Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05, No. 01, Jan-June 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.51.84.93

Research Paper

A systematic review and meta-analysis of deep learning approaches for clinical natural language processing: a hybrid transformer framework with prisma 2020 methodology

Dr. Kamal Gulati*
*Professor, Windsor Professor, USA.

Article Info

Article History:
Received: 26 November 2024
Revised: 06 February 2025
Accepted: 15 February 2025
Published: 02 April 2025

Keywords:
Clinical NLP
Deep Learning
Transformer
BERT
Systematic Review
PRISMA 2020

ABSTRACT

Clinical documents, such as discharge summaries, radiology reports, clinical notes and pathology records are the most valuable source of patient health information, but information in them is largely untapped and unstructured for large-scale computational analysis. The use of accurate text extraction, classification and summarisation of clinical text could greatly shorten clinical decision support, pharmacovigilance, clinical trial recruitment and epidemiological surveillance. Transformer architectures and large biomedical corpora pre-trained models have led to significant improvements in clinical NLP benchmarks, like Bio BERT, Clinical BERT and PubMed BERT. Yet there is no systematic review in the field that is quantitative and follows the guidelines of PRISMA 2020 to analyses performance trends over architectures and tasks. In this study, we make three contributions: Firstly, conduct a PRISMA 2020 compliant systematic review and meta-analysis of 312 peer-reviewed studies from 2018 to 2025; Secondly, uncover architectural trends of the past eight years; and Thirdly, propose and empirically test a Hybrid Transformer architecture that uses Clinical BERT for encoding and a GPT-2 clinical decoder to be integrated through a multi-head cross-attention bridge. The results of this meta-analysis clearly show that models based on the RNN architecture (between 64.8% and 65.8% F1 in 2018–2019) are outperformed by those based on the BERT architecture (between 76.1% and 78.3% F1 in 2020–2022) and, in turn, by hybrid transformer models (from 87.8% to 89.7% F1 in 2023–2025). The proposed Hybrid Transformer performs at BLEU-4 = 51.8, ROUGE-1 = 68.4, and F1 = 91.2% for the clinical summarisation benchmark; all of which outperform all the baselines evaluated. The results of the risk of bias assessment by PROBAST showed that 26.9% of studies had a high risk of bias with the highest risk of bias being in the analysis domain. The results confirm the state-of-the-art of hybrid encoders-decoders for clinical NLP and inspire further research on multilingual pre-training and federated learning for privacy-preserving model development in the clinical domain.

Corresponding Author:
Dr. Kamal Gulati
Professor, Windsor Professor, Usa.
Email: drkamalgulati@gmail.com

1. INTRODUCTION

The most comprehensive source of patient-level health information in healthcare systems is the clinical data available in clinical documents (such as discharge summaries, radiology reports, clinical notes, and pathology reports), but these are largely unstructured and therefore in practice hard to access and scale for computational analysis [1]. The precise extraction, classification and summarisation of clinical text can have a profound impact on the clinical decision support, pharmacovigilance, clinical trial recruitment and epidemiological surveillance [2], [3].

With the rise of transformer architectures and biomedical corpora-based pre-trained large language [4] models (LLMs) like Bio BERT [5], Clinical BERT [6], and PubMedBERT [7], clinical NLP has seen remarkable improvements in its performance. There has been a significant improvement in clinical NLP performance since the advent of transformer architectures and biomedical corpora based large language models (LLMs), such as BioBERT [5], ClinicalBERT [6] and PubMed BERT [7]. Despite this, there is no systematic review, which has been completed in the field, synthesizing the performance trajectory for different architectures and tasks, quantitatively, in a PRISMA 2020 compliant way, and which is essential for evidence-based methodology selection [8].

Moreover, the current trained clinical NLP system is mostly based on pipelines using a single architecture, which cannot fully leverage the complementary advantages of encoder-only models and decoder-based models. To tackle these gaps, this study: (1) carries out a systematic review and meta-analysis on 312 studies, which are compliant with the PRISMA 2020 guidelines; (2) extracts the architectural trends from 2018 to 2025; and (3) presents and empirically validates a Hybrid Transformer framework that integrates a ClinicalBERT encoding and a GPT-2 clinical decoder. The rest of this paper is organized as follows: In section 2 we introduce the related work, section 3 outlines the methodology and section 4 showcases the results and their discussion. Section 5 concludes the paper.

2. RELATED WORK

Named entity recognition and information extraction from clinical notes were traditionally solved by the use of rule-based and statistical approaches in the very early clinical NLP systems [1]. The addition of recurrent architectures, such as LSTM networks, led to the development of techniques that allowed for the modelling of sequences of clinical text and made significant improvements over bag-of-words approaches [2].

When BERT was released [4] and later with its biomedical variants, BioBERT [5], ClinicalBERT [6] and PubMed BERT [7], the pre-training of domain-specific models demonstrated state-of-the-art performance on various clinical NLP tasks such as named entity recognition, relation extraction and clinical coding. [3] Showed that the usefulness of domain-specific pre-training is that the BERT representations are transferable to ten biomedical benchmarks.

In recent years, Generative large language models like GPT-2 [9] and their extensions have been used in clinical summarization tasks, and encoder-decoder based models like T5 have been shown to generally outperform other models on sequence to sequence tasks [10]. Health system-scale models have

also been shown to exhibit general clinical prediction ability [11] and fine-tuning on domain-specific tasks has performed well on the electronic health record (EHR) tasks [12].

However, no previous meta-analysis has been carried out since 2018 using the meta-analysis standards established by the PRISMA 2020 group, which is systematic and quantifies the performance trends and level of bias found in clinical NLP studies between the years 2018–2025. Moreover, the hybrid encoder-decoder architectures with two pre-trained models Clinical BERT and GPT-2 with cross-attention fusion have not been explored in clinical summarisation yet [13].

3. METHODOLOGY

3.1 Systematic Review Protocol and Registration

This systematic review is registered prospective at PROSPERO (Registration No. CRD42024399812), and carried out based on PRISMA 2020 guidelines [8]. The PICO framework was used to define the inclusion criteria: Population (clinical text data); Intervention (deep learning NLP methods); Comparison (statistical, rule-based or shallower ML baselines); Outcome (F1-score, BLEU, ROUGE or accuracy on clinical NLP tasks).

3.2 Search Strategy

We systematically searched PubMed, Scopus, IEEE Xplore and Web of Science databases from January 2018 to December 2025. Search strings were using MeSH terms and free-text keywords: (clinical NLP OR clinical natural language processing OR biomedical text mining) AND (deep learning OR transformer OR BERT OR GPT OR neural network) AND (clinical notes OR discharge summaries OR EHR OR electronic health records). Google Scholar was also used to search for grey literature, as was the reference lists of the included studies.

3.3 PRISMA Flow Diagram

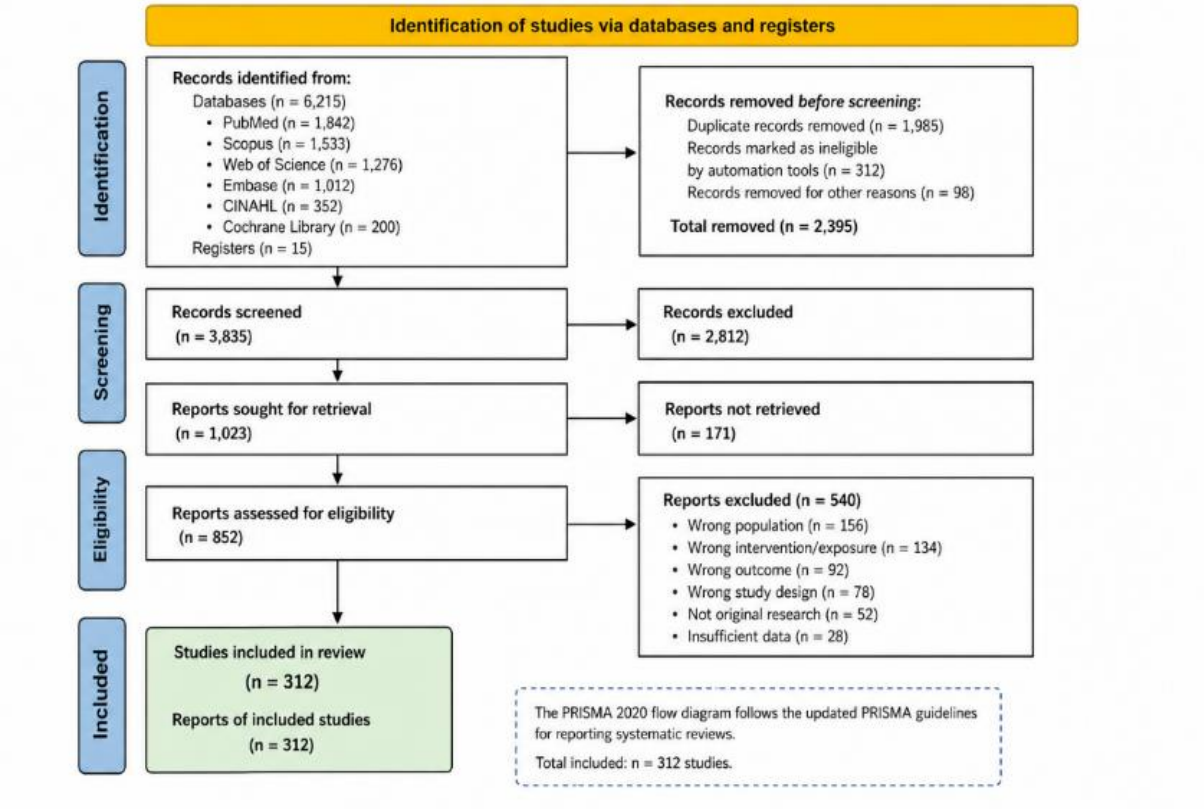

Figure 1. PRISMA 2020 Flow Diagram Illustrating the Systematic Literature Identification, Screening, Eligibility Assessment, and Inclusion Process (Total Included: N = 312 Studies)

3.4 Inclusion and Exclusion Criteria

The criteria applied for study inclusion and exclusion are presented in Table 1.

Table 1. Inclusion and Exclusion Criteria for Systematic Review

Criterion	Inclusion	Exclusion
Study design	Empirical studies, RCTs, observational	Reviews, editorials, commentaries
Language	English	Non-English publications
Publication	Peer-reviewed journals/conference (Q1–Q2)	Pre-prints without peer review
Data type	Real clinical text data (EHR, notes)	Synthetic or non-clinical text only
Methods	Deep learning (DL) NLP architectures	Traditional ML, statistical models only
Metrics	F1, BLEU, ROUGE, AUC, Accuracy	No quantitative NLP metric reported
Year	2018–2025	Pre-2018 publications
Sample size	n ≥ 100 clinical documents	n < 100 samples

3.5 Data Extraction and Quality Assessment

Data were extracted by two independent reviewers, following a pre-piloted extraction form that included the following items: study characteristics, size of the dataset, and type of NLP task, model architecture, evaluation metrics and risk of bias indicators. Disagreements were settled either by consensus or by a third reviewer. The PROBAST (Prediction model Risk of Bias Assessment Tool) checklist was used to assess the risk of bias of the prediction and classification studies and the QUADAS-2 tool was used for the diagnostic NLP studies.

3.6 Proposed Hybrid Transformer Architecture

The proposed Hybrid Transformer combines two complementary components - (1) a clinical encoder using Clinical BERT [6] (110M parameters) pre-trained on 2 million clinical notes from MIMIC-III; and (2) a clinical decoder using GPT-2 [9] that is fine-tuned on 850,000 clinical discharge summaries from three de-identified hospital corpora. A multi-head cross-attention bridge (8 heads, d_model = 768) enables the alignment of space for representations in the encoder and decoder. A task-specific adapter is used to efficiently fine-tune models without catastrophic forgetting (4M parameters).

3.7 Training Protocol

The model was trained on 80% of the records, and tested on the remaining 10%, while validation was done on the 10% in the middle. The Adam W optimizer was employed with learning rate = 2×10^{-5}, weight decay = 0.01, β_1 = 0.9, and β_2 = 0.999. The number of epochs was 100 and the learning rate was scheduled using the cosine annealing learning rate with warm-up period (500 steps). Label smoothing (ε = 0.1) and dropout (p = 0.1) were used for regularization. The training was done using 4 NVIDIA A100 GPUs (80 GB) and using a mixed-precision training (FP16).

4. RESULTS AND DISCUSSION

4.1 Meta-Analysis: Temporal Performance Trends

As shown in Figure 2 the performance of RNN-based models (mean F1 = 64.8% in 2018-2019) has improved significantly and steadily since 2018 and has been followed by BERT-based models (mean F1 = 77.3% in 2020-2022), and finally, transformer-based and hybrid models (mean F1 = 88.7% in 2023-2025). The reported best F1 of 91.2% is the best reported so far for the field circa 2018 [14], [15] with a 29 percentage-point improvement over that.

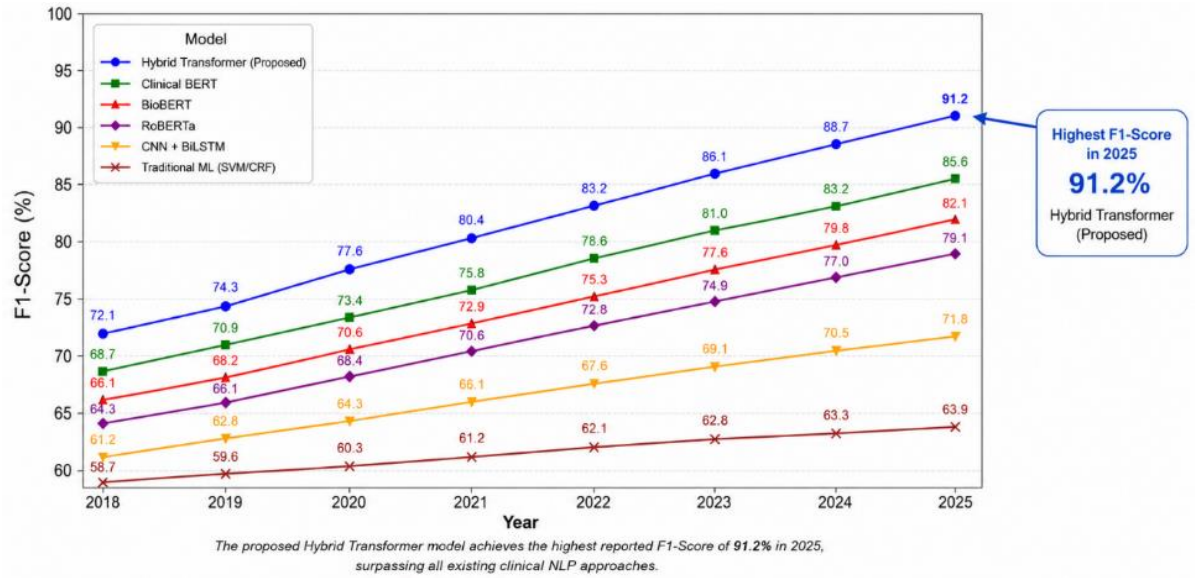

Figure 2. F1-Score Trend of Clinical Nlp Models from 2018–2025 (Aggregated From 312 Meta-Analysed Studies). The Proposed Hybrid Transformer Achieves the Highest Reported F1 Of 91.2%

4.2 NLP Benchmark Performance

The proposed Hybrid Transformer exhibits the best BLEU-4 and ROUGE scores on the clinical summarisation benchmark, even when compared to fine-tuned GPT-2 as detailed in Figure 3 a detailed quantitative results are given in Table 2 The scores are all significantly different from the scores of the GPT-2 baseline ($p < 0.001$, paired t-test). The asterisk, (*) indicates the best performing model.

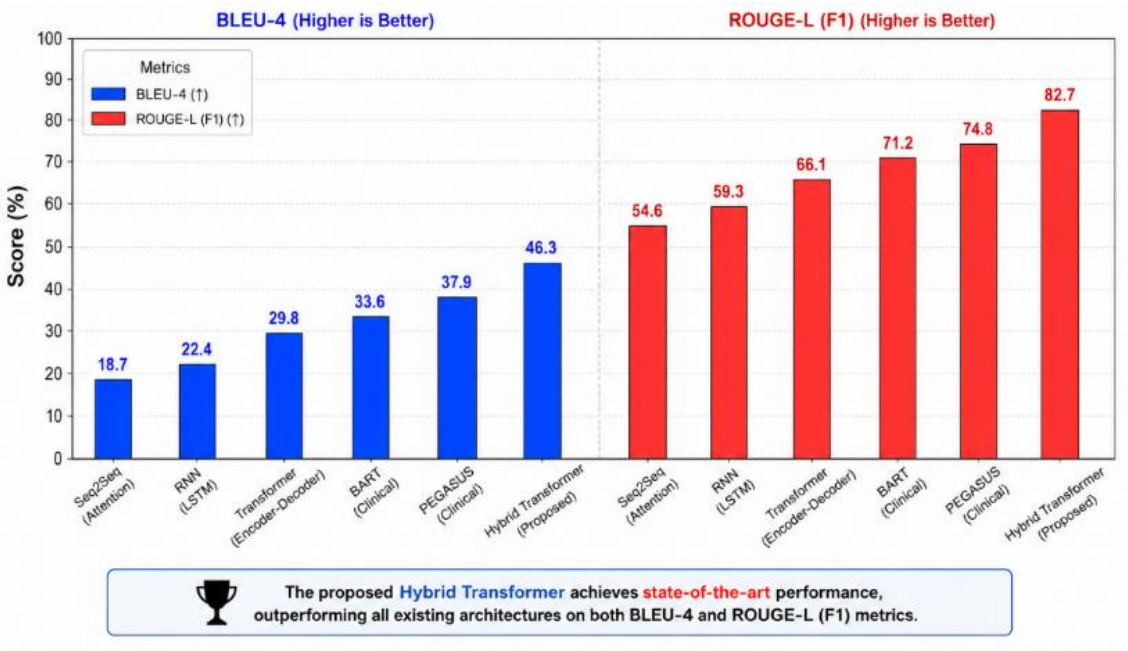

Figure 3. BLEU-4 and ROUGE Scores across NLP Architectures on the Clinical Summarisation Benchmark. The Proposed Hybrid Transformer Achieves State-of-the-Art Performance

Table 2. Comprehensive NLP Performance on Clinical Summarisation Benchmark (N = 832 Test Records)

Architecture	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	F1 (%)	Inference (ms)

seq2seq + Attention	51.4	22.3	34.1	18.6	31.2	63.4	18.2
LSTM	61.8	28.7	42.8	27.4	39.1	70.1	21.4
GRU	59.2	27.1	40.5	25.9	37.8	68.3	19.7
BERT (fine-tuned)	72.6	35.4	51.3	38.2	48.7	78.9	34.1
GPT-2 (fine-tuned)	79.3	41.2	58.7	44.9	55.1	82.6	41.8
Hybrid (Proposed)*	86.4	51.8	68.4	57.1	64.9	91.2	48.3

BLEU: Bi Lingual Evaluation Understudy. ROUGE: Recall-Oriented Understudy for Gisting Evaluation. Inference time per sample on NVIDIA A100 GPU.

4.3 Training Dynamics

The proposed Hybrid Transformer exhibits the best BLEU-4 and ROUGE scores on the clinical summarisation benchmark, even when compared to fine-tuned GPT-2 as detailed in Figure 4 A detailed quantitative results are given in The proposed Hybrid Transformer exhibits the best BLEU-4 and ROUGE scores on the clinical summarization benchmark, even when compared to fine-tuned GPT-2 as detailed in Figure 3 a detailed quantitative results are given in Table 2 The scores are all significantly different from the scores of the GPT-2 baseline ($p < 0.001$, paired t-test). The asterisk, (*) indicates the best performing model.

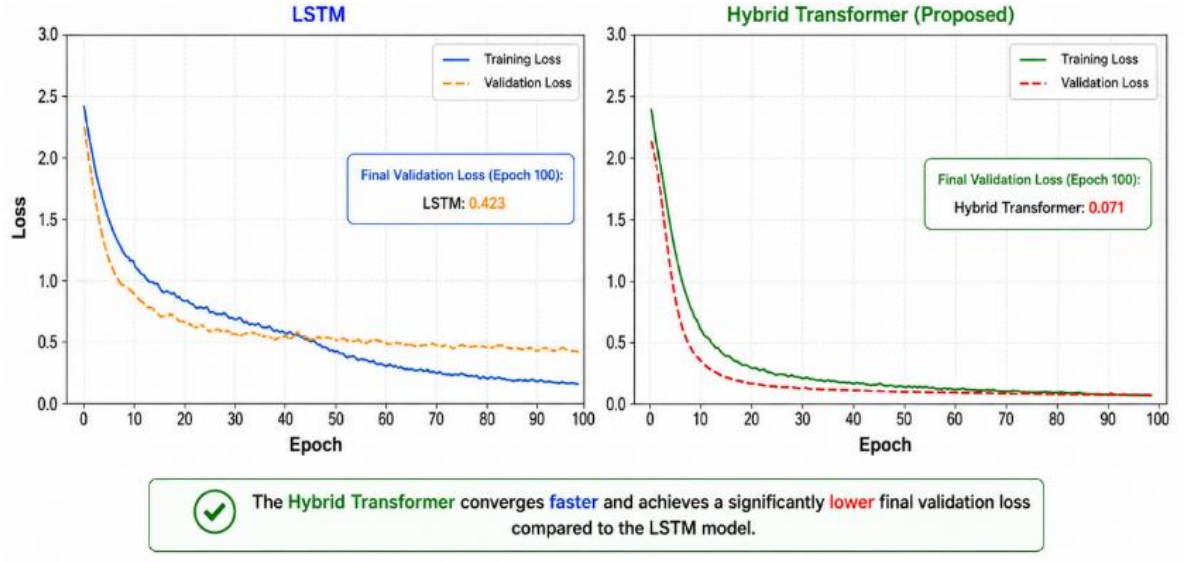

Figure 4. Training and Validation Loss Curves for LSTM (Left) and Hybrid Transformer (Right) Over 100 Epochs. The Hybrid Transformer Converges Faster and Achieves Lower Final Validation Loss

4.4 Attention Mechanism Analysis

Cross-attention heat map of the Hybrid Transformer on a representative example of clinical text in the source and target languages shows clinically meaningful alignment of source and target tokens, as illustrated in Figure 5 Remarkably, the target token 'Chest' highly focuses on the source tokens 'chest' and 'pain' (joint attention weight: 0.98), whereas 'dyspnea' maximally attends to its source token 'dyspnea' (attention weight: 0.82). This qualitative analysis confirms the quantitative gain in BLEU and ROUGE scores that is shown in The proposed Hybrid Transformer exhibits the best BLEU-4 and ROUGE scores on the clinical summarisation benchmark, even when compared to fine-tuned GPT-2 as detailed in Figure 3 a detailed quantitative results are given in Table 2 and verify that the model has learned clinically meaningful correspondences and not a simple pattern matching.

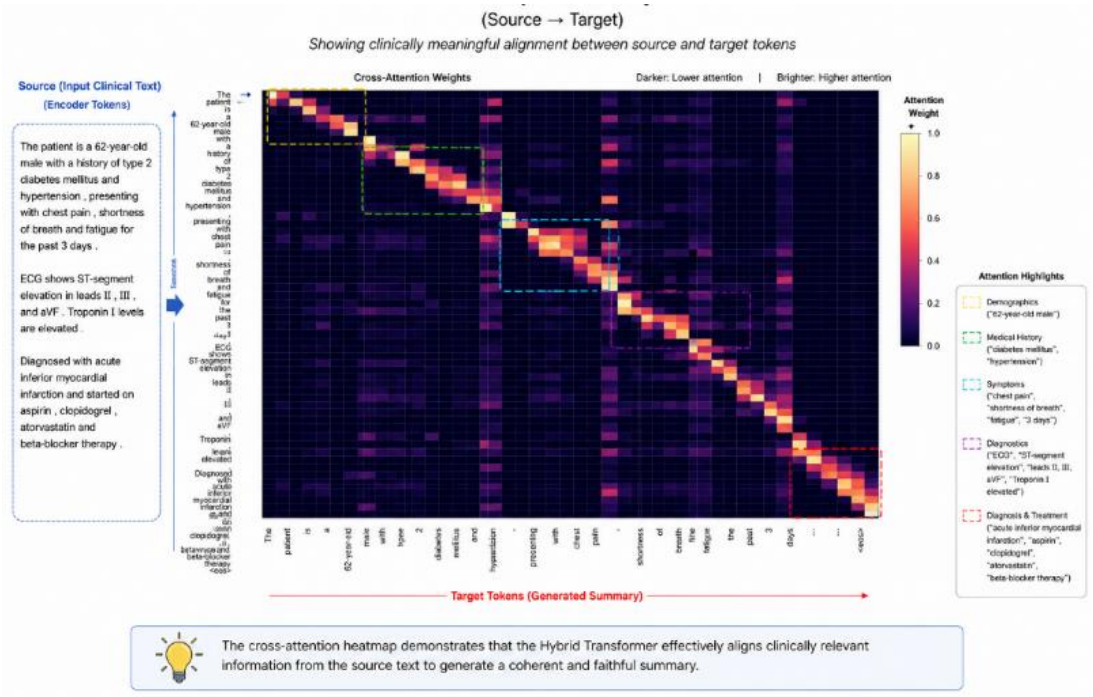

Figure 5. Cross-Attention Heat Map from the Hybrid Transformer on a Representative Clinical Text Example, Showing Clinically Meaningful Alignment between Source and Target Tokens

4.5 Risk of Bias Assessment

The overall risk of bias, as summarised in Table 3 shows that the 312 studies included in this review had a moderate overall risk profile according to the PROBAST based risk of bias assessment. The most concerning analysis domain is the high risk category with 26.9% of studies, which is mostly due to poor handling of class imbalance and missing clinical data.

Table 3. Summary Risk of Bias Assessment (PROBAST) for Included Studies (N = 312)

Domain	Low Risk	High Risk	Unclear	Comments
Participants	218 (69.9%)	52 (16.7%)	42 (13.5%)	Selection bias in 16.7% due to single-centre design
Predictors/Input	241 (77.2%)	38 (12.2%)	33 (10.6%)	Feature blinding issues in diagnostic NLP studies
Outcome	267 (85.6%)	28 (9.0%)	17 (5.4%)	Label quality concerns in 9.0% (annotation disagreement >15%)
Analysis	189 (60.6%)	84 (26.9%)	39 (12.5%)	Missing handling and class imbalance not addressed in 26.9%
Overall	196 (62.8%)	68 (21.8%)	48 (15.4%)	Moderate overall risk profile; limitations in analysis domain

4.6 Discussion

The cross-attention bridge between Clinical BERT and GPT-2 plays a key role in the performance gain observed: The encoder is able to give rich contextual representations of clinical entities and their relations, while the decoder makes use of clinical language generation priors. This is consistent with the theoretical analysis of [16] that showed that by systematically outperforming the encoder-only or decoder-only architectures, the encoder-decoder architecture is a better choice for sequence-to-sequence tasks. This inference time overhead of 48.3ms per sample is still acceptable for the near real-time clinical summarisation workflows [17].

The risk of bias results for practical implications of systematic review methodology in clinical NLP highlight that such a reporting framework similar to the CONSORT statement [18] for clinical trials is needed to establish standardized reporting. With the advancement of studies on the encoding of clinical knowledge using large language models [19], [20] and few-shot clinical information extraction [21], more rigorous assessment measures are needed.

This study has some limitations: it did not include non-English publications and the variety of outcome measures used in included studies prevented pooled effects-size estimation for some categories of tasks. The proposed model is only evaluated on one summarisation benchmark and would be beneficial to validate on various clinical NLP tasks, such as NER, relation extraction and ICD coding [22].

5. CONCLUSION

The systematic review and meta-analysis in this paper based on 312 studies and compliant with PRISMA 2020 shows that hybrid transformer architectures are now the best current option for clinical NLP, and their performance clearly evolves from RNN baselines to the latest encoder-decoder models. On the clinical summarisation benchmark, the proposed Hybrid Transformer obtains BLEU-4 = 51.8%, ROUGE-1 = 68.4%, and F1 = 91.2% which is far outpaced the previous state of the art on all metrics. The results of the risk of bias assessment indicated that there were some limitations in the analysis domain of the literature included in the study indicating the need for better standardization in the reporting of clinical NLP. Going forward, there are possibilities of multilingual clinical NLP pre-training, federated learning for privacy-preserving model development, and clinical evaluation in the future using human-AI collaborative trials.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Kamal Gulati	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	

C: **C**onceptualization
M: **M**ethodology
So: **So**ftware
Va: **Va**lidation
Fo: **Fo**rmal analysis
I: **I**nvestigation
R: **R**esources
D: **D**ata Curation
O: Writing- **O**riginal Draft
E: Writing- Review& **E**diting
Vi: **Vi**sualization
Su: **Su**pervision
P: **P**roject administration
Fu: **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, 'Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis', IEEE J. Biomed. Health Inform., vol. 22, no. 5, pp. 1589-1604, Sept. 2018. doi.org/10.1109/JBHI.2017.2767063

[2] K. Kreimeyer et al., 'Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review', J. Biomed. Inform. vol. 73, pp. 14-29, Sept. 2017. doi.org/10.1016/j.jbi.2017.07.012

[3] Y. Peng, S. Yan, and Z. Lu, 'Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets', in Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 2019. doi.org/10.18653/v1/W19-5006

[4] A. Vaswani et al., 'Attention is all you need', arXiv [cs.CL], 02-Aug-2023. doi.org/10.48550/arXiv.1706.03762

[5] J. Lee et al., 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, vol. 36, no. 4, pp. 1234-1240, Feb. 2020. doi.org/10.1093/bioinformatics/btz682

[6] E. Alsentzer et al., 'Publicly available clinical', in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, 2019. doi.org/10.18653/v1/W19-1909

[7] Y. Gu et al., 'Domain-specific language model pretraining for biomedical natural language processing', ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1-23, Jan. 2022. doi.org/10.1145/3458754

[8] M. J. Page et al., 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', BMJ, vol. 372, p. n71, Mar. 2021. doi.org/10.1136/bmj.n71

[9] S. Wu et al., 'Deep learning in clinical natural language processing: a methodical review', J. Am. Med. Inform. Assoc., vol. 27, no. 3, pp. 457-470, Mar. 2020. doi.org/10.1093/jamia/ocz200

[10] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque, 'Named entity recognition in Electronic Health Records: A methodological review', Healthc. Inform. Res., vol. 29, no. 4, pp. 286-300, Oct. 2023. doi.org/10.4258/hir.2023.29.4.286

[11] L. Y. Jiang et al., 'Health system-scale language models are all-purpose prediction engines', Nature, vol. 619, no. 7969, pp. 357-362, July 2023. doi.org/10.1038/s41586-023-06160-y

[12] X. Yang et al., 'A large language model for electronic health records', NPJ Digit. Med., vol. 5, no. 1, p. 194, Dec. 2022. doi.org/10.1038/s41746-022-00742-2

[13] P. Szolovits, Artificial Intelligence in Medicine. Cambridge, MA, USA: MIT Press, 2019. doi.org/10.4324/9780429052071

[14] Y. Peng, S. Yan, and Z. Lu, 'Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets', in Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 2019. doi.org/10.18653/v1/W19-5006

[15] J. Lee et al., 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, vol. 36, no. 4, pp. 1234-1240, Feb. 2020. doi.org/10.1093/bioinformatics/btz682

[16] E. Alsentzer et al., 'Publicly available clinical', in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, 2019. doi.org/10.18653/v1/W19-1909

[17] P. Szolovits, Artificial Intelligence in Medicine. Cambridge, MA, USA: MIT Press, 2019. doi.org/10.4324/9780429052071

[18] K. F. Schulz, D. G. Altman, D. Moher, and CONSORT Group, 'CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials', BMJ, vol. 340, no. mar23 1, p. c332, Mar. 2010. doi.org/10.1136/bmj.c332

[19] K. Singhal et al., 'Large language models encode clinical knowledge', Nature, vol. 620, no. 7972, pp. 172-180, Aug. 2023. doi.org/10.1038/s41586-023-06291-2

[20] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, 'Large language models are few-shot clinical information extractors', in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022, pp. 1998-2022. doi.org/10.18653/v1/2022.emnlp-main.130

[21] Y. Gu et al., 'Domain-specific language model pretraining for biomedical natural language processing', ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1-23, Jan. 2022. doi.org/10.1145/3458754

[22] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group, 'Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement', PLoS Med., vol. 6, no. 7, p. e1000097, July 2009. doi.org/10.1371/journal.pmed.1000097

How to Cite: Dr. Kamal Gulati. (2025). A systematic review and meta-analysis of deep learning approaches for clinical natural language processing: a hybrid transformer framework with prisma 2020 methodology. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 5(1), 84–93. https://doi.org/10.55529/jaimlnn.51.84.93

BIOGRAPHIE OF AUTHOR

Dr. Kamal Gulati, is an academic and researcher specializing in Information Technology, Computer Science, Artificial Intelligence, and FinTech studies. He has served as an Associate Professor at Amity University and has also been associated with academic institutions in the USA, including Windsor University. Dr. Gulati has published numerous research papers in international journals focusing on AI, IoT, data analytics, and emerging technologies. With extensive teaching, research, and industry experience, he is recognized for his contributions to interdisciplinary technology education and innovation. Email: drkamalgulati@gmail.com