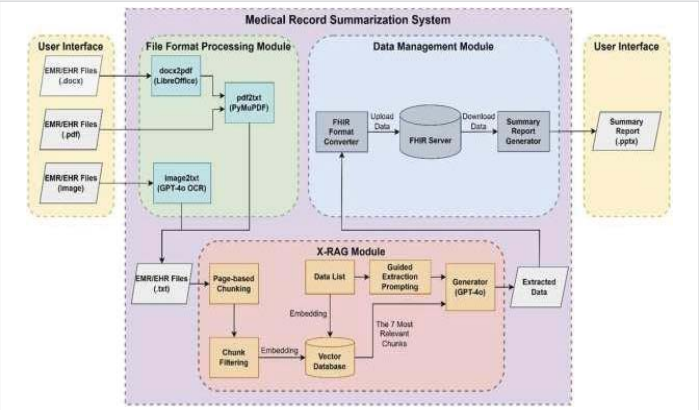


Figure 12: ICU EHR Preprocessing and Predictive Modeling Pipeline.

- ROUGE-L (sequence similarity)
- METEOR
- BERTScore (semantic similarity)

Clinical metrics

Clinical evaluation metrics

Besides traditional natural language generation metrics, domain-specific metrics need to be used to assess the clinical reliability of the stroke report generated by Transformer.

Clinical entity accuracy: Assesses the accuracy of representation of diagnosis, medications, neurological findings, and other clinical entities in the generated report.

NIHSS consistency: Assesses whether the documented severity scores for the neurologic system are maintained in a consistent manner without a numeric change or missing.

Numerical accuracy: Evaluates the degree of agreement between the laboratory values, vital signs, medication dosages, and other quantitative observations generated and those in the source records.

Temporal consistency: Verifies that the produced report maintains correct chronological order and timing of the important clinical events.

Hallucination rate: Number of clinically supported statements or findings that are made that are not substantiated in the source EHR.

Clinical completeness: Judges if clinically relevant details are missing from the created text.

Treatment documentation accuracy: Evaluates the accuracy of written interventions, medications, and events related to treatments.

These measures should be assessed against clear criteria of annotation and using clinician-reviewed referral information. None of the automatic metrics are sufficient on their own to ensure clinical safety.

- Medical entities are factually correct (content aspect)
- NIHSS consistency
- Hallucination rate
- Treatment accuracy

Human evaluation

- Clinical correctness
- Fluency
- Completeness
- Usefulness for decision-making

Baseline models

Table 4 summarizes the architecture, main application domain, pre-training objective, position embedding, number of parameters, and strengths and weaknesses of T5, BERT and GPT-2. T5 is a type of encoder-decoder model that is ideal for sequence-to-sequence applications like EHR-to-report. BERT is a type of encoder-only model which is best suited for text understanding, classification and entity extraction. The decoder-only model GPT-2 is more effective for fluent text generation, while having limited bidirectional understanding.

Key findings from literature

New resResearch has proven:

- Transformers are better than RNN-based systems for the clinical text generation task.
- The FactAccuracy was enhanced by domain-specific pretraining.
- In contrast, under controlled conditions, LLM's can become nearly as fluent as humans!
- The studThe use of report generation for stroke is under-researched.

A clinical report generation example from Rouhizadeh, et al. (2026) [15] showed ROUGE-1 scores of ~0.70 and BERTScore of ~0.83, and the expert-rated quality of these reports was comparable to human performance.

Research gaps

- While progress has been made, there are still a number

Table 4: Comparison of Transformer-based models for clinical NLP and report generation

Model	Type	Role
Rule-based templates	Symbolic	Baseline
LSTM Seq2Seq	RNN	Early neural model
T5	Transformer	Strong baseline
BART	Transformer	Summarization
GPT-4	LLM	Upper-bound
ClinicalBERT	Encoder	Feature extraction

of shortcomings: No large-scale stroke EHR-to-report paired dataset. There are no standardised evaluation benchmarks.

- Reports with hallucination in text generation The clinical validity and reliability in real hospitals are weak. The output of the Transformer is hard to interpret.

Future directions

Future studies should include: Multimodal learning (EHR + MRI + CT)

- Retrieval-augmented generation (RAG)

However, privacy can be preserved using federated learning.

- Stroke-specific LLM pretraining
- Clinician-in-the-loop validation systems

Conclusion

The transformer models are one of the most significant advances in automatic clinical report generation from structured EHR information. These systems show great promise in supporting efficiency and consistency in stroke care documentation. But the availability of datasets, control of hallucination, and clinical validation are the limitations that prevent real-world deployment. Important areas for future research will be stroke-specific data, multimodal integration, and clinically relevant evaluation frameworks.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. NeurIPS. 2017.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. 2019.
3. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR. 2020;21(140):1–67.
4. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation. ACL. 2020.
5. Johnson AEW, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. doi: 10.1038/sdata.2016.35
6. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2021;8:1–9. Available from: https://dx.doi.org/10.1038/s41597-021-01036-9
7. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. Available from: https://dx.doi.org/10.1038/sdata.2018.178
8. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. ICLR. 2020.
9. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. ACL Workshop. 2004.
10. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. ACL. 2002.
11. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv preprint. 2019.
12. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240. Available from: https://dx.doi.org/10.1093/bioinformatics/btz682
13. Liu CF, Li J, Kim G, et al. Automatic comprehensive ASPECTS reports in clinical acute stroke MRIs. Sci Rep. 2023;13:3784. Available from: https://dx.doi.org/10.1038/s41598-023-30242-6
14. Liu CF, Zhao Y, Yedavalli V, et al. Automatic comprehensive radiological reports for clinical acute stroke MRIs. Commun Med. 2023;3:95. Available from: https://dx.doi.org/10.1038/s43856-023-00327-4
15. Rouhizadeh H, et al. The Detectability Paradox: Bilingual Medical Report Generation with Open-Weight Models and the Limits of Human Oversight. J Am Med Inform Assoc. 2026. Available from: https://dx.doi.org/10.1093/jamia/ocag070
16. Duarte KT, Sidhu AS, Bakshi M, et al. Systematic Review of Large Language Models and NLP in Stroke Care: Applications, Challenges, and Future Directions. Stroke Vasc Interv Neurol. 2026;6(3). Available from: https://dx.doi.org/10.1161/SVIN.125.002261
17. Lehman E, Johnson A. Clinical-T5: Large Language Models Built Using MIMIC Clinical Text. PhysioNet. 2023. Available from: https://dx.doi.org/10.13026/rj8x-v335
18. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–e220. Available from: https://dx.doi.org/10.1161/01.CIR.101.23.e215