Received: 11 October 2024 | Accepted: 4 December 2024

DOI: 10.1002/ird3.115

COMMENTARY

iRADIOLOGY

Natural language processing for chest X-ray reports in the transformer era: BERT-like encoders for comprehension and GPT-like decoders for generation

Han Yuan

Duke-NUS Medical School, National University of Singapore, Singapore

Correspondence
Han Yuan, Duke-NUS Medical School, National University of Singapore, 8 College Rd, 169857, Singapore.
Email: yuan.han@u.duke.nus.edu

KEYWORDS
bidirectional encoder representations from transformers, chest X-ray report, generative pre-trained transformer, large language model, natural language processing

1 | TRANSFORMER FOR CHEST X-RAY REPORT ANALYSIS

Natural language processing (NLP) has gained widespread use in computer-assisted chest X-ray (CXR) report analysis, particularly since the renaissance of deep learning (DL) in the 2012 ImageNet challenge. While early endeavors predominantly employed recurrent neural networks (RNN) and convolutional neural networks (CNN) [1], the revolution is brought by the transformer [2] and its success can be attributed to three key factors [3]. First, its self-attention mechanism enables simultaneous processing of multiple parts of an input sequence, offering significantly greater efficiency compared to earlier models such as RNN [4]. Second, its architecture exhibits exceptional scalability, supporting models with over 100 billion parameters to capture intricate linguistic relationships in human language [5]. Third, the availability of vast internet-based corpus and advances in computational power have made the pre-training and fine-tuning of large-scale transformer-based models feasible [6]. The development of the transformer enables the resolution of previously intractable problems and achieves expert-level performance across a broad range of CXR report analytical tasks, such as name entity recognition, question answering, and extractive summarization [7]. In this commentary, we conducted a comprehensive literature search in PubMed (Figure 1) to illustrate the current landscape, adoption barriers, and potential solutions for the transformer-based tools from the perspective of the transformer's two integral components: encoder handling comprehension and decoder managing generation. As our primary focus is NLP, the classification criteria for encoder or decoder was based on text modules and we excluded research purely focusing on vision transformers (ViT).

Abbreviations: BERT, bidirectional encoder representations from transformers; CNN, convolutional neural networks; DL, deep learning; GPT, generative pre-trained transformer; LSTM, long short-term memory; NLP, natural language processing; RNN, recurrent neural networks; ViT, vision transformers.

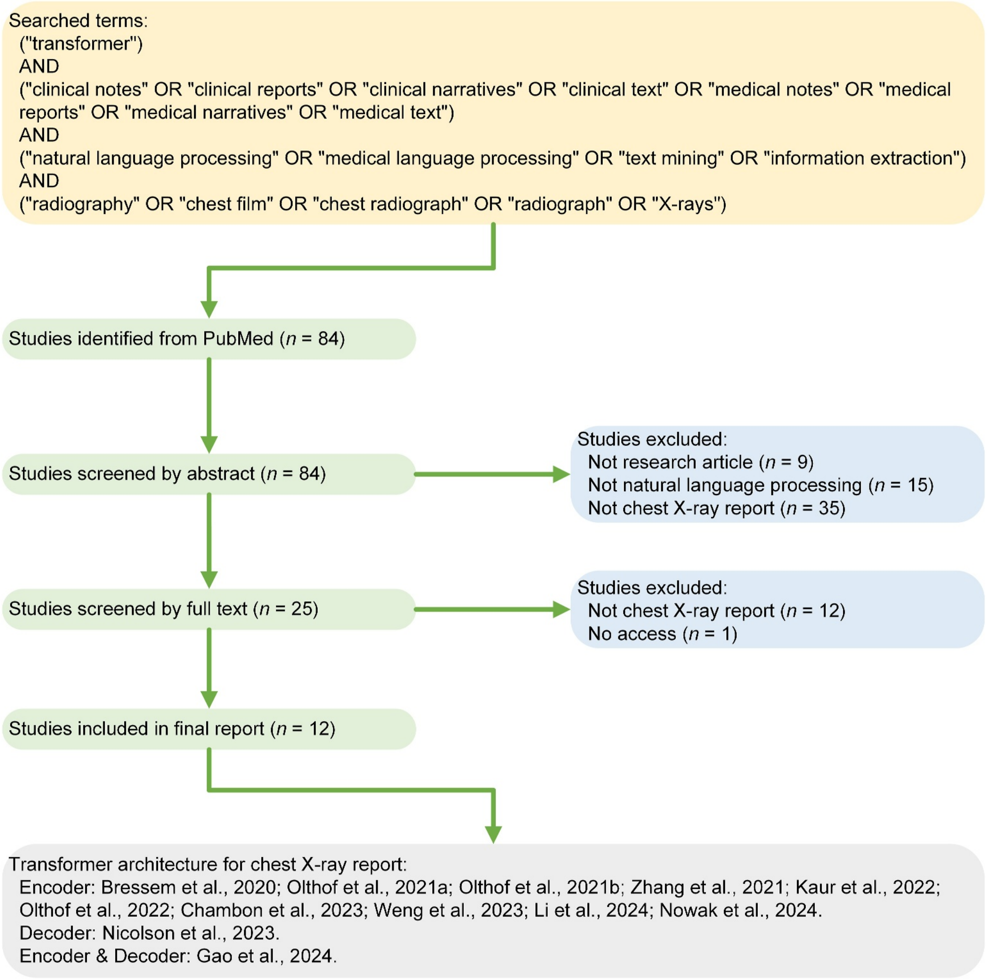
Searched terms:
("transformer")
AND
("clinical notes" OR "clinical reports" OR "clinical narratives" OR "clinical text" OR "medical notes" OR "medical reports" OR "medical narratives" OR "medical text")
AND
("natural language processing" OR "medical language processing" OR "text mining" OR "information extraction")
AND
("radiography" OR "chest film" OR "chest radiograph" OR "radiograph" OR "X-rays")
Studies identified from PubMed (n = 84)
Studies screened by abstract (n = 84)
Studies excluded:
Not research article (n = 9)
Not natural language processing (n = 15)
Not chest X-ray report (n = 35)
Studies screened by full text (n = 25)
Studies excluded:
Not chest X-ray report (n = 12)
No access (n = 1)
Studies included in final report (n = 12)
Transformer architecture for chest X-ray report:
Encoder: Bressem et al., 2020; Olthof et al., 2021a; Olthof et al., 2021b; Zhang et al., 2021; Kaur et al., 2022; Olthof et al., 2022; Chambon et al., 2023; Weng et al., 2023; Li et al., 2024; Nowak et al., 2024.
Decoder: Nicolson et al., 2023.
Encoder & Decoder: Gao et al., 2024.

FIGURE 1 Literature search pipeline in PubMed to identify relevant articles published from June 12, 2017, when the transformer model was first introduced, to October 4, 2024. We followed previous systematic reviews [3, 8, 9] to design the groups of keywords: (1) "transformer"; (2) "clinical notes", "clinical reports", "clinical narratives", "clinical text", "medical notes", "medical reports", "medical narratives", or "medical text"; (3) "natural language processing", "medical language processing", "text mining", or "information extraction"; (4) "radiography", "chest film", "chest radiograph", "radiograph", or "X-rays".

2 | BERT-LIKE ENCODERS FOR COMPREHENSION

As the primary means of communication between radiologists and referring physicians, CXR reports contain high-density information on patients' conditions [10]. Much like physicians interpreting CXR reports, the first step of NLP analysis is understanding the content and an important application of the transformer encoder is explicitly converting it into a format suitable for subsequent tasks. One notable encoder is BERT [11], which stands for bidirectional encoder representations from transformers. In contrast to predecessors that rely on large amounts of expert annotations for supervised learning [12], BERT undergoes self-supervised training on large-scale unlabeled datasets to understand language patterns and is subsequently fine-tuned with a small set of annotations on the target task [12, 13], yielding superior performance in text classification [14], name entity recognition [15], extractive summarization [16], and semantics optimization [17]. In the context of healthcare, Olthof et al. [18] built a comprehensive pipeline to evaluate BERT across datasets of varying complexities, disease prevalence, and sample sizes, demonstrating that BERT statistically outperformed conventional DL models of RNN and CNN, in terms of area under the curve and F_1-score, with t-test p-values less than 0.05. Beyond the superior performance of BERT compared to conventional DL models, adapting it to domain-specific corpus can further enhance the effectiveness across various tasks. Yan et al. [19] adapted four BERT-like encoders using millions of radiology reports to tackle three key tasks: identifying sentences that describe abnormal findings, assigning diagnostic codes, and extracting key sentences that summarize the reports. Their results demonstrated that domain-specific adaptation yielded statistically significant improvements in accuracy, F_1-score, and ROUGE metrics across all three tasks.

Most BERT-relevant studies predominantly focus on sentence-, paragraph-, or report-level predictions, while BERT-like encoders are also well-suited for word-level pattern recognition. Chambon et al. [20] leveraged PubMed BERT [21], a biomedical-specific adaptation of BERT, to evaluate the probability of individual tokens containing protected health information, and replaced identified sensitive tokens with synthetic surrogates to ensure privacy preservation. Similarly, Weng et al. [22] developed a system utilizing ALBERT [23], a lite BERT with reduced parameters, to identify diagnostic keywords unrelated to abnormal findings, thereby reducing false-positive alarms and outperforming regular expression-, syntactic grammar-, and conventional DL-based baselines.

BERT-derived labels can also be applied to develop DL models targeting other modalities [12, 13]. Nowak et al. [24] systematically explored the utility of BERT-generated silver labels for CXR reports and subsequently linked them to the corresponding radiographs to develop image classifiers. Compared to models trained exclusively on radiologist-annotated gold labels, integrating silver and gold labels led to improved discriminability. In a further macro-averaged analysis, synchronous training on silver and gold labels proved effective in settings with limited gold labels, whereas training first with silver, followed by gold labels was better in cases with abundant gold labels. Zhang et al. [25] introduced a novel approach to extracting more generalizable labels from CXR reports for image classifiers, rather than relying on predefined categories: first, they used BERT to extract linguistic entities and relationships; second, they constructed a knowledge graph based on these extractions; third, radiologists refined the graph using their domain expertise. Unlike traditional multiclass labels, the established knowledge graph not only categorized each sample but also revealed interpretable relationships between categories, such as those linking anatomical regions with abnormal signs. In addition to deriving classification labels, BERT and its advanced comprehension capabilities introduced an unprecedented innovation: the direct supervision of pixel-level segmentation models using medical text [26]. Li et al. [26] proposed a text-augmented lesion segmentation paradigm that integrated BERT-based textual information to compensate for the deficiency in chest radiograph quality and refine pseudo annotations for semi-supervision. These studies highlight the strength of BERT-like encoders in comprehending healthcare-related content and their potential to enhance annotation systems for multi-modality beyond text. Meanwhile, researchers have identified the failures of BERT models in handling complex clinical tasks. Sushil et al. [27] demonstrated that BERT implementations for clinical language inference achieved a test accuracy of 0.778. While domain-specific adaptations using medical textbooks or PubMed articles improved accuracy to 0.833, this performance still fell short of that achieved by medical experts. Potential limitations of BERT-like encoders lie in their relatively modest parameter size, although larger than earlier DL models, and their reliance on inadequate training corpora, such as books, Wikipedia, and selected text databases [28]. Consequently, their ability to learn human knowledge remains constrained. These shortcomings are being alleviated by GPT-like decoders, which incorporate hundreds of billions of parameters and are trained on an internet-scale corpora [29].

3 | GPT-LIKE DECODERS FOR GENERATION

Following the advent of BERT-like encoders, generative pre-trained transformer (GPT) [30], the next groundbreaking leap, breaks technical barriers by enabling non-experts to perform NLP tasks through a freely conversational format without any coding. CvT2DistilGPT2 [31], a prominent report generator in the transformer era, utilizes a convolutional ViT as the image encoder and GPT-2 as the text decoder. Their comprehensive experiments indicated that ViT outperformed CNN and GPT surpassed BERT in encoder–decoder architectures for CXR report generation. In specific generation applications, state-of-the-art methods integrate BERT-like encoders with GPT-like decoders. TranSQ [32] is such an advanced framework. Compared with earlier models, it emulates the diagnostic reasoning process of radiologists when generating reports: (1) formulating diagnostic hypothesis embeddings that represent implicit clinical intentions, (2) querying relevant visual features extracted by a ViT and synthesizing semantic embeddings through the cross-modality fusion, and (3) transforming the semantic embeddings into candidate sentences based on DistilGPT [33]. Finally, TranSQ attained a BLEU-4 score of 0.205 and a ROUGE score of 0.409. In comparison, the best-performing baseline among 17 retrieval and generation models achieved a BLEU-4 score of 0.188 and a ROUGE score of 0.383, highlighting the superior capability of the unified transformer architecture in multi-modality.

Though GPT-like decoders have dominated text generation in the general domain, the RNN family such as long short-term memory (LSTM) [34] still achieves good performance in generating medical reports, partially because of highly templated characteristics in the clinical text [32]. Kaur and Mittal [35] employed classical encoder–decoder architectures, utilizing CNN for visual feature extraction, and LSTM for textual token generation. They also integrated transformer modules, not GPT-like decoders but BERT-like encoders, to generate numerical representations as LSTM inputs prior to report generation and to shortlist disease-relevant sentences afterward. Results presented that their proposed solution achieved a BLEU-4 score of 0.767 and a ROUGE score of 0.897, suggesting that conventional approaches remain a viable candidate backbone for CXR report generation in specific scenarios. In addition to quantitative metrics by comparing GPT outputs with ground truth reports, model-generated reports should be supplemented with evaluation by medical experts. Boag et al. [36] conducted a broad study on automated CXR report generation, highlighting a divergence between quantitative metrics and clinical accuracy. A discrepancy between quantitative metrics and report readability has also been reported [37]. Accordingly, we emphasize the involvement of human rating in the evaluation of CXR report generation to ensure clinical correctness and readability.

4 | ADOPTION BARRIERS AND POTENTIAL SOLUTIONS

In previous sections, we reviewed the current applications of transformer for various CXR report analytical tasks. Although the remarkable performance of BERT-like encoders and GPT-like decoders has been well-established, these applications still face domain-specific problems. Some of these can be alleviated through the integration of advanced technical methods and specialized medical expertise [31, 38], while others necessitate further research for resolution.

4.1 | Computational burdens

First, the computational demands in the transformer era are substantial. For example, the large version of BERT contains 334 million parameters and GPT-3 has 175 billion. In contrast, traditional DL models, such as support vector machines [39] and random forests [40], require only a few hundred to a few thousand parameters. As a result, many healthcare providers cannot afford the computational costs of tailoring models from scratch. To address this, we offer several recommendations. For model development, we suggest researchers leverage pre-trained open-access models and focus on fine-tuning rather than building models from scratch. For fine-tuning, considering the varying parameter scales, we recommend parameter-efficient fine-tuning for BERT-like encoders, a technique that updates only a small subset of the model's parameters while leaving the majority of pre-trained weights unchanged [41]. An exemplificative study conducted by Taylor et al. [42] empirically validated the effectiveness of various parameter-efficient fine-tuning techniques on BERT-like encoders within the healthcare domain. For GPT-like decoders, we advocate prompt engineering techniques, such as retrieval-augmented generation, which emphasize crafting informative and instructive inputs to guide the decoders' output without changing model parameters [43]. For example, Ranjit et al. [44] proposed a method to retrieve the most relevant sentences from prior CXR reports as contextual prompts for GPT-like decoders, enabling the generation of concise and accurate reports retaining critical clinical entities. Last but not least, obtaining approval from ethics committees to share anonymous data can facilitate collaboration with external technical partners, helping to alleviate resource burdens.

4.2 | Interpretability concerns

Second, the interpretability of transformer models, including both BERT-like encoders and GPT-like decoders, is critical in healthcare applications, where decisions directly impact patients' lives. While traditional DL approaches have often been regarded as black-box models, their relatively few parameters and simple architectures render them more explainable compared to modern transformers with over 100 billion parameters. For example, individual layers and neurons in CNN can be dissected and visualized, providing insights into their functionality [45–48]. In contrast, understanding the behavior of neurons in transformer models remains a significant challenge due to the computational complexity associated with the exponential scaling of neuron numbers [49]. For BERT-like encoders, though the internal neuron activations remain challenging to interpret, preliminary experiments focusing on identifying key tokens and analyzing their influence on the model's outputs have demonstrated a high degree of alignment with medical expert assessments [50, 51]. For GPT-like decoders, a key strength lies in their flexibility to generate content and align with human instructions. This capability allows users to not only obtain expected outputs for predefined tasks but also request explanations for these outputs, fostering enhanced interpretability and usability [52, 53]. For readers seeking a more comprehensive overview of techniques or detailed insights, we recommend referring to these surveys [54–56].

4.3 | Ethical issues

Third, ethical considerations are paramount in the era of transformers, given their powerful ability to extract nuanced patterns from training datasets. These concerns are particularly pressing when datasets contain sensitive private information or are not representative of the target population. To address patient privacy, we recommend anonymizing input data during both model development and deployment stages to ensure that sensitive information is neither learned by the model [57] nor inadvertently disclosed under certain prompts [58]. Dataset representativeness is also a critical issue, as underrepresentation of minority groups in training data can exacerbate performance disparities and perpetuate inequities [59]. To mitigate this risk, developers should prioritize inclusivity during data collection, and maintainers should continuously monitor model performance to ensure equitable outcomes [60].

4.4 | Hallucination problems

Fourth, although GPT-like decoders have demonstrated remarkable capability in generating coherent responses to diverse user prompts and solving a wide range of tasks in a conversational format [61], they are developed on the predictive probability of tokens from the internet corpora instead of contextual radiological language and well-defined logic [62]. Therefore, they continue to suffer from hallucinations, a phenomenon where model-generated content appears coherent and plausible but is factually incorrect, nonsensical, or unrelated to users' inputs [63]. Current efforts to reducing hallucination can be broadly categorized into methods applied during training and post-training stages. During training, key strategies include supervised fine-tuning on in-house CXR reports and reinforcement learning guided by radiologists' feedback [31, 64]. Post-training methods encompass hallucination detection, integration of external knowledge, multi-agent collaboration, and radiologist-in-the-loop frameworks [62, 65]. Due to space constraints, we encourage readers to refer to these reviews [62, 66–68] for comprehension of these strategies.

4.5 | Malpractice and legal liabilities

Lastly, even after these technical refinements, transformer may still present risks of malpractice, potentially leading to medical errors and legal liabilities [69]. Errors can arise from various sources, including inaccurate transformer outputs, clinician nonadherence to correct transformer recommendations, and poor integration of the transformer into clinical workflows [70]. Consequently, determining legal responsibility in cases of adverse outcomes remains a critical issue for various stakeholders, including software developers, maintenance teams, radiology departments, and radiologists [71]. A report by the European Commission focuses on the safety and liability implications of artificial intelligence, which applies medical device laws to DL models, and demonstrates that liability generally falls into two categories: civil and product liability [71]. Civil liability typically pertains to radiologists and radiology departments, while product liability applies to software developers. However, the report stops short of providing a strict and definitive framework for liability due to the inherent complexity and ambiguity of DL algorithms [71]. As a result, legal questions surrounding liability will likely continue to be addressed through courts and case law. Under existing legal frameworks, we recommend radiologists to follow the standard of care, utilizing DL models as supplementary confirmatory tools rather than substitutes for standard medical practice to ensure beneficial outcomes for all stakeholders [69]. Additionally, for radiology departments seeking to implement transformer-based NLP tools, we suggest that they should involve radiologists, the most important stakeholders, throughout the entire development cycle [72], and prepare in-depth training programs to familiarize radiologists with transformer-based tools, which differ significantly from routine statistical tests and are often black boxes that resist full interpretation [73]. Moreover, managing radiologists' expectations is important: both unrealistic optimism, where transformer is seen as a replacement for expert expertise, and undue pessimism, where transformer is perceived as offering no utility, should be avoided [74–77].

AUTHOR CONTRIBUTIONS

Han Yuan: Conceptualization; data curation; formal analysis; investigation; project administration; validation; visualization; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The author declares that he has no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing does not apply to this study as no datasets were generated or analyzed.

ETHICS STATEMENT

This study is exempt from review by the ethics committee because it does not involve human participants, animal subjects, or sensitive data collection.

INFORMED CONSENT

Not applicable.

ORCID

Han Yuan https://orcid.org/0000-0002-2674-6068

REFERENCES

[1] Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med. 2019;97:79–88. https://doi.org/10.1016/j.artmed.2018.11.004

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Proceedings of the advances in neural information processing systems; 2017. p. 6000–10. https://doi.org/10.48550/arXiv.1706.03762

[3] Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language processing of clinical notes on chronic diseases: systematic review. JMIR Medical Informatics. 2019;7(2):e12239. https://doi.org/10.2196/12239

[4] Brauwers G, Frasincar F. A general survey on attention mechanisms in deep learning. IEEE Trans Knowl Data Eng. 2023;35(4):3279–98. https://doi.org/10.1109/TKDE.2021.3126456

[5] Wang H, Ma S, Dong L, Huang S, Zhang D, Wei F. DeepNet: scaling transformers to 1, 000 layers. IEEE Trans Pattern Anal Mach Intell. 2024;46(10):6761–74. https://doi.org/10.1109/TPAMI.2024.3386927

[6] Zeng G, Yang W, Ju Z, Yang Y, Wang S, Zhang R, et al. MedDialog: large-scale medical dialogue datasets. In: Proceedings of the conference on empirical methods in natural language processing; 2020. p. 9241–50. https://doi.org/10.18653/v1/2020.emnlp-main.743

[7] Li I, Pan J, Goldwasser J, Verma N, Wong W, Nuzumlalı M, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev. 2022;46:100511. https://doi.org/10.1016/j.cosrev.2022.100511

[8] Li Y, Zhang Z, Dai C, Dong Q, Badrigilan S. Accuracy of deep learning for automated detection of pneumonia using chest X-ray images: a systematic review and meta-analysis. Comput Biol Med. 2020;123:103898. https://doi.org/10.1016/j.compbiomed.2020.103898

[9] Siebra CA, Kurpicz-Briki M, Wac K. Transformers in health: a systematic review on architectures for longitudinal data analysis. Artif Intell Rev. 2024;57(2):32. https://doi.org/10.1007/s10462-023-10677-z

[10] Olthof AW, van Ooijen PM, Cornelissen LJ. The natural language processing of radiology requests and reports of chest imaging: comparing five transformer models' multilabel classification and a proof-of-concept study. Health Inf J. 2022;28(4):1–26. https://doi.org/10.1177/14604582221131198

[11] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the. Association for Computational Linguistics; 2018. p. 4171–86.

[12] Olthof AW, Shouche P, Fennema EM, Ijpma FFA, Koolstra RHC, Stirler VMA, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Progr Biomed. 2021;208:106304. https://doi.org/10.1016/j.cmpb.2021.106304

[13] Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics. 2021;36(21):5255–61. https://doi.org/10.1093/bioinformatics/btaa668

[14] Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inf Assoc. 2019;26(12):1632–6. https://doi.org/10.1093/jamia/ocz164

[15] Li X, Zhang H, Zhou X-H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inf. 2020;107:103422. https://doi.org/10.1016/j.jbi.2020.103422

[16] Wang Y, Zhang J, Yang Z, Wang B, Jin J, Liu Y. Improving extractive summarization with semantic enhancement through topic-injection based BERT model. Inf Process Manag. 2024;61(3):103677. https://doi.org/10.1016/j.ipm.2024.103677

[17] Yang N, Jo J, Jeon M, Kim W, Kang J. Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models. Expert Syst Appl. 2022;190:116209. https://doi.org/10.1016/j.eswa.2021.116209

[18] Olthof AW, van Ooijen PA, Cornelissen LJ. Deep learning-based natural language processing in radiology: the impact of report complexity, disease prevalence, dataset size, and algorithm type on model performance. J Med Syst. 2021;45(10):91. https://doi.org/10.1007/s10916-021-01761-4

[19] Yan A, McAuley J, Lu X, Du J, Chang E, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4(4):e210258. https://doi.org/10.1148/ryai.210258

[20] Chambon PJ, Wu C, Steinkamp JM, Adleberg J, Cook TS, Langlotz CP. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. J Am Med Inf Assoc. 2023;30(2):318–28. https://doi.org/10.1093/jamia/ocac219

[21] Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns. 2023;4(4):100729. https://doi.org/10.1016/j.patter.2023.100729

[22] Weng KH, Liu CF, Chen CJ. Deep learning approach for negation and speculation detection for automated important finding flagging and extraction in radiology report: internal validation and technique comparison study. JMIR Med Inform. 2023;11:e46348. https://doi.org/10.2196/46348

[23] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricu R. ALBERT: a lite BERT for self-supervised learning of language

representations. In: Proceedings of the international conference on learning representations; 2020. p. 1–17.

[24] Nowak S, Schneider H, Layer YC, Theis M, Biesner D, Block W, et al. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. Eur Radiol. 2024;34(5):2895–904. https://doi.org/10.1007/s00330-023-10373-0

[25] Zhang Y, Liu M, Hu S, Shen Y, Lan J, Jiang B, et al. Development and multicenter validation of chest X-ray radiography interpretations based on natural language processing. Commun Med. 2021;1:43. https://doi.org/10.1038/s43856-021-00043-x

[26] Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, et al. LViT: language meets vision transformer in medical image segmentation. IEEE Trans Med Imag. 2024;43(1):96–107. https://doi.org/10.1109/TMI.2023.3291719

[27] Sushil M, Suster S, Daelemans W. Are we there yet? Exploring clinical domain knowledge of BERT models. In: Proceedings of the 20th workshop on biomedical language processing; 2021. p. 41–53. https://doi.org/10.18653/v1/2021.bionlp-1.5

[28] Turchin A, Masharsky S, Zitnik M. Comparison of BERT implementations for natural language processing of narrative medical documents. Inform Med Unlocked. 2023;36:101139. https://doi.org/10.1016/j.imu.2022.101139

[29] Le Mens G, Kovács B, Hannan MT, Pros G. Uncovering the semantics of concepts using GPT-4. Proc Natl Acad Sci USA. 2023;120(49):e2309350120. https://doi.org/10.1073/pnas.2309350120

[30] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the international conference on neural information processing systems; 2020. p. 1877–901. https://doi.org/10.5555/3495724.3495883

[31] Nicolson A, Dowling J, Koopman B. Improving chest X-ray report generation by leveraging warm starting. Artif Intell Med. 2023;144:102633. https://doi.org/10.1016/j.artmed.2023.102633

[32] Gao D, Kong M, Zhao Y, Huang J, Huang Z, Kuang K, et al. Simulating doctors' thinking logic for chest X-ray report generation *via* transformer-based semantic query learning. Med Image Anal. 2024;91:102982. https://doi.org/10.1016/j.media.2023.102982

[33] Victor S, Lysandre D, Julien C, Thomas W. DistilBERT, A distilled version of BERT: smaller, faster, cheaper and lighter. In: Proceedings of the international workshop on energy efficient machine learning and cognitive computing; 2019. p. 1–5.

[34] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[35] Kaur N, Mittal A. RadioBERT: a deep learning-based system for medical report generation from chest X-ray images using contextual embeddings. J Biomed Inf. 2022;135:104220. https://doi.org/10.1016/j.jbi.2022.104220

[36] Boag W, Hsu T-MH, Mcdermott M, Berner G, Alesentzer E, Szolovits P. Baselines for chest X-ray report generation. In: Proceedings of the machine learning for healthcare conference; 2020.

[37] Liu G, Hsu TH, McDermott M, Boag W, Weng W, Szolovits P, et al. Clinically accurate chest X-ray report generation. In: Proceedings of the machine learning for healthcare conference; 2019.

[38] Yuan H, Kang L, Li Y, Fan Z. Human-in-the-Loop machine learning for healthcare: current progress and future opportunities in electronic health records. Med Adv. 2024;2(3):318–22. https://doi.org/10.1002/med4.70

[39] Mitra V, Wang CJ, Banerjee S. Text classification: a least square support vector machine approach. Appl Soft Comput. 2007;7(3):908–14. https://doi.org/10.1016/j.asoc.2006.04.002

[40] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests. Pattern Recogn. 2011;44(2):330–49. https://doi.org/10.1016/j.patcog.2010.08.011

[41] Lu Q, Dou D, Nguyen TH. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In: Proceedings of the conference on empirical methods in natural language processing; 2021. p. 3855–65. https://doi.org/10.18653/v1/2021.findings-emnlp.325

[42] Taylor N, Ghose U, Rohanian O, Nouriborjiet M, Kormilitzin A, Clifton DA, et al. Efficiency at scale: investigating the performance of diminutive language models in clinical tasks. Artif Intell Med. 2024;157:103002. https://doi.org/10.1016/j.artmed.2024.103002

[43] Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res. 2024;26:e60501. https://doi.org/10.2196/60501

[44] Ranjit M, Ganapathy G, Manuel R, Ganu T. Retrieval augmented chest X-ray report generation using OpenAI GPT models. In: Proceedings of the machine learning for healthcare conference; 2023. p. 650–66.

[45] Ibrahim R, Shafiq MO. Explainable convolutional neural networks: a taxonomy, review, and future directions. ACM Comput Surv. 2023;55(10):206–37. https://doi.org/10.1145/3563691

[46] Yuan H, Hong C, Jiang PT, Zhao G, Tran NTA, Xu X, et al. Clinical domain knowledge-derived template improves post Hoc AI explanations in pneumothorax classification. J Biomed Inf. 2024;156:104673. https://doi.org/10.1016/j.jbi.2024.104673

[47] Yuan H. Anatomic boundary-aware explanation for convolutional neural networks in diagnostic radiology (forthcoming). iRADIOLOGY. https://doi.org/10.1002/ird3.113

[48] Yuan H, Hong C, Tran NTA, Xu X, Liu N. Leveraging anatomical constraints with uncertainty for pneumothorax segmentation. Health Care Sci. 2024;3(6):456–74. https://doi.org/10.1002/hcs2.119

[49] Sajjad H, Durrani N, Dalvi F. Neuron-level interpretation of deep NLP models: a survey. Trans Assoc Comput Linguist. 2022;10:1285–303. https://doi.org/10.1162/tacl_a_00519

[50] Lee S, Lee J, Park J, Park J, Kim D, Lee J, et al. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. Am J Emerg Med. 2024;77:29–38. https://doi.org/10.1016/j.ajem.2023.11.063

[51] Malhotra A, Jindal R. XAI transformer based approach for interpreting depressed and suicidal user behavior on online social networks. Cognit Syst Res. 2024;84:101186. https://doi.org/10.1016/j.cogsys.2023.101186

[52] Li Z, Zhang J, Zhou W, Zheng J, Xia Y. GPT-agents based on medical guidelines can improve the responsiveness and explainability of outcomes for traumatic brain injury rehabilitation. Sci Rep. 2024;14(1):7626. https://doi.org/10.1038/s41598-024-58514-9

[53] Mazumdar H, Chakraborty C, Sathvik M, Mukhopadhyay S, Panigrahi PK. GPTFX: a novel GPT-3 based framework for mental health detection and explanations. IEEE J Biomed Health Inform. 2023:1–8. https://doi.org/10.1109/JBHI.2023.3328350

[54] Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. ACM Trans Intell Syst Technol. 2024;15(2):1–38. https://doi.org/10.1145/3639372

[55] Schneider J. Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda. Artif Intell Rev. 2024;57(11):289. https://doi.org/10.1007/s10462-024-10916-x

[56] El Zini J, Awad M. On the explainability of natural language processing deep models. ACM Comput Surv. 2022;55(5):1–31. https://doi.org/10.1145/3529755

[57] Yoon J, Drumright LN, van der Schaar M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE J Biomed Health Inform. 2020;24(8):2378–88. https://doi.org/10.1109/JBHI.2020.2980262

[58] Humphreys D, Koay A, Desmond D, Mealy E. AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business. AI Ethics. 2024;4:791–804. https://doi.org/10.1007/s43681-024-00443-4

[59] Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inf. 2024;154:104646. https://doi.org/10.1016/j.jbi.2024.104646

[60] Yuan H. Toward real-world deployment of machine learning for health care: external validation, continual monitoring, and randomized clinical trials. Health Care Sci. 2024;3(5):360–4. https://doi.org/10.1002/hcs2.114

[61] Arora A, Arora A. The promise of large language models in health care. Lancet. 2023;401(10377):641. https://doi.org/10.1016/S0140-6736(23)00216-7

[62] Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMS: a review on debiasing and dehallucinating in large language models. Artif Intell Rev. 2024;57(9):243. https://doi.org/10.1007/s10462-024-10896-y

[63] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625–30. https://doi.org/10.1038/s41586-024-07421-0

[64] Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology. 2024;310(1):e232756. https://doi.org/10.1148/radiol.232756

[65] Yuan H. Clinical decision making: evolving from hypothetico-deductive model to knowledge-enhanced machine learning. Med Adv. 2024;2(4):375–9. https://doi.org/10.1002/med4.83

[66] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. https://doi.org/10.1145/3571730

[67] Guerreiro NM, Alves DM, Waldendorf J, Haddow B, Birch A, Colombo P, et al. Hallucinations in large multilingual translation models. Trans Assoc Comput Linguist. 2023;11:1500–17. https://doi.org/10.1162/tacl_a_00615

[68] Ahmadi A. Unravelling the mysteries of hallucination in large language models: strategies for precision in artificial intelligence language generation. AJCST. 2024;13(1):1–10. https://doi.org/10.70112/ajcst-2024.13.1.4144

[69] Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. J Am Med Assoc. 2019;322(18):1765–6. https://doi.org/10.1001/jama.2019.15064

[70] Mello MM, Guha N. Understanding liability risk from using health care artificial intelligence tools. N Engl J Med. 2024;390(3):271–8. https://doi.org/10.1056/NEJMhle2308901

[71] Schneeberger D, Stöger K, Holzinger A. Machine learning and knowledge extraction. The European Legal Framework for Medical AI. 2020:209–26.

[72] Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open. 2020;3(2):167–72. https://doi.org/10.1093/jamiaopen/ooz046

[73] Yuan H, Yu K, Xie F, Liu M, Sun S. Automated machine learning with interpretation: a systematic review of methodologies and applications in healthcare. Med Adv. 2024;2(3):205–37. https://doi.org/10.1002/med4.75

[74] Shuaib A, Arian H, Ali S. The increasing role of artificial intelligence in health care: will robots replace doctors in the future? Int J Gen Med. 2020;13:891–6. https://doi.org/10.2147/IJGM.S268093

[75] Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med. 2018;48:e13–4. https://doi.org/10.1016/j.ejim.2017.06.017

[76] Zhang Z. When doctors meet with AlphaGo: potential application of machine learning to clinical medicine. Ann Transl Med. 2016;4(6):1–2. https://doi.org/10.21037/atm.2016.03.25

[77] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7

How to cite this article: Yuan H. Natural language processing for chest X-ray reports in the transformer era: bidirectional encoder representations from transformers-like encoders for comprehension and generative pre-trained transformer-like decoders for generation. iRADIOLOGY. 2025;1–8. https://doi.org/10.1002/ird3.115