iScience

CellPress
OPEN ACCESS

Review

Language model and its interpretability in biomedicine: A scoping review

Daoming Lyu,[1,2] Xingbo Wang,[1,2] Yong Chen,[3] and Fei Wang[1,2,*]

SUMMARY

With advancements in large language models, artificial intelligence (AI) is undergoing a paradigm shift where AI models can be repurposed with minimal effort across various downstream tasks. This provides great promise in learning generally useful representations from biomedical corpora, at scale, which would empower AI solutions in healthcare and biomedical research. Nonetheless, our understanding of how they work, when they fail, and what they are capable of remains underexplored due to their emergent properties. Consequently, there is a need to comprehensively examine the use of language models in biomedicine. This review aims to summarize existing studies of language models in biomedicine and identify topics ripe for future research, along with the technical and analytical challenges w.r.t. interpretability. We expect this review to help researchers and practitioners better understand the landscape of language models in biomedicine and what methods are available to enhance the interpretability of their models.

INTRODUCTION

Recent progress made in large language models, i.e., GPT,[1] BERT,[2] and ChatGPT, presents a chance to rethink artificial intelligence (AI) systems, with language as a means to facilitate interaction between humans and AI. Generally, a language model is a probability distribution $p\,(w_1, w_2, \ldots, w_M)$ over a sequence of word tokens, with $w_m \in \Omega$ and Ω being a vocabulary, as shown in Figure 2A. But why would you want to compute such a probability of a word sequence? In the application scenario, the goal is to produce word sequences as output. For example, the goal of text summarization is to convert long texts into concise summaries. By computing the probability distribution over utterances, the word sequence can be generated by sampling tokens from this learned probability distribution.

A simple approach to computing the probability distribution of word sequence is to use statistical techniques, such as relative frequency counts. However, it is very data-intensive and suffers from high variance: even grammatical sentences will have a zero probability if they have not occurred in the training data. An alternative way is to compute the probability in the product format. N-gram models make a crucial simplifying approximation by conditioning on only the last $n - 1$ words. However, those traditional probabilistic language models require smoothing techniques to avoid the situation $p\,(w_1, w_2, \ldots, w_M) = 0$ when there is a rare or unseen word. Besides, these models are computationally intensive for large histories of text and cannot capture the long-range dependencies in language. Neural language models use neural networks or deep neural networks to model languages, such as feedforward neural networks, recurrent neural networks, and transformer neural networks. Neural language models have significant advantages over traditional probabilistic language models. Compared to n-gram models, neural language models are not constrained by the restricted context and can incorporate contexts from arbitrarily distant words, while remaining computationally and statistically tractable. Besides, neural language models can generalize better over contexts of similar words and are more accurate at word prediction. In this survey, we will focus on the neural language models and use the term "language model" (LM) to refer to the neural language models.

LMs usually use (low-dimensional) latent feature representation to implicitly capture the syntactic or semantic features of the language. The representation needs to be learned afresh for each new natural language processing (NLP) task, and in many cases, the size of the training data limits the quality of the latent feature representation. Given that the nuances of language are common to all NLP tasks, one could posit that we could learn generic latent feature representations from some generic tasks once and then share it across all NLP tasks. Language modeling, where the model needs to learn how to predict the next word given previous words, is such a generic task with abundant naturally occurring text to pre-train such a model (hence the name pre-trained language models). There are some benefits in pre-training, including (i) learning a universal representation through the massive corpus for downstream tasks, (ii) achieving an improved generalization ability and faster convergence with model initialization, and (iii) mitigating the overfitting issues in scenarios with limited data. There are several classes of pre-trained language models: autoregressive language models (GPT,[1] GPT-2,[3] ELMo[4]), masked language models (BERT,[2] XLM,[5] T5,[6] MASS[7]), permuted language models (XLNet[8]), and denoising autoencoders (BART,[9] mBART[10]), which are categorized by their ways of masking tokens, overcoming the mismatch issue, and recovering back the inputs. Besides, the pre-trained language models can also be categorized from

[1] Institute of Artificial Intelligence for Digital Health, Weill Cornell Medicine, New York, NY, USA
[2] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
[3] Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
*Correspondence: few2001@med.cornell.edu

https://doi.org/10.1016/j.isci.2024.109334

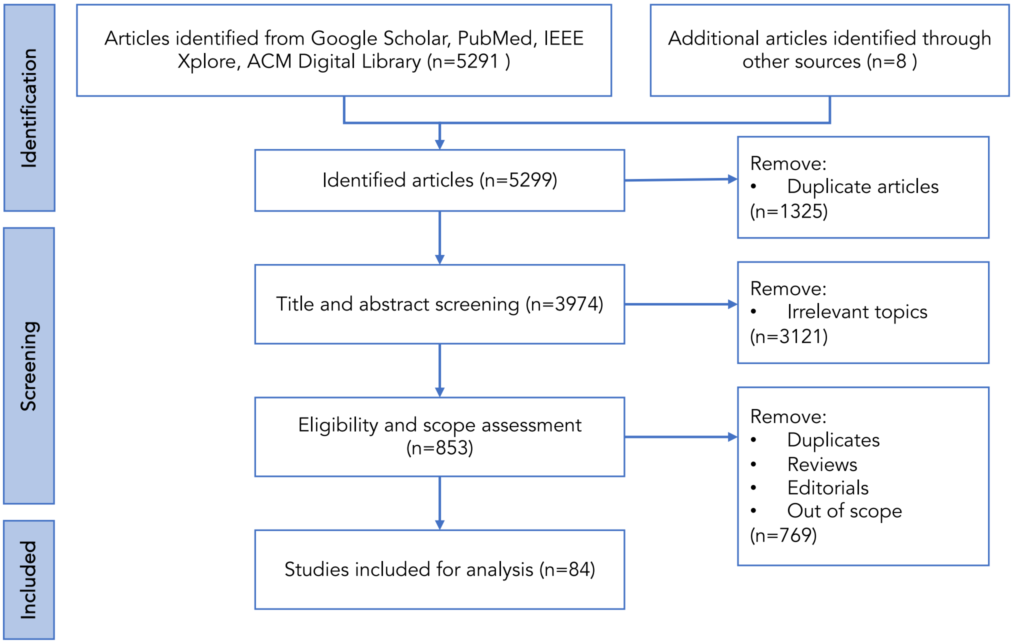
Identification
Screening
Included
Articles identified from Google Scholar, PubMed, IEEE Xplore, ACM Digital Library (n=5291)
Additional articles identified through other sources (n=8)
Identified articles (n=5299)
Remove:
• Duplicate articles
(n=1325)
Title and abstract screening (n=3974)
Remove:
• Irrelevant topics
(n=3121)
Eligibility and scope assessment (n=853)
Remove:
• Duplicates
• Reviews
• Editorials
• Out of scope
(n=769)
Studies included for analysis (n=84)

Figure 1. PRISMA flow diagram of study selection: language models in healthcare and biomedical research

other perspectives. For example, they can be divided into non-contextual and contextual models according to the representation used for downstream tasks. According to various scenarios, they can be categorized as knowledge-enriched LMs, multilingual or language-specific LMs, multi-model LMs, domain-specific LMs, and compressed LMs.

Healthcare and biomedicine represent vast domains of application, encompassing diverse areas of focus. Healthcare entails the delivery of care to patients via diagnosis, treatment, and health administration, while biomedical research concentrates on the scientific understanding of disease and the discovery of new therapeutic approaches. Both areas necessitate significant resources, time, and comprehensive medical knowledge. Language models can be trained on diverse sources or modalities of data in the biomedical domain, which have the potential to serve as a central storage of medical knowledge. In this way, they can be accessed and queried by medical professionals (e.g., healthcare providers and biomedical researchers) and by the public. By leveraging their strong adaptability through fine-tuning or prompting, language models can be effectively tailored to suit various specific tasks within healthcare and biomedicine. Despite the imminent widespread adoption of these models, our current understanding of how they work, when they fail, and what they are even capable of remains underexplored due to their emergent properties and complexity. Consequently, there is a need to examine the utilization of language models in healthcare and biomedicine.

Interpretability, often used interchangeably with explainability, refers to the ability to explain or provide meaning to model predictions. In particular, interpretability aims to describe the inner structure of a model in a manner that is easily understandable by humans.[11] In the medical domain, for example, there are great challenges in clinical decision support, such as diagnostic/prognostic/treatment uncertainties, and imbalanced, heterogeneous, noisy, sparse, high-dimensional datasets. Due to their powerful capacity, language models can be used for various use cases, including predicting the future diagnosis of depression in a temporal manner for mental health research,[12] recommending medications,[13] extracting cancer phenotypes,[14] and predicting a patient's likelihood of readmission to the hospital.[15] In these high-stakes decisions, however, one of the concerns in the deployment of such models is that there can still be high model misclassification. Besides, it has been widely shown that such models are not robust and may encounter failures in the presence of both artificial and natural noise.[16] Due to the black-box nature of such models, there is no easily discernible logic connecting the data to the decisions of the models. Therefore, providing explanations is critical to holding people/institutes accountable when models malfunction and gaining scientific understanding about models. To reach a level of explainable and usable machine intelligence, we need to not only learn from data, extract knowledge, generalize, and mitigate the curse of dimensionality but also disentangle the underlying explanatory factors of the data.

Therefore, the purpose of this scoping review is to map different types of corpora and language models used in existing healthcare and biomedical literature to their application tasks. Further, it seeks to identify topics ripe for future research, along with the technical and analytical challenges w.r.t. the interpretability. The processing and reporting of the results of this review were guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, as shown in Figure 1. We performed the literature search from various resources to find relevant articles published between Jan. 2015 and Dec. 2022: (i) the primary databases including Google Scholar, IEEE Xplore, ACM Digital Library, and PubMed; and (ii) the additional resources such as ACL Anthology. The search strategy for "language models for healthcare and biomedical research" is: ("language models" OR "Transformer" OR "deep neural networks" OR "pre-trained models") AND ("health" OR

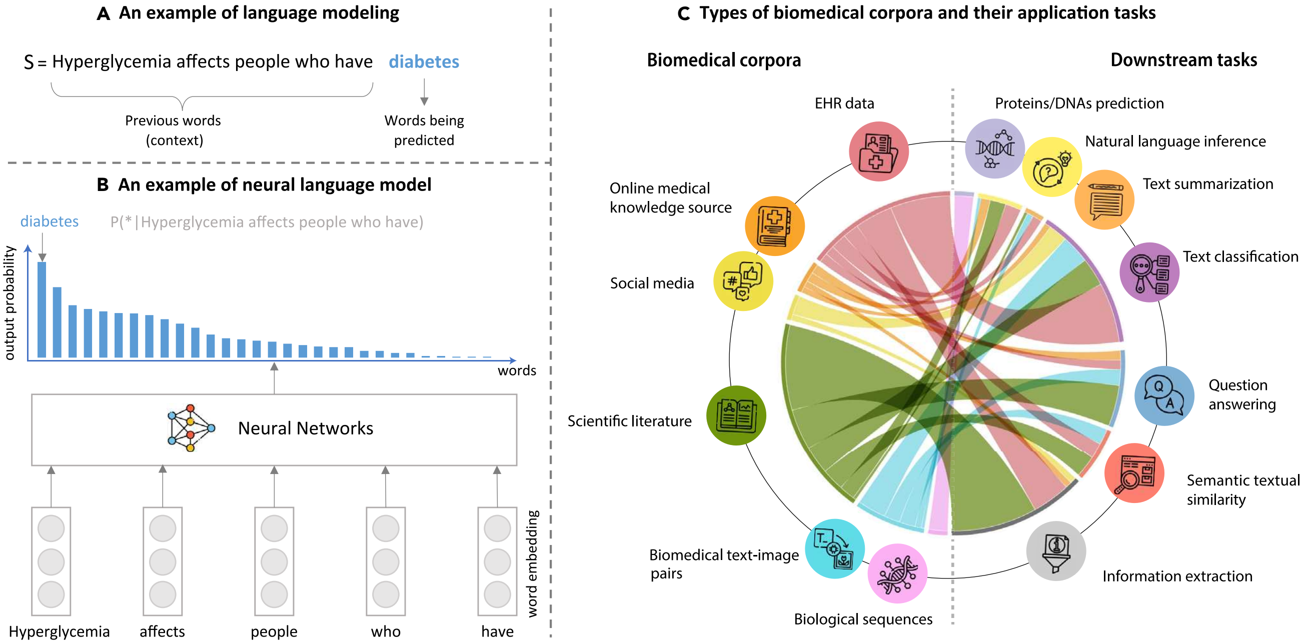

Figure 2. Overview of language model and its applications in biomedicine

(A) An example of language modeling that aims to predict the token of "diabetes" in the context of "Hyperglycemia affects people who have"; (B) An example of language model based on deep neural networks, which transforms each word token as word embedding, takes them as inputs, and computes the probability for the word being predicted through the language model; (C) Different types of corpus for language models in healthcare and biomedical research, and their associated application tasks.

"biomedical" OR "biomedicine"). The search strategy for "interpretability of language models" is: ("language models" OR "Transformer" OR "deep neural networks" OR "pre-trained models") AND ("health" OR "biomedical" OR "biomedicine") AND ("explainability" OR "interpretability"). Exclusions for the study selection were: (a) articles were not published in English; (b) commentaries or editorials; (c) the full text of the article is not accessible; (d) the language models are not based on deep neural networks; and (e) the outcome is not related to healthcare and biomedical research. But there might be a few limitations in this study: (i) we focused on the language models and limited several corpora as listed in the Results section, without including other types of corpora, such as speech data, audio recordings, video recordings, physiological data, medical robotic data, etc.; and (ii) the searched studies are all in English, which might result in the underrepresentation of language model applications in non-English-speaking countries. Despite of these, our review provides a landscape of the current literature on the language model and its interpretability in biomedicine.

RESULTS

Language models for healthcare and biomedical research

In this subsection, we classify the biomedical corpora used to train the language models into six types, followed by a presentation of each category in detail (as shown in Figure 2C). Besides, we make an overview table listing the various examined categories as shown in Table 1.

Electronic health records

Electronic health records (EHRs) have been utilized to store patient's health records from admission to discharge. These records contain a wealth of clinical data that can be leveraged to enhance patient care through knowledge discovery and the development of advanced algorithms. EHR data encompass both structured data (e.g., lab results and medical codes) and unstructured data (e.g., clinical notes, medication instructions, progress notes, or discharge summaries). Medical Information Mart for Intensive Care III (MIMIC-III) is the largest publicly available dataset of medical records, which consists of 58,976 unique hospital admissions from 38,597 patients in the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. Among EHR data, clinical notes contain valuable patient information but are challenging and costly to manually extract. Consequently, there is a need to effectively leverage the information embedded in clinical notes for research and practical applications. Zhu et al.[20] aimed to automatically extract clinical concepts by training ELMo on a corpus of clinical notes from MIMIC-III. Si et al.[19] proposed to pre-train BERT on clinical notes from MIMIC-III for clinical concept extraction. Shang et al.[13] proposed to combine Graph Neural Networks and BERT for medication recommendation where their model was pre-trained on the MIMIC-III dataset. Huang et al.[29] proposed the Clinical XLNet on the MIMIC-III dataset, in order to predict prolonged mechanical ventilation. Huang et al.[15] developed the ClinicalBERT and pre-trained the model on the clinical notes from MIMIC-III for the task of predicting hospital readmission.

Table 1. A summary of the selected studies in this review

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Zhang et al., 2019[17]	EHR	VetTag	https://github.com/yuhui-zh15/VetTag	CSU test data: F1 (66.2%), Precision (72.1%), Recall (63.1%), ExactMatch (26.2%)	text classification	saliency method
Liu et al., 2022[18]	EHR	MedM-PLM	https://git.openi.org.cn/liusc/3-6-liusicen-multi-modal-pretrain	2010-i2b2: F1 (86.29%); medication recommendation: AUC (95.57%); 30-day readmission prediction: AUC (74.7%); ICD coding: AUC (87.46%)	information extraction; classification	–
Huang et al., 2019[15]	EHR	ClinicalBERT	https://github.com/kexinhuang12345/clinicalBERT	clinical word similarity: Pearson correlation (67.0%); 30-day readmission prediction: AUC (71.4%)	semantic textual similarity; classification	attention weight
Si et al., 2019[19]	EHR	BERTbase, BERTlarge	https://huggingface.co/models?sort=trending&search=bert	i2b2 2010: F1 (90.25%); i2b2 2012: F1 (80.91%); Semeval 2014 Task 7: F1 (80.74%); Semeval 2015 Task 14: F1 (81.65%)	information extraction	–
Zhu et al., 2018[20]	EHR, Online Medical Knowledge Sources	Clinical ELMo	https://github.com/noc-lab/clinical_concept_extraction	2010 i2b2/VA: Precision (89.34%), Recall (87.87%), F1 (88.60%)	information extraction	–
Alsentzer et al., 2019[21]	EHR	Clinical BERT, Discharge Summary BERT	https://github.com/EmilyAlsentzer/clinicalBERT	i2b2 2010: Exact F1 (87.8%); i2b2 2012: Exact F1 (78.9%); MedNLI: Accuracy (82.7%)	information extraction; natural language inference	–
Shang et al., 2019[13]	EHR	G-BERT	https://github.com/jshang123/G-Bert	Jaccard (45.65%), PR-AUC (69.60%), F1 (61.52%)	classification	–
Rasmy et al., 2021[22]	EHR	Med-BERT	https://github.com/ZhiGroup/Med-BERT	DHF-Cerner: AUC (85.39%); PaCa-Cerner: AUC (82.23%); PaCa-Truven: AUC (80.57%)	classification	attention weights
Li et al., 2020[23]	EHR	BEHRT	–	AUC (90.4%), average precision score (21.6%)	classification	attention weights
Lewis et al., 2020[24]	EHR, Scientific literature	Bio-LM	https://github.com/facebookresearch/bio-lm	I2B2-2010: F1 (89.7%); HOC: Macro-F1 (86.6%); MedNLI: Accuracy (88.5%)	Information extraction; classification; natural language inference	–

(Continued on next page)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Peng et al., 2019[25]	EHR, Scientific literature	BlueBERT	https://github.com/ncbi-nlp/bluebert	MedSTS: Pearson (84.8%); BC5CDR: F1 (93.5%); i2b2 2010: F1 (76.4%); HOC: F1 (87.3%); MedNLI: Accuracy (84.0%)	Semantic textual similarity; Information extraction; text classification; natural language inference	–
Agrawal et al., 2022[26]	EHR	GPT-3+R	–	Biomedical Evidence Extraction: Accuracy (85%), F1 (61%); Medication status classification: Conditional Accuracy (89%), Conditional Macro F1 (71%)	Information extraction; classification	–
Chang et al., 2020[27]	EHR	Clinical BERT	https://github.com/dchang56/chief_complaints	Top-5 accuracies of 0.92 and 0.94 on datasets comprised of 434 and 188 labels, respectively	classification	–
Yang et al., 2022[28]	EHR, Scientific literature	GatorTron	https://github.com/uf-hobi-informatics-lab/GatorTron	2010 i2b2: F1 (89.96%); 2018 n2c2: F1 (96.27%); 2019 n2c2: Pearson correlation (89.03%); MedNLI71: Accuracy (90.20%); emrQA medication: F1 (74.08%), Exact Match (31.55%)	Information extraction; semantic textual similarity; natural language inference; question answering	–
Huang et al., 2019[29]	EHR	Clinical XLNet	https://github.com/lindvalllab/clinicalXLNet	prolonged mechanical ventilation: AUC (66.3%); 90-day mortality: AUC (77.9%)	classification	–
Zhou et al., 2022[14]	EHR	CancerBERT	https://github.com/zhang-informatics/CancerBERT	macro F1 scores equal to 0.876 (95% CI, 0.873–0.879) and 0.904 (95% CI, 0.902–0.906) for exact match and lenient match, respectively.	information extraction	–
Michalopoulos et al., 2020[30]	EHR, Online Medical Knowledge Sources	UmlsBERT	https://github.com/gmichalo/UmlsBERT	MedNLI: Accuracy (83.0%); i2b2 2010: F1 (88.6%)	natural language inference; information extraction	–
Kades et al., 2021[31]	EHR	Enhanced BERT	–	2019 n2c2: Pearson correlation (88.3%)	semantic textual similarity	–
Yang et al., 2020[32]	EHR	RoBERTa-MIMIC	https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER	2010 i2b2: F1 (89.94%); 2012 i2b2: F1 (80.53%); 2018 n2c2: F1 (89.07%)	information extraction	–
Meng et al., 2021[12]	EHR	BRLTM	https://github.com/lanyexiaosa/brltm	depression prediction: PRAUC (76%)	classification	attention weights

(*Continued on next page*)

Table 1. ***Continued***

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Chen et al., 2020[33]	EHR	AlphaBERT	https://github.com/wicebing/AlphaBERT	AUC (94.7%); ROUGE-L (69.3%)	text summarization	–
Wang et al., 2021[34]	EHR	CHMBERT	–	disease prediction: Top-1 F1 (61.95%), Top-5 F1 (91.58%), Top-1 F1 (96.83%),	classification	–
Zhang et al., 2020[35]	EHR	MC-BERT	https://github.com/alibaba-research/ChineseBLUE	cEHRNER: F1 (90%); cMedQA: F1 (82.3%); cMedTC: F1 (82.1%)	information extraction; question answering; text classification	–
Kraljevic et al., 2021[36]	EHR	MedGPT	–	NER+L: F1 (93%)	information extraction	saliency method
Khin et al., 2018[37]	EHR	ELMo	–	i2b2-PHI: F1 (89.87%–98.74%)	information extraction	–
Yang et al., 2020[38]	EHR	RoBERTa	https://github.com/uf-hobi-informatics-lab/2019_N2C2_Track1_ClinicalSTS	Pearson correlation (90.65%)	semantic textual similarity	attention weights
Xiong et al., 2020[39]	EHR	BERT-based model	–	2019 n2c2: Pearson correlation (86.8%)	semantic textual similarity	–
Mahajan et al., 2020[40]	EHR	ClinicalBERT	–	2019 n2c2: Pearson correlation (90.1%)	semantic textual similarity	–
Yan et al., 2022[41]	EHR	RadBERT	–	abnormal sentence classification: Accuracy (96.1%), F1 (95.6%); report coding: Accuracy (96.1%), F1 (96.0%); report summarization: ROUGE-1 (16.18%);	text summarization; classification	–
Lau et al., 2022[42]	EHR	BERTrad	https://github.com/wilsonlau-uw/BERT-EE	90.9%–93.4% F1 for finding triggers; 72.0%–85.6% F1 for arguments role extraction	Information extraction	–
Meng et al., 2020[43]	EHR	BERT-based model	–	Precision (97.0%), Recall (93.3%), F-measure (95.1%)	classification	–
Bressem et al., 2021[44]	EHR	FS-BERT & RAD-BERT	https://github.com/rAIdiance/bert-for-radiology	chest radiograph reports: AUC (97%–99%); CT reports: pooled AUC/AUPRC of 88%/80%	classification	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Naseem et al., 2022[45]	Biomedical image-text pairs	TraP-VQA	–	Overall: Accuracy (64.82%), Open-ended: Accuracy (37.72%), Close-ended: Accuracy (93.57%)	visual question answering	Gradient-weighted Class Activation Mapping (Grad-CAM), Shapley additive explanations (SHAP), attention weights
Li et al., 2020[46]	Biomedical image-text pairs	V + L models	https://github.com/YIKUAN8/Transformers-VQA	OpenI: averaged AUC (98.5%)	visual question answering	visualization of attention maps
Khare et al., 2021[47]	Biomedical image-text pairs	MMBERT	https://github.com/VirajBagal/MMBERT	VQA-Med 2019: overall Accuracy (67.2%), BLEU (69%); VQA-RAD: overall Accuracy (72%)	visual question answering	visualization of attention maps
Moon et al., 2022[48]	Biomedical image-text pairs	MedViLL	https://github.com/SuperSupermoon/MedViLL	diagnosis classification (Open-I): AUC (89.2%), F1 (40.7%); VQA-RAD: accuracy of 59.5%/77.7% for open-ended and close-ended questions, respectively	visual question answering; classification	visualization of attention maps
Chen et al., 2022[49]	Biomedical image-text pairs	Med-VLP	https://github.com/zhjohnchan/ARL	VQA-2019: overall Accuracy (80.32%); VQA-RAD: overall Accuracy (79.16%); MELINDA: Accuracy (80.51%)	visual question answering; classification	–
Chen et al., 2022[50]	Biomedical image-text pairs	M3AE	https://github.com/zhjohnchan/M3AE	VQA-RAD: overall Accuracy (77.01%); VQA-2019: overall Accuracy (79.87%); MELINDA: Accuracy (78.50%)	visual question answering; classification	–
Monajatipoor et al., 2022[51]	Biomedical image-text pairs	BERTHop	https://github.com/masoud-monajati/BERTHop	OpenI: AUC (98.12%)	classification	–
Boecking et al., 2022[52]	Biomedical image-text pairs	BioViL	https://huggingface.co/microsoft/BiomedVLP-BioViL-T	RadNLI: Accuracy (65.21%)	natural language inference	–
Lee et al., 2020[53]	Scientific literature	BioBERT	https://github.com/dmis-lab/biobert	2010 i2b2: F1 (86.73%), NCBI disease: F1 (89.71%), BC5CDR: F1 (93.47%), BC2GM: F1 (84.72%), ChemProt: F1 (76.46%), BioASQ 5b: Strict Accuracy (46%)	information extraction; question answering	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Shin et al., 2020[54]	Scientific literature	BioMegatron	https://github.com/NVIDIA/NeMo	BC5CDR-chem: F1 (92.9%), BC5CDR-disease: F1 (88.5%), NCBI-disease: F1 (87.8%), ChemProt: F1 (77.0%), BioASQ-7b-factoid: Strict Accuracy (47.4%)	information extraction; question answering	–
Gu et al., 2021[55]	Scientific literature	PubMedBERT	–	BC5-chem: F1 (93.33%), BC5-disease: F1 (85.62%), NCBI-disease: F1 (87.82%), BC2GM: F1 (84.52%), ChemProt: Micro F1 (77.24%), DDI: Micro F1 (82.36%), BIOSSES: Pearson (92.30%), HoC: Micro F1 (82.32%), PubMedQA: Accuracy (55.84%), BioASQ: Accuracy (87.56%),	information extraction; text classification; question answering; semantic textual similarity	–
Luo et al., 2022[56]	Scientific literature	BioGPT	https://github.com/microsoft/BioGPT	KD-DTI: F1 (38.42%), BC5CDR: F1 (46.17%), DDI: F1 (40.76%), PubMedQA: Accuracy (78.2%), HoC: F1 (85.12%)	information extraction; text classification; question answering	–
Kanakarajan et al., 2021[57]	Scientific literature	BioELECTRA	https://github.com/kamalkraj/BioELECTRA	BC5-chem: F1 (93.60%), BC5-disease: F1 (85.84%), NCBI-disease: F1 (89.38%), BC2GM: F1 (84.69%), ChemProt: Micro F1 (78.20%), DDI: Micro F1 (82.76%), BIOSSES: Pearson (92.49%), HoC: Micro F1 (83.50%), PubMedQA: Accuracy (64.02%), BioASQ: Accuracy (88.57%), MedNLI: Accuracy (86.34%)	information extraction; text classification; natural language inference; question answering; semantic textual similarity	–
Yasunaga et al., 2022[58]	Scientific literature	BioLinkBERT	https://github.com/michiyasunaga/LinkBERT	BC5-chem: F1 (94.04%), BC5-disease: F1 (86.39%), NCBI-disease: F1 (88.76%), BC2GM: F1 (85.18%), ChemProt: Micro F1 (79.98%), DDI: Micro F1 (83.35%), BIOSSES: Pearson (93.63%), HoC: Micro F1 (84.87%), PubMedQA: Accuracy (72.18%), BioASQ: Accuracy (94.82%)	information extraction; text classification; question answering; semantic textual similarity	–

(*Continued on next page*)

Table 1. ***Continued***

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Miolo et al., 2021[59]	Scientific literature	ELECTRAMed	https://github.com/gmpoli/electramed	NCBI-disease: F1 (87.54%), BC5CDR: F1 (90.03%), ChemProt: Micro F1 (72.94%), DDI: Micro F1 (79.13%), BioASQ: MRR (47.95%)	information extraction; question answering	–
Taylor et al., 2022[60]	Scientific literature	Galactica	https://github.com/paperswithcode/galai	BioASQ: Accuracy (94.3%), PubMedQA: Accuracy (77.6%), MedMCQA dev: Accuracy (52.9%)	question answering	attention visualization
Jin et al., 2019[61]	Scientific literature	BioELMo	https://github.com/Andy-jqa/bioelmo	BC2GM-Probe: F1 (88.4%), MedNLI-Probe: Accuracy (75.5%)	information extraction; natural language inference	–
Naseem et al., 2022[62]	Scientific literature, EHR	BioALBERT	https://github.com/usmaann/BioALBERT	BC5CDR-chem: F1 (98.08%), BC5CDR-disease: F1 (97.78%), NCBI-disease: F1 (97.18%), BC2GM: F1 (96.97%), ChemProt: F1 (78.32%), DDI: F1 (84.05%), i2b2: F1 (76.86%), BIOSSES: Pearson (92.80%), MedSTS: Pearson (85.70%), HoC: F1 (87.92%), BioASQ 4b-factoid: Accuracy (48.90%), BioASQ 5b-factoid: Accuracy (62.31%), BioASQ 6b-factoid: Accuracy (62.88%), MedNLI: Accuracy (79.52%)	information extraction; text classification; natural language inference; question answering; semantic textual similarity	–
Yuan et al., 2021[63]	Scientific literature, Online medical knowledge sources	KeBioLM	https://github.com/GanjinZero/KeBioLM	BC5-chem: F1 (93.3%), BC5-disease: F1 (86.1%), NCBI-disease: F1 (89.1%), BC2GM: F1 (85.1%), ChemProt: F1 (77.5%), DDI: F1 (81.9%), GAD: F1 (84.3%)	Information extraction	–
Tinn et al., 2021[64]	Scientific literature	PubMedELECTRA	https://huggingface.co/microsoft	BC5-chem: F1 (93.32%), BC5-disease: F1 (85.16%), NCBI-disease: F1 (87.73%), BC2GM: F1 (83.79%), ChemProt: F1 (76.74%), DDI: F1 (81.09%), BIOSSES: Pearson (92.01%), HoC: F1 (82.57%), BioASQ: Accuracy (92.07%), PubMedQA: Accuracy (67.64%)	information extraction; text classification; question answering; semantic textual similarity	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Ozyurt, 2020[65]	Scientific literature	Bio-ELECTRA	https://github.com/SciCrunch/bio_electra	BioASQ 8b-factoid: Exact Match (57.93%), BC4CHEMD: F1 (83.80%), BC2GM: F1 (72.55%), NCBI Disease: F1 (81.13%), LINNAEUS: F1 (85.02%), BioASQ 5b based: MRR (33.5%), GAD: F1 (80.96%), ChemProt: F1 (64.22%)	information extraction; question answering	–
Moradi et al., 2020[66]	Scientific literature	BERT-based-Summ	https://github.com/BioTextSumm/BERT-based-Summ	ROUGE-1 (75.04%), ROUGE-2 (33.12%)	text summarization	–
Xie et al., 2022[67]	Scientific literature	KeBioSum	–	CORD-19: ROUGE-1 (32.04%), PubMed-Long: ROUGE-1 (36.39%), s2orc: ROUGE-1 (37.44%), PubMed-Short: ROUGE-1 (43.98%)	text summarization	–
Du et al., 2020[68]	Scientific literature	BioBERTSum	–	PubMed: ROUGE-1 (37.45%), CNN/DailyMail: ROUGE-1 (43.13%)	text summarization	attention visualization
Wallace et al., 2021[69]	Scientific literature	BART-based model	–	XSUM: ROUGE-L (26.5%), Pretrain: ROUGE-L (26.9%), Decorate: ROUGE-L (26.6%), Sort by N·RoB: ROUGE-L (26.7%), Decorate and sort: ROUGE-L (26.5%)	text summarization	–
Guo et al., 2021[70]	Scientific literature	BART-based model	https://github.com/qiuweipku/Plain_language_summarization	ROUGE-1 (53.02%), ROUGE-2 (22.06%), ROUGE-L (50.24%)	text summarization	–
Kieuvongngam et al., 2020[71]	Scientific literature	BERT&GPT-2 based model	https://github.com/VincentK1991/BERT_summarization_1	extractive summary: ROUGE-1 (20%–70%), abstractive summary: ROUGE-1 (20%–45%)	text summarization	attention visualization
Chakraborty et al., 2020[72]	Scientific literature	BioMedBERT	https://github.com/BioMedBERT/biomedbert	GAD: F (79.92%), SQuAD v1.1: F1 (92.46%), EM (86.12%), NCBI Disease: F (87.51%), BC5CDR-Disease: F (87.51%), BC5CDR-chem: F (92.21%), BC4CHEMD: F (86.41%), BC2GM: F (82.32%),	information extraction; question answering	–
Oniani & Wang, 2020[73]	Scientific literature	GPT-2-based model	https://github.com/oniani/covid-19-chatbot	overall average rating score: 4.023	question answering	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Liévin et al., 2022[74]	Online Medical Knowledge Sources	CODEX 5-SHOT COT	https://github.com/vlievin/medical-reasoning	USMLE: accuracy (60%), PubMedQA: accuracy (78%)	question answering	–
He et al., 2020[75]	Online Medical Knowledge Sources	diseaseBERT	https://github.com/heyunh2015/diseaseBERT	MEDIQA-2019: MRR (90.00%), Accuracy (79.49%); TRCEQA-2017: MRR (57.21%), Accuracy (80.10%); MEDNLI: Accuracy (86.15%); BC5CDR: F1 (86.52%); NCBI: F1 (88.30%)	information extraction; question answering; natural language inference	–
Hao et al., 2020[76]	Online Medical Knowledge Sources	Clinical Kb-BERT/ALBERT	https://github.com/noc-lab/clinical-kb-bert	MedNLI: Accuracy (84.4%); i2b2 2010: Exact F1 (89.7%); i2b2 2012: Exact F1 (81.9%)	information extraction; natural language inference	–
Liu et al., 2020[77]	Online Medical Knowledge Sources	SapBERT	https://github.com/cambridgeltl/sapbert	NCBI: Accuracy (92.5%), BC5CDR-d: Accuracy (93.6%), BC5CDR-c: Accuracy (96.8%), AskAPatient: Accuracy (87.6%), COMETA: Accuracy (77.0%),	Information extraction	–
Singhal et al., 2022[78]	Online Medical Knowledge Sources	Flan-PaLM and Med-PaLM	https://huggingface.co/google/flan-t5-xl	67.6% accuracy on MedQA	question answering	–
Naseem et al., 2022[79]	Social Media	PHS-BERT	https://huggingface.co/publichealthsurveillance/PHS-BERT	Suicide Ideation: F1 (30.28%), Stress Detection: F1 (88.82%), Health Mention: F1 (87.38%), Depression Detection: F1 (76.49%), Vaccine Sentiment: F1 (81.10%), COVID Related: F1 (94.34%)	classification	–
Müller et al., 2020[80]	Social Media	CT-BERT	https://github.com/digitalepidemiologylab/covid-twitter-bert	CC: F1 (94.9%), VC: F1 (86.9%), MVS: F1 (74.8%), SST-2: F1 (94.4%), SE: F1 (65.4%)	classification	–
Tutubalina et al., 2021[81]	Social Media	RuDR-BERT& EnRuDR-BERT	https://github.com/cimm-kzn/RuDReC	sentence classification: Macro F1 (68.82%), Drug and disease recognition: Macro F1 (74.85%)	information extraction; classification	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Ji et al., 2021[82]	Social Media	MentalBERT & MentalRoBERTa	https://huggingface.co/mental	eRisk T1: F1 (93.38%), CLPsych T1: F1 (69.71%), Depression Reddit: F1 (94.23%), UMD: F1 (58.58%), T-SID: F1 (89.01%), SWMH: F1 (72.16%), SAD: F1 (68.44%), Dreaddit: F1 (81.76%),	classification	–
Papanikolaou et al., 2020[83]	Scientific literature	DARE (GPT-2)	https://openai.com/research/gpt-2-1-5b-release	CDR: F1 (73%), DDI2013: F1 (78%), ChemProt: F1 (73%)	Information extraction	–
Papanikolaou et al., 2019[84]	Scientific literature	BERT model	–	CDR: F1 (62.2%), GAD: F1 (69.8%), EUADR: F1 (81.2%), Healx CD: F1 (81.4%)	Information extraction	–
Wang et al., 2020[85]	Scientific literature	GLRE	https://github.com/nju-websoft/GLRE	CDR: F1 (68.5%), DocRED: F1 (57.4%)	Information extraction	–
Cabot et al., 2021[86]	Scientific literature	REBEL (BART)	https://github.com/babelscape/rebel	CONLL04: F1 (71.97%), NYT: F1 (91.76%), DocRED: F1 (41.84%), ADE: F1 (81.69%), Re-TACRED: F1 (90.39%),	Information extraction	–
Weber et al., 2022[87]	Scientific literature	transformer-based LM	https://github.com/leonweber/drugprot	F1 score of 79.73% on the hidden DrugProt test set	Information extraction	–
Heinzinger et al., 2019[88]	Biological sequence	SeqVec	https://github.com/Rostlab/SeqVec	Per-residue predictions: CASP12: Accuracy (76.5%), TS115: Accuracy (82.4%), CB513: Accuracy (80.7%)	Proteins/DNA prediction	–
Rives et al., 2021[89]	Biological sequence	ESM-1b Transformer	https://github.com/facebookresearch/esm	CB513: accuracy (71.6%), CASP13: accuracy (72.5%)	Proteins/DNA prediction	–
Xiao et al., 2021[90]	Biological sequence	ProteinLM	https://github.com/THUDM/ProteinLM	contact prediction: P@L/5 (75%), remote homology: Top 1 Accuracy (30%), Secondary Structure: Accuracy (79%), fluorescence: Spearman's rho (68%)	Proteins/DNA prediction	–
Brandes et al., 2022[91]	Biological sequence	ProteinBERT	https://github.com/nadavbra/protein_bert	Secondary structure - 3 state: accuracy (74%), Remote homology: accuracy (22%), Fluorescence: Spearman's ρ (66%),	Proteins/DNA prediction	attention visualization
Weissenow et al., 2022[92]	Biological sequence	EMBER2	https://doi.org/10.5281/zenodo.6412497	SetTst29: TM score (50%)	Proteins/DNA prediction	–

(*Continued on next page*)

Table 1. *Continued*

Authors, year	Biomedical Corpora	Model name	Open source (model)	Model performance	Application tasks	Interpretability technique
Ji et al., 2021[93]	Biological sequence	DNABERT	https://github.com/jerryji1993/DNABERT	predicts promoter regions: TATA (accuracy [92.2%], F1 [91.4%]), non-TATA (accuracy [97%], F1 [97%]); identifies transcription factor binding sites: both mean and median accuracy and F1 > 0.9	Proteins/DNA prediction	attention visualization
Yamada & Hamada, 2022[94]	Biological sequence	BERT-RBP	https://github.com/kkyamada/bert-rbp	154 RBPs: AUC (0.786%)	Proteins/DNA prediction	attention visualization
Mock et al., 2022[95]	Biological sequence	BERTax	https://github.com/f-kretschmer/bertax	loosely related dataset: accuracy (94.78% for superkingdom and 85.55% for phylum); distantly related dataset: accuracy (88.95% for superkingdom and 60.10% for phylum)	Proteins/DNA prediction	attention weights
Heinzinger et al., 2023[96]	Biological sequence	ProstT5	https://github.com/mheinzinger/ProstT5	secondary structure: accuracy@Q3 (89.4%); binding residues: F1 (37%); subcellular localization: accuracy@Q10 (57.3%); conservation: accuracy@Q9 (30.9%);	Proteins/DNA prediction	–

Chang et al.[27] aimed to derive a compact and computationally useful representation for free-text chief complaints by using the clinical BERT pre-trained on the MIMIC corpus. Kraljevic et al.[36] developed MedGPT with MIMIC-III and other EHR data for predicting the next disorder in a patient's timeline. Liu et al.[18] proposed to pre-train the model of MedM-PLM on the MIMIC-III dataset and evaluate its effectiveness on clinical tasks of medication recommendation, readmission prediction, and ICD coding. There are other language models[21,24–26,30,32] developed on MIMIC-III datasets.

In addition to MIMIC-III, there are many works using private sources of EHR data for pre-training language models.[12,14,17,22,23,31,33–35,37–40,97,98] For example, Li et al.[23] introduced the model of BEHRT to predict the likelihood of 301 conditions in one's future visits. Wang et al.[98] proposed the MEB model based on BERT for medication recommendation. Meng et al.[12] proposed the BRLTM model to predict future diagnoses of depression in mental health. Wang et al.[34] developed a Chinese BERT model for disease prediction and department recommendation tasks. Rasmy et al.[22] proposed the Med-BERT model to predict the diseases, such as diabetes, heart failure, and pancreatic cancer, by leveraging the structured EHR data. Danilov et al.[97] used neurosurgical data to predict the inpatient length of stay. Zhou et al.[14] proposed the CancerBERT model in order to extract breast cancer phenotypes from EHR data. Besides, there is some work using radiology reports as the corpus for pre-training the language models.[20,41–44]

Online medical knowledge sources

Online medical knowledge sources contain medicine and health-related information that is created and maintained by medical professionals. For example, the Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine, which includes the NCBI taxonomy, the Medical Subject Headings, Gene Ontology, OMIM, and the Digital Anatomist Symbolic Knowledge Base. There are over 2 million names for 900,000 concepts from more than 60 families of biomedical vocabulary, as well as 12 million relations among these concepts in UMLS. Liu et al.[77] aimed to capture fine-grained semantic relationships in the biomedical domain and proposed the SAPBERT model to self-align the representation space of biomedical entities by leveraging a massive collection of biomedical ontologies from UMLS. He et al.[75] integrated BERT-like pre-trained language models with disease knowledge for solving a variety of medical domain tasks, such as answering health questions, medical language inference, and disease name recognition. Hao et al.[76] introduced adding knowledge base information from UMLS into language model pre-training and obtained Clinical KB-BERT and Clinical KBALBERT for downstream tasks. Yuan et al.[63] proposed a biomedical pre-trained language model, KeBioLM, that can explicitly leverage knowledge from the UMLS knowledge bases. Michalopoulos et al.[30] incorporated domain knowledge into the pre-training process for clinical concept extraction by using a knowledge augmentation strategy with UMLS Metathesaurus. Besides, Zhu et al.[20] proposed to pre-train the ELMo model on Wiki pages using a domain-specific ontology such as SNOMED CT, to extract clinical concepts. Singhal et al.[78] proposed the Med-PaLM model to encode clinical knowledge from the medical question-answering datasets. Liévin et al.[74] investigated answering medical questions by performing reasoning and leveraging the expert-domain knowledge from medical exam question datasets.

Biomedical image-text pairs

This type of corpus contains two different data modalities, such as the image and text, in the biomedical domain. There are some popular sources for the corpus. For instance, the MIMIC Chest X-ray is a large publicly available dataset of chest radiographs with free-text radiology reports[99] from the Beth Israel Deaconess Medical Center. ROCO is a large-scale medical and multimodal imaging dataset where images and their corresponding captions are from publications available on PubMed Central. MedICaT is another dataset of medical image-caption pairs extracted from PubMed Central. Different from ROCO, 74% of its images are compound figures, including several sub-figures. In particular, there are 217,060 figures from 131,410 open-access biomedical papers, 7507 subcaptions, and subfigure annotations for 2,069 compound figures and inline references for around 25,000 figures in the ROCO dataset. IU X-ray has a collection of chest X-ray images from the Indiana University hospital network which includes the radiology images and XML reports. OpenI is another publicly available chest X-ray dataset collected by Indiana University, which has 3,996 radiology reports associated with 8,121 images. Li et al.[46] investigated different vision-and-language models for the visual question-answering task, with joint pre-training on chest X-ray radiographs and associated reports. Kaur et al.[100] proposed the RadioBERT model to generate radiological reports from chest X-ray images. Moon et al.[48] proposed the MedViLL model based on BERT for the tasks of diagnosis classification, medical image-report retrieval, medical visual question answering, and radiology report generation. Chen et al.[49] proposed to pre-train the medical vision-and-language model with medical domain knowledge for various downstream tasks. Monajatipoor et al.[51] proposed a vision-and-language model of BERTHop for chest X-ray disease diagnosis. Chen et al.[50] proposed a multimodal masked auto-encoder method for the medical vision-and-language understanding tasks. Boecking et al.[52] proposed the BioViL model for self-supervised multi-modal learning on paired image-text radiology data. Naseem et al.[45] aimed the pathology visual question-answering task by utilizing high- and low-level interactions on the pathology image (vision) and question (language) to generate an answer.

Social media

Users often post information on social media platforms and recent studies have shown that health-related social media data are useful in many applications to provide better health-related services. For example, Twitter is a social media platform where users post and interact with messages known as "tweets." Müller et al.[80] proposed the COVID-Twitter-BERT model by pre-training on a large corpus of COVID-19-related tweets. Zhang et al.[101] pre-trained language models on HPV vaccine-related tweets for the sentiment analysis of the HPV vaccination task. Naseem et al.[79] proposed the PHS-BERT model for tasks related to public health surveillance on social media by pre-training on health-related tweets. For Reddit, it is a social news aggregation, web content rating, and discussion website. Ji et al.[82] proposed MentalBERT and MentalRoBERTa for depression detection and other mental disorders classification with the mental health posts on Reddit. Besides, Tutubalina et al.[81] proposed the RuDR-BERT model for drug reactions and effectiveness detection by pre-training the model on the health-related user-generated texts collected from social media in Russian.

Scientific literature

As valuable knowledge is discovered from biomedical literature, biomedical researchers begin to develop pre-trained language models to handle biomedical text. PubMed and PubMed Central (PMC) are the two popular sources of biomedical text. PubMed contains only biomedical literature citations and abstracts only while PMC contains full-text biomedical articles. There is a large portion of work pre-training the proposed model on the corpus from PubMed and PMC[25,63,53–62,64,65] for biomedical information extraction. Moradi et al.[66] proposed a BERT-based model for biomedical text summarization with pre-training on PubMed, PMC, and Wiki. Du et al.[68] proposed the BioBERTSum model to better capture token-level and sentence-level contextual representation for extractive summarization tasks in the biomedical domain. Wallace et al.[69] and Guo et al.[70] both proposed BART-based models for biomedical text summarization with pre-training on the corpus of Cochrane systematic reviews indexed in PubMed.

BREATHE is another large and diverse dataset collection of biomedical research articles that contains titles, abstracts, and full-body texts. The primary advantage of the BREATHE dataset is its source diversity, including BMJ, arXiv, medRxiv, bioRxiv, CORD-19, Springer Nature, NCBI, JAMA, and BioASQ. Kieuvongngam et al.[71] proposed to use BERT and GPT-2 for the text summarization of COVID-19 medical research articles from CORD-19. Chakraborty et al.[72] proposed the BioMedBERT model for the task of question-answering by pre-training the model on the BREATHE dataset. Oniani et al.[73] proposed a GPT-2-based model for the task of question-answering for COVID-19 with pre-training on the corpus of CORD-19. Xie et al.[67] proposed the KeBioSum model for biomedical text summarization with the corpus of CORD-19 and PubMed. Taylor et al.[60] developed the Galactica model pre-trained on a large scientific corpus of papers that can perform the task of medical question answering. Besides, there are some works pre-training the models on the corpus of chemical disease relation or drug and adverse effects for the task of biomedical relation extraction.[83–87]

Biological sequences

In addition to the text or image data, the biological sequence data can be another corpus for pre-training language models. For example, the structure of each protein is fully determined by a sequence of amino acids; however, these amino acids are from a limited-size amino acid vocabulary, of which 20 are commonly observed. This is similar to text that is composed of words in a lexicon vocabulary. The Pfam dataset is a large collection of protein families, in which each protein is represented by multiple sequence alignments using hidden Markov models. Xiao et al.[90] proposed the model of ProteinLM for the protein prediction task with the preprocessed Pfam. Heinzinger et al.[88] proposed the SeqVec model to predict the protein function and structure from sequences and they further presented the ProstT5 model by combining 1D sequence with 3D structure.[96] Rives et al.[89] proposed to use the language model for the tasks of remote homology detection, prediction of secondary structure, long-range residue-residue contacts, and mutational effect for protein sequences. Brandes et al.[91] proposed the ProteinBERT model for protein sequences designed to capture local and global representations of proteins in a natural way. Weissenow et al.[92] proposed the EMBER2 model for protein structure prediction without requiring any multiple sequence alignments. Besides, Ji et al.[93] proposed the DNABERT model to predict the promoters, splice sites, and transcription factor-binding sites with the DNA sequence. Yamada et al.[94] proposed the BERT-RBP model to predict RNA and RNA-binding protein interactions by adapting the BERT architecture pre-trained on a human reference genome. Mock et al.[95] proposed the BERTax model to taxonomically classify the superkingdom and phylum of DNA sequences.

In the following, we categorize various biomedical downstream tasks, as shown in Figure 2C.

Information extraction

Information extraction plays an important role in automatically extracting structured biomedical information from unstructured biomedical text data ranging from biomedical scientific literature, and EHR data, to biomedical-related social media corpus, etc. It generally refers to several important sub-tasks in this review, including named entity recognition and relation extraction. For instance, named entity recognition is the first step in unlocking valuable information in unstructured text data that aims to identify the concept or entity names in biomedical texts. Extracting clinical concepts, such as types of diagnosis, test, treatment, clinical department, medication, adverse drug events, etc., is useful for EHR corpus,[14,20,19,21,24–26,30,32,35,42,28] while extracting biomedical entities, such as disease entity, drug-chemical entity, drug-protein entity, species entity, etc., is meaningful to discover knowledge in scientific literature,[25,63,53–55,57–59,61,62,64,65,102] online medical knowledge corpus,[30,63,75–77] or social media posts.[81] Relation extraction aims to identify the relationship or semantic correlation between biomedical entities mentioned in texts and generally be considered as a classification problem to predict the possible relation type of two identified entities in a given sentence.[25,42,77,76,63,53–59,62,64,65,83–87]

Text classification

Text classification aims to assign one of the predefined labels to variable-length texts like phrases, sentences, paragraphs, or documents in the corpus like EHR data,[24–26,35,41,44] biomedical scientific literature,[55–58] and social media data.[80,79,81,82,64]

Semantic textual similarity

Semantic textual similarity aims to measure the degree of semantic similarity between two phrases or sentences.[25,55,57,58,62,64] Typically, it can be formulated into a regression problem to predict the similarity score for each pair. In the clinical domain,[28,31,38,39,40] semantic textual similarity has the potential to facilitate clinical decision processes, such as highlighting crucial text snippets in a report, query databases for similar reports, assessing the quality of reports, or being used in question-answering applications.

Question answering

Question answering (QA) aims to extract answers for the given queries. QA can facilitate seeking information in clinical notes,[28,35] biomedical scientific literature,[53–60,62,64,72,73] biomedical image-text corpus,[28,45–51] and online medical knowledge corpus,[74,78] and thus save time for the clinicians and biomedical researchers.

Text Summarization

Typically, the clinical notes, scientific literature, and radiology reports could be lengthy in nature. However, clinicians or biomedical researchers need to go through a large number of biomedical documents, which is time-consuming. In this context, there is a need for automatic text summarization, in order to reduce the effort and time required by clinicians or biomedical researchers. Text summarization falls into two broad categories, namely extractive summarization,[33,66,67,68,71] which identifies the most relevant sentences in the document, while abstractive summarization[41,56,69–71] generates new text, which represents the summary of the document.

Natural language inference

Natural language inference (NLI) aims to identify the semantic correlation between a pair of sentences, i.e., whether the second sentence entails or contradicts or is neutral with the first one.[21,24,25,28,30,52,57,61,62,76] Since NLI requires sentence-level semantics, it is particularly useful in tasks like paraphrase mining and information retrieval in the general domain and medical concept normalization, semantic relatedness, and question answering in the biomedical domain.

Proteins/DNAs prediction

Protein can be associated with almost every life process. Consequently, analyzing the biological structure and property of protein sequences and understanding their functions[88–92,96] becomes crucial to the study of life science as well as disease detection and drug discovery. Since only a fraction of all species are available in today's databases, it is important to accurately assign DNA sequences to their origin particularly when there are no closely related species in databases.[95] Deciphering the language of non-coding DNA is also one of the fundamental problems in genome research.[93] Besides, identifying RNA and RNA-binding protein interactions[94] can help to understand the biological roles in regulating cellular functions.

Interpretability of language models

Language models, particularly large language models like BERT, have become highly widespread. The increase in model complexity is driven by a general correlation between model size and model performance. A growing concern is therefore whether these models are reliable and trustworthy in downstream applications. Explainability can offer evidence and justification for decision-making, which is also critical in the healthcare and biomedical domains. We summarize the explanation techniques used in the language models as shown in the following section.

Attention-based methods use attention weights as the importance scores.[103,104] They appeal to human intuition and can help indicate where the model is "focusing."[12,15,22,23,38,94,95,105] For example, Huang et al.[15] aimed to predict 30-day hospital readmission by developing the model of ClinicalBERT with clinical notes. Further, the predictions generated from ClinicalBERT can be interpreted by its model's attention weights, revealing which terms in clinical notes are predictive of patient readmission. Meng et al.[12] aimed to predict a future diagnosis of depression by proposing a bidirectional representation learning model with a Transformer architecture on EHR data. Besides, the model's interpretability was boosted by the quantitative analysis of self-attention weights of EHR sequences, demonstrating the inner relationship between various topic features and diagnosis codes. Córdova Sáenz and Becker[106] proposed a framework to classify stances expressed in tweets regarding COVID-19 vaccination using BERT-based models and an interpretation mechanism that obtains the most relevant words in terms of attention weights for model decision-making. Shi et al.[107] proposed a corpus-level explanation approach, which aimed at capturing causal relationships between keywords and model predictions via learning the importance of keywords for predicted labels across a training corpus based on a hierarchical attention network. Chrysostomou and Aletras[108] aimed to improve the faithfulness of attention-based explanations for text classification by proposing a new family of task-scaling mechanisms, which can learn task-specific non-contextualized information to scale the original attention weights. Bacco et al.[109] proposed two different transformer-based methodologies by exploiting the inner hierarchy of the documents to perform a sentiment analysis task while extracting the most important sentences (with regard to the model decision) to build a summary as the explanation of the output. Niu et al.[110] proposed the method of jointly embedding words and labels whereby attention modules learn the weights of words from medical notes according to their relevance to the names of risk prediction labels. Tutek and Šnajder[111] proposed to improve the faithfulness of attention based on regularization methods that promote the retention of word-level information. Liu et al.[112] proposed a novel practical framework by utilizing a two-tier attention architecture to decouple the complexity of explanation and the decision-making process. Rigotti et al.[113] proposed the generalization of attention from low-level input features to high-level concepts as a mechanism to ensure the interpretability of attention scores. In particular, they designed the ConceptTransformer that exposes explanations of the output of a model in which it is embedded in terms of attention over user-defined high-level concepts.

Shapley additive explanation (SHAP) is to compute shapely values for each combination of the features (a power set of the features) by training a linear model. But, it will be computationally expensive to train 2^M models for M set of features. For example, Attanasio et al.[114] investigated the SHAP-based explainability approach on Transformer-based models.

Visualization plays an essential role in understanding how a neural model works.[115] It can be applied with any of the feature importance-based methods. With visualization, we can project the feature importance weights using heatmap, partial dependency plot, etc. Saliency has been primarily used to visualize the importance scores of different types of elements in XAI learning systems,[36,17] such as showing input-output word alignment,[116] highlighting words in input text,[117] or displaying extracted relations.[118] Ding and Koehn[119] investigated the gradient-based saliency methods on different language models based on the perspective of plausibility and faithfulness. Malkiel et al.[120] proposed the BTI approach to explain paragraph similarities inferred by pre-trained BERT models. Specifically, the proposed approach can identify important words that dictate each paragraph's semantics, match between the words, and retrieve the most important pairs by utilizing activation and saliency maps. Natural language explanation is verbalized in human-readable natural language. The natural language can be generated using sophisticated deep learning models, e.g., by training a language model with human natural language explanations and coupling with a deep generative model.[121] It can also be generated by using simple template-based approaches.[122] Brand et al.[123] developed the E-BART model by jointly making a veracity prediction and providing an explanation within the same model. Sammani et al.[124] proposed the NLX-GPT that can simultaneously predict an answer and explain it by formulating the answer prediction as a text generation task along with the explanation. Besides, there are other visualization techniques for the purpose of interpretability. For example, Dunn et al.[125] proposed a context-sensitive visualization method with Leave-N-Out that leads to heatmaps that include more of the relevant information pertaining to the classification, as well as more accurately highlighting the most important words from the input text. Li et al.[126] developed a visual analysis method to enable a unified understanding of models for text classification. Specifically, the mutual information-based measure was used to provide quantitative explanations on how each layer of a model maintains the information of input words in a sample.

There are also some works that aim to improve the interpretability of the Transformer-based vision and language (multimodal) model. For example, Naseem et al.[45] aimed to develop a model that can answer a medical question posed by pathology images. They proposed TraP-VQA that embeds the image and question features, coupled with domain-specific contextual information, via a transformer for PathVQA. Grad-Cam and SHAP were used to interpret the retrieved answers visually to indicate which area of the image contributed to the predicted answer. Visualization of the transformers' attention showed proposed model assigns more weight to the relevant words and explains the reason for the retrieved answer. Aflalo et al.[127] proposed the VL-InterpreT method that can provide interactive visualizations for interpreting the attention and hidden representations in multimodal transformers.

DISCUSSION

Language models, particularly pre-trained language models, provide great promise in their ability to learn a generally useful representation from the knowledge encoded in the corpora by being repurposed with minimal effort for diverse downstream tasks in the biomedical domains. Interpreting the decision mechanism of a pre-trained language model can help understand the rationale behind its success and its limitations. In this section, we further discuss the challenges in the aforementioned explanation methods, and uncover the gaps and future research directions toward the interpretability in language models.

Other interpretability techniques

In addition to the attention-based method, SHAP, and visualization method, there are some other interpretability techniques that could be used in language models. For example, knowledge graphs can enhance language representation since knowledge graphs have high entity/concept coverage and strong semantic expression ability. Further, knowledge graphs can also be used to improve interpretability. Yan et al.[128] proposed a sentiment analysis knowledge graph-BERT model by combining both the knowledge graph and the language representation model of BERT together. Further, the interpretability can be improved by injecting triples from the knowledge graph into sentences as domain knowledge. Islam et al.[129] developed the method of AR-BERT, which is a two-level global-local entity embedding scheme that allows efficient joint training of knowledge graphs (KG)-based aspect embeddings and aspect-level sentiment classification models. Interpretability was enhanced by the semantic relations between aspects extracted from KGs.

Interpretability can be achieved through counterfactual explanation and adversarial examples (AE). A counterfactual explanation involves generating an instance that is similar to the original instance but leads to a different model prediction. This counterfactual instance helps understand what changes in the input features would result in a different model output. For AE, one can know the scenario in which its model is going to generate an incorrect output. It will provide an explanation that which type of edit has led to the change in the output. In order to secure the model from AE attacks, models can be trained on adversarial data. Feder et al.[130] proposed the framework of CausaLM that can produce causal model explanations using counterfactual language representation models. Taylor et al.[131] proposed to apply the model of InfoCal to the task of predicting hospital readmission using hospital discharge notes, where the model can produce extractive rationales for its predictions by using the adversarial-based technique. Li et al.[132] proposed a joint classification and rationale extraction model for both explainability and robustness. Specifically, the mixed Adversarial Training was designed to use various perturbations in discrete and embedding space to improve the model's robustness, and the Boundary Match Constraint was to locate rationales more precisely with the guidance of boundary information.

Neurosymbolic methods can produce an answer to a complex query by chaining these operations together, passing inputs from one module to another. This has the benefit of producing an interpretable trace of intermediate computations, in contrast to the "black box" computations common to end-to-end deep learning approaches. Creswell et al.[133] proposed a selection inference framework that exploits pre-trained large LMs as general processing modules, and alternates between selection and inference to generate a series of interpretable, symbolic reasoning steps leading to the final answer.

Layer-wise relevance propagation is another way to attribute relevance to features computed in any intermediate layer of a neural network (NN). Definitions are available for most common NN layers including fully connected layers, convolution layers, and recurrent layers. Layer-wise relevance propagation has been used to, for example, enable feature importance explainability[134] and example-based explainability.[135] Aken et al.[136] presented a layer-wise analysis of BERT's hidden states to understand their internal functioning. They focused on models fine-tuned on the task of QA as an example of a downstream task and inspected how QA models transform token vectors in order to find the correct answer. Aken et al.[137] proposed the VisBERT that can visualize the contextual token representations within BERT for the task of (multi-hop) QA. Interpretability can be provided by observing how the semantic representations are transformed throughout the layers of the model. Sevastjanova et al.[138] aimed to explain models by exploring the continuum between function and content words with respect to contextualization in BERT. Specifically, they utilized the similarity-based score to measure contextualization and integrate it into a visual analytics technique, presenting the model's layers simultaneously and highlighting intra-layer properties and inter-layer differences.

Advantages and disadvantages of interpretability techniques

Gradient-based interpretability vs. layer-wise relevance propagation-based interpretability: Gradient-based methods treat the gradient (or some variant of it) of the model output w.r.t. each input feature as its relative importance.[139] The feature can typically be a pixel in an image or a token in the text. Intuitively, the gradient represents how much difference a tiny change in the input will apply to the output. Regarding layer-wise relevance propagation-based methods, they are a more generalized solution by using a high-level relevance conservation constraint, i.e., the total incoming relevance into a neuron should equal the total outgoing relevance from it. They have been applied to sentence classification tasks to explain which tokens are most important to the prediction. Compared to gradient-based methods, there are some advantages in layer-wise relevance propagation-based methods. First, they do not require the differentiability or smoothness properties of neuron activations. Second, it provides a way to quantitatively assess its faithfulness via a perturbation-based evaluation.[140] However, there are also some drawbacks in the layer-wise relevance propagation-based methods, such as suffering from the saturation problem[141] and no principled way to decide which rule to choose for which type of layer. Overall, the strengths of these two types of methods are: (i) they generate a spectrum of feature relevance scores, which is easily understandable for all kinds of target users and (ii) they are easy to compute—gradient-based methods require only a few calls to the model's backward function while layer-wise relevance propagation involves a custom implementation of the backward pass. Their weaknesses are obvious as well: (i) most existing work targets low-level features, and it is non-intuitive how to compute any gradient w.r.t. higher-level features like semantic role, syntax dependency, and discourse relations; (ii) it is questionable how to apply such methods to non-classification tasks, especially when there is no single output of the model, e.g., text generation or structured prediction; and (iii) the explanation might be unstable, i.e., minimally different inputs can lead to drastically different relevance maps.[142,143]

Attention-based interpretability: As Transformers has become the backbone architecture for many language models, the attention mechanism in Transformers, a.k.a. self-attention, is widely used as well. Simply, self-attention is a function of the affine transformation between an input sequence of vectors and an output sequence of vectors. Its weights are called attention weights, intuitively representing how much the model "attends to" each input vector when computing the weighted average. Therefore, it is appealing to interpret attention weights as the importance of input tokens to the output. Such types of understanding have been used (implicitly or explicitly) as evidence for model interpretability in different tasks and domains, such as text classification,[144] knowledge base induction,[118] and medical code prediction.[117] Despite these intuitive findings, there is a debate on whether the attention mechanism can be a faithful model explanation. For example, prior work[103] contends that attention weights do not correlate well with other feature importance-based explanation methods. Also, it is possible to construct an adversarial attention distribution, i.e., one that is maximally different from the original distribution but has little influence on the model output. There are also some counter-arguments:[104] (i) attention weights can provide an explanation, but that does not have to be the only explanation. In practice, most tasks considered in the study by Jain and Wallace[103] are binary classification, which means that it is possible to construct adversarial attention distributions that differ significantly from the original distribution but have little effect on the model's output. This may suggest that attention weights are not always a reliable indicator of feature importance. (ii) Adversarial distributions are not adversarial weights. The adversarial attention distributions are artificially constructed by humans, but not learned by models through training. Overall, its strengths are: (i) the visualization of model-internal structures is intuitive and readable to humans, especially end users; (ii) the attention mechanism can capture the interaction between features, whereas many other methods can only capture the influence of individual features themselves; and (iii) attention weights are easily accessible and computationally efficient, compared to other methods. For its weaknesses: (i) it is questionable to what extent attention weights represent causal contribution, as mentioned in the debate; (ii) simply focusing on attention weights in a single layer and/or from a single token position may reflect how much the model attends to each input position locally, but not taking the whole computation path into account. So the attention mechanism in hierarchical architecture might mitigate the issue and improve the interpretability.

Counterfactual intervention methods explain the causal effect between a feature/concept/example and the prediction by erasing or perturbing it and observing the change in the prediction. Counterfactual examples, therefore, refer to the outcome of perturbations. Although counterfactual examples and adversarial examples look similar in the robustness literature, they differ in this context: (i) the goal of the former is to explain the model's reasoning mechanism, while that of the latter is to examine model robustness; (ii) the former should be meaningfully different in the perturbed feature to the original example while the latter should be similar to or even indistinguishable from it; and (iii) the former can lead to changes in the ground truth label, whereas the latter should not.[145] However, generating high-quality counterfactual examples is non-trivial, as they need to simultaneously accord with the counterfactual target label, be semantically coherent, and only differ from the original example in the intended feature. In existing work, the most reliable (yet expensive) approach to collecting counterfactual examples is still manual creation.[145,146] Besides, counterfactual intervention can directly happen on the level of examples, such as the methods of influence functions. Influence functions are based on counterfactual reasoning – if a training example were absent or slightly changed, then how would the prediction change? Since it is impractical to retrain the model after erasing/perturbing every single training example, influence functions provide an approximation by directly recomputing the loss function. However, it is found in the existing work[147] that influence functions can become fragile and the approximation accuracy can vary significantly depending on a variety of factors, such as network architecture, depth, width, the extent of model parameterization and regularization techniques, and the examined checkpoints, as models become more complex. Counterfactual intervention can also happen in the feature representations in the model, such as the work of Amnesic Probing[148] and CausalLM.[130] They both aim to answer the more insightful question – is some high-level feature, e.g., syntax tree, used in prediction? They exploit different algorithms to erase the target feature from the model representation and then measure the change in the prediction. The larger the change, the more strongly it indicates that the feature has been used by the original model. In terms of faithfulness, only CausalLM is validated with a white-box evaluation, whereas no explicit evaluation is provided for Amnesic Probing. Causal inference can also be used for interpretability. However it requires a more rigorous formalization of the causal framework, e.g., a causal model, which is usually task- or even dataset specific and needs to be designed by domain experts. Therefore, there are still important challenges such as how to automatically derive causal models from data and how to make them more generalizable across tasks. Overall, counterfactual interventions can capture causal relationships instead of mere correlational effects between inputs and outputs and are more often explicitly evaluated in terms of faithfulness. However, counterfactual intervention is relatively more expensive in computational cost, normally requiring multiple forward passes or modifications to the model representation. Searching for the right targets to intervene in can also be costly. Interventions are often overly specific to the particular example and this calls for more insights into the scale of such explanations.[149] Counterfactual intervention may suffer from hindsight bias, which questions the foundation of counterfactual reasoning.[150]

Surrogate models for post hoc interpretability: SHAP is one of the widely adopted surrogate-model-based methods that can be thought of as using additive surrogate models as an explanation. Shapley values are theoretically shown to be locally faithful, but there is no empirical evidence on whether this property is maintained after the SHAP approximation. Subsequent work also finds other limitations: linear surrogate models have limited expressivity. For example, if the decision boundary is a circle and the target example is inside the circle, it is impossible to derive a locally faithful linear approximation. Besides, they can result in nonsensical inputs or representations, which sometimes allow adversaries to manipulate the explanation.[151] What's more important, one major concern of using SHAP in the medical domain is that the Shapley value was originally derived from economics tasks, where the cost is additive. However, clinical features are usually heterogeneous, and the Shapley values derived from the model may not be meaningful.[152]

Faithfulness and plausibility of interpretability

In addition to explanation methods, interpretability can be evaluated from the trustworthy aspects: how faithful the explanation is and how understandable the explanation is to humans, a.k.a., faithfulness and plausibility. Specifically, faithfulness measures the degree to which the rationales in fact influence the corresponding predictions,[153,154] while plausibility measures how much the rationales provided by models align with human-annotated rationales.[153,155] These two aspects are often at odds with each other. This is because a complex model decision might require a rather complex explanation to cover all of the possible aspects of the model's behaviors on different inputs, which might not look easy to understand to humans. Regarding faithfulness, a perfectly faithful interpretation accurately represents the decision-making of the model being explained. If the explanation is constrained to agree with the model's behavior on all possible inputs, then no simpler explanation than the original model is possible. When applying an explanation method to black-box models trained on biomedical data, it is necessary to consider: (i) the concordance between the explanation method and the original model. If the concordance is low, then the model is not faithful; (ii) if changing the feature importance based on the explanation would alter the original predictions; (iii) if the same model might produce different explanations for the same pair of input-outputs over multiple runs. Regarding plausibility, we discuss it from the different perspectives of human expert users. Like any other data-driven machine learning approach, language models for biomedical problems aim to further improve performance by learning much more complex representations from raw features while sacrificing model transparency. Explanation methods may provide human-understandable explanations, yet it is crucial that the explanations should be aligned with our knowledge to be trustable, especially for real-world deployment in the biomedical domain. From a clinical perspective, for example, it is necessary and critical to have clinically relevant features that align with medical knowledge and clinical practice. However, current deployments with explanation methods mainly focus on helping to debug the model for engineers rather than the real-world use for end users.[156] For model developers, they evaluate their use of interpretation methods with different levels of model transparency from both quantitative and qualitative (visualization) perspectives. But they usually overtrust the methods and this may lead to their misuse since good visualization may sway human thought but may not fully explain the behavior of the system and may be incorrectly interpreted by developers. So the appropriate explanation methods should be selected and evaluated both to help model developers (data scientists and machine learning practitioners) understand how their models behave and to assist clinicians and biomedical researchers to understand the rationale for predictions produced by the model.

Case studies in healthcare and biomedicine

In healthcare, language models can be used to improve the efficiency and accuracy of care provided by health professionals. For example, EHR, including clinical notes, lab tests, radiology reports, and discharge summaries, contains significant clinical values since it can provide a richer picture of the patient by describing symptoms, reasons for diagnosis, radiology results, daily activities, and patient history. Making accurate clinical predictions might require health professionals to spend unnecessary time reading and analyzing EHR. In these settings, language models can be adapted to help predict the diagnosis, suggest treatments and discharges, generate summaries of patient visitation, and predict hospital readmissions. Further, interpretability could be used to disentangle the underlying explanatory factors of the data, such as uncovering which terms in clinical notes are predictive of patient readmission[15] or demonstrating the relationship between the topic features and diagnosis codes.[12] Besides, language models can be adapted to answer medical questions,[45] along with the relevant medical explanatory information. With interpretability, it would significantly enhance the trust of both the health professionals and patients in outputs produced by such models.

In biomedicine, it is critical to first identify a target (e.g., proteins, DNA, RNA) and search for molecules that bind to the target before discovering a drug or a therapeutic that treats the disease.[157] Language models in these settings can be adapted to improve the search space and efficiency, which reduces the amount of experiments and discovers new drugs. Although these biological sequence data have exhibited similarity to human language, ranging from alphabets and lexicons to grammar and phonetics, it remains largely unknown how the semantics (i.e., functions) vary across different contexts (locations of sequences). Interpretability is therefore critical to help find important patterns in sequences and understand their relationship within contexts.[93]

Legal and ethical regulations

Despite successful applications of language models in healthcare and biomedicine, there are some concerns about legal and ethical issues due to the potential risks posed by the models. Practical or actionable principles/guidelines of AI ethics have also been raised to address the issues.[158–161] For example, regarding safety, predictions produced by the models must be factually accurate with established knowledge and defer to an expert when uncertain.[162] For the privacy of health data, the use of patient health data must observe regulations, such as HIPAA in the US. For fairness, language models can create unfair discrimination and representation due to existing social inequalities. On the one hand, it must ensure that the training and evaluation data for language models are sufficiently representative of different sexes, races, and socio-economic backgrounds. On the other hand, debiasing methods are needed to ensure fairness when data are extremely imbalanced and scarce. Nevertheless, the interpretability of the model is still essential in healthcare and biomedicine since it provides evidence and logical steps for decision-making. It enables to detect the risks of harm in the model and avoid users overestimating the capabilities of the model. Tracing a given output or harm to its origins in the model can be key to resolving such harms. Although it remains an open challenge to define what constitutes a good explanation, various researchers have suggested the interpretability of language models is critical to ensure these systems are fair, ethical, and safe.[163]

ACKNOWLEDGMENTS

The work was supported by the U.S. National Science Foundation under Grant Numbers 1750326 and 2212175 and the U.S. National Institutes of Health under Grant Numbers R01MH124740, R01AG076448, R01AG080991, R01AG076234, RF1AG072449, and R01AG080624.

AUTHOR CONTRIBUTIONS

F.W. contributed to the conceptualization and reviewing and editing of the manuscript; D.L. contributed to the investigation, drafting, and reviewing and editing of the manuscript; X.W. contributed to the visualization and reviewing and editing of the manuscript; Y.C. contributed to reviewing and editing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training.
2. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
3. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog *1*, 9.
4. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Held in New Orleans, Louisiana (Association for Computational Linguistics), pp. 2227–2237.
5. Conneau, A., and Lample, G. (2019). Cross-lingual language model pretraining. Adv. Neural Inf. Process. Syst. *32*.
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. *21*, 5485–5551.
7. Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: masked sequence to sequence pre-training for language generation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.02450.

8. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., and Le, Q.V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. *32*.
9. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.13461.
10. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Ling. *8*, 726–742.
11. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1702.08608.
12. Meng, Y., Speier, W., Ong, M.K., and Arnold, C.W. (2021). Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J. Biomed. Health Inform. *25*, 3121–3129. https://doi.org/10.1109/jbhi.2021.3063721.
13. Shang, J., Ma, T., Xiao, C., and Sun, J. (2019). Pre-training of graph augmented transformers for medication recommendation. Preprint at arXiv. https://doi.org/10.24963/ijcai.2019/825.
14. Zhou, S., Wang, N., Wang, L., Liu, H., and Zhang, R. (2022). CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J. Am. Med. Inform. Assoc. *29*, 1208–1216.
15. Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: modeling clinical notes and predicting hospital readmission. Preprint at arXiv. https://doi.org/10.48550/arXiv.1904.05342.
16. Jin, D., Jin, Z., Zhou, J.T., and Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment, *34*, pp. 8018–8025.
17. Zhang, Y., Nie, A., Zehnder, A., Page, R.L., and Zou, J. (2019). VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. NPJ Digit. Med. *2*, 35.
18. Liu, S., Wang, X., Hou, Y., Li, G., Wang, H., Xu, H., Xiang, Y., and Tang, B. (2023). Multimodal data matters: language model pre-training over structured and unstructured electronic health records. IEEE J. Biomed. Health Inform. *27*, 504–514.
19. Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. J. Am. Med. Inform. Assoc. *26*, 1297–1304.
20. Zhu, H., Paschalidis, I.C., and Tahmasebi, A. (2018). Clinical concept extraction with contextual word embedding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.10566.
21. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. Preprint at arXiv. https://doi.org/10.48550/arXiv.1904.03323.
22. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit. Med. *4*, 86.
23. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). BEHRT: transformer for electronic health records. Sci. Rep. *10*, 7155.
24. Lewis, P., Ott, M., Du, J., and Stoyanov, V. (2020). Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-Of-The-Art, pp. 146–157.
25. Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.05474.
26. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large Language Models Are Few-Shot Clinical Information Extractors, pp. 1998–2022.
27. Chang, D., Hong, W.S., and Taylor, R.A. (2020). Generating contextual embeddings for emergency department chief complaints. JAMIA Open *3*, 160–166.
28. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., et al. (2022). A large language model for electronic health records. NPJ Digit. Med. *5*, 194.
29. Huang, K., Singh, A., Chen, S., Moseley, E.T., Deng, C.-Y., George, N., and Lindvall, C. (2019). Clinical xlnet: modeling sequential clinical notes and predicting prolonged mechanical ventilation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.11975.
30. Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., and Wong, A. (2020). Umlsbert: clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. Preprint at arXiv. https://doi.org/10.18653/v1/2021.naacl-main.139.
31. Kades, K., Sellner, J., Koehler, G., Full, P.M., Lai, T.Y.E., Kleesiek, J., and Maier-Hein, K.H. (2021). Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR Med. Inf. *9*, e22795.
32. Yang, X., Bian, J., Hogan, W.R., and Wu, Y. (2020). Clinical concept extraction using transformers. J. Am. Med. Inform. Assoc. *27*, 1935–1942.
33. Chen, Y.-P., Chen, Y.-Y., Lin, J.-J., Huang, C.-H., and Lai, F. (2020). Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med. Inf. *8*, e17787.
34. Wang, J., Zhang, G., Wang, W., Zhang, K., and Sheng, Y. (2021). Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J. Cloud Comput. *10*, 1–12.
35. Zhang, N., Jia, Q., Yin, K., Dong, L., Gao, F., and Hua, N. (2020). Conceptualized representation learning for chinese biomedical text mining. Preprint at arXiv. https://doi.org/10.48550/arXiv.2008.10813.
36. Kraljevic, Z., Shek, A., Bean, D., Bendayan, R., Teo, J., and Dobson, R. (2021). MedGPT: medical concept prediction from clinical narratives. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.03134.
37. Khin, K., Burckhardt, P., and Padman, R. (2018). A deep learning architecture for de-identification of patient notes: implementation and evaluation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.01570.
38. Yang, X., He, X., Zhang, H., Ma, Y., Bian, J., and Wu, Y. (2020). Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models. JMIR Med. Inf. *8*, e19735.
39. Xiong, Y., Chen, S., Chen, Q., Yan, J., and Tang, B. (2020). Using character-level and entity-level representations to enhance bidirectional encoder representation from transformers-based clinical semantic textual similarity model: ClinicalSTS modeling study. JMIR Med. Inf. *8*, e23357.
40. Mahajan, D., Poddar, A., Liang, J.J., Lin, Y.-T., Prager, J.M., Suryanarayanan, P., Raghavan, P., and Tsou, C.-H. (2020). Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning. JMIR Med. Inf. *8*, e22508.
41. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., and Hsu, C.-N. (2022). RadBERT: Adapting transformer-based language models to radiology. Radiol. Artif. Intell. *4*, e210258.
42. Lau, W., Lybarger, K., Gunn, M.L., and Yetisgen, M. (2023). Event-based clinical finding extraction from radiology reports with pre-trained language model. J. Digit. Imaging *36*, 91–104.
43. Meng, X., Ganoe, C.H., Sieberg, R.T., Cheung, Y.Y., and Hassanpour, S. (2020). Self-supervised contextual language representation of radiology reports to improve the identification of communication urgency. AMIA Jt. Summits Transl. Sci. Proc. *2020*, 413–421.
44. Bressem, K.K., Adams, L.C., Gaudin, R.A., Tröltzsch, D., Hamm, B., Makowski, M.R., Schüle, C.Y., Vahldiek, J.L., and Niehues, S.M. (2020). Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics *36*, 5255–5261.
45. Naseem, U., Khushi, M., and Kim, J. (2022). Vision-language transformer for interpretable pathology visual question answering. IEEE J. Biomed. Health Inform. *27*, 1681–1690.
46. Li, Y., Wang, H., and Luo, Y. (2020). A Comparison of Pre-trained Vision-And-Language Models for Multimodal Representation Learning across Medical Images and Reports (IEEE), pp. 1999–2004.
47. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., and Jawahar, C. (2021). MMBERT: Multimodal BERT Pretraining for Improved Medical VQA, pp. 1033–1036.
48. Moon, J.H., Lee, H., Shin, W., Kim, Y.-H., and Choi, E. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE J. Biomed. Health Inform. *26*, 6070–6080.
49. Chen, Z., Li, G., and Wan, X. (2022). Align, Reason and Learn: Enhancing Medical Vision-And-Language Pre-training with Knowledge, pp. 5152–5161.
50. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., and Chang, T.-H. (2022). Multi-modal Masked Autoencoders for Medical

Vision-And-Language Pre-training (Springer), pp. 679–689.

51. Monajatipoor, M., Rouhsedaghat, M., Li, L.H., Jay Kuo, C.-C., Chien, A., and Chang, K.-W. (2022). Berthop: An Effective Vision-And-Language Model for Chest X-Ray Disease Diagnosis (Springer), pp. 725–734.
52. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., and Alvarez-Valle, J. (2022). Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing (Springer), pp. 1–21.
53. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240.
54. Shin, H.-C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., and Mani, R. (2020). Biomegatron: larger biomedical domain language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.06060.
55. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. *3*, 1–23.
56. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief. Bioinform. *23*, bbac409.
57. Kanakarajan, K.R., Kundumani, B., and Sankarasubbu, M. (2021). BioELECTRA: Pretrained Biomedical Text Encoder Using Discriminators, pp. 143–154.
58. Yasunaga, M., Leskovec, J., and Liang, P. (2022). Linkbert: pretraining language models with document links. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.15827.
59. Miolo, G., Mantoan, G., and Orsenigo, C. (2021). Electramed: A new pre-trained language representation model for biomedical nlp. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.09585.
60. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: a large language model for science. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.09085.
61. Jin, Q., Dhingra, B., Cohen, W.W., and Lu, X. (2019). Probing biomedical embeddings from language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.1904.02181.
62. Naseem, U., Dunn, A.G., Khushi, M., and Kim, J. (2022). Benchmarking for biomedical natural language processing tasks with a domain specific albert. BMC Bioinf. *23*, 144.
63. Yuan, Z., Liu, Y., Tan, C., Huang, S., and Huang, F. (2021). Improving biomedical pretrained language models with knowledge. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.10344.
64. Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. Patterns *4*, 100729.
65. Ozyurt, I.B. (2020). On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. Preprint at bioRxiv. https://doi.org/10.18653/v1/2020.sdp-1.12.
66. Moradi, M., Dorffner, G., and Samwald, M. (2020). Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. Comput. Methods Programs Biomed. *184*, 105117.
67. Xie, Q., Bishop, J.A., Tiwari, P., and Ananiadou, S. (2022). Pre-trained language models with domain knowledge for biomedical extractive summarization. Knowl. Base Syst. *252*, 109460.
68. Du, Y., Li, Q., Wang, L., and He, Y. (2020). Biomedical-domain pre-trained language model for extractive summarization. Knowl. Base Syst. *199*, 105964.
69. Wallace, B.C., Saha, S., Soboczenski, F., and Marshall, I.J. (2021). Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. AMIA Jt. Summits Transl. Sci. Proc. *2021*, 605–614.
70. Guo, Y., Qiu, W., Wang, Y., and Cohen, T. (2021). Automated Lay Language Summarization of Biomedical Scientific Reviews, *1*, pp. 160–168.
71. Kieuvongngam, V., Tan, B., and Niu, Y. (2020). Automatic text summarization of covid-19 medical research articles using bert and gpt-2. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.01997.
72. Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., and Mosconi, F. (2020). BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR, pp. 669–679.
73. Oniani, D., and Wang, Y. (2020). A Qualitative Evaluation of Language Models on Automatic Question-Answering for Covid-19, pp. 1–9.
74. Liévin, V., Hother, C.E., and Winther, O. (2022). Can large language models reason about medical questions?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2207.08143.
75. He, Y., Zhu, Z., Zhang, Y., Chen, Q., and Caverlee, J. (2020). Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.03746.
76. Hao, B., Zhu, H., and Paschalidis, I.C. (2020). Enhancing Clinical Bert Embedding Using a Biomedical Knowledge Base.
77. Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2020). Self-alignment pretraining for biomedical entity representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11784.
78. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., and Pfohl, S. (2022). Large language models encode clinical knowledge. Preprint at arXiv. https://doi.org/10.48550/arXiv.2212.13138.
79. Naseem, U., Lee, B.C., Khushi, M., Kim, J., and Dunn, A.G. (2022). Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. Preprint at arXiv. https://doi.org/10.18653/v1/2022.nlppower-1.3.
80. Müller, M., Salathé, M., and Kummervold, P.E. (2023). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. Front. Artif. Intell. *6*, 1023281.
81. Tutubalina, E., Alimova, I., Miftahutdinov, Z., Sakhovskiy, A., Malykh, V., and Nikolenko, S. (2021). The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. Bioinformatics *37*, 243–249.
82. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2021). Mentalbert: publicly available pretrained language models for mental healthcare. Preprint at arXiv. https://doi.org/10.48550/arXiv.2110.15621.
83. Papanikolaou, Y., and Pierleoni, A. (2020). Dare: Data augmented relation extraction with gpt-2. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.13845.
84. Papanikolaou, Y., Roberts, I., and Pierleoni, A. (2019). Deep bidirectional transformers for relation extraction without supervision. Preprint at arXiv. https://doi.org/10.48550/arXiv.1911.00313.
85. Wang, D., Hu, W., Cao, E., and Sun, W. (2020). Global-to-local neural networks for document-level relation extraction. Preprint at arXiv. https://doi.org/10.48550/arXiv.2009.10359.
86. Cabot, P.-L.H., and Navigli, R. (2021). REBEL: Relation Extraction by End-To-End Language Generation, pp. 2370–2381.
87. Weber, L., Sänger, M., Garda, S., Barth, F., Alt, C., and Leser, U. (2022). Chemical–protein relation extraction with ensembles of carefully tuned pretrained language models. Database *2022*, baac098.
88. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinf. *20*, 1–17.
89. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA *118*. e2016239118.
90. Xiao, Y., Qiu, J., Li, Z., Hsieh, C.-Y., and Tang, J. (2021). Modeling protein using large-scale pretrain language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2108.07435.
91. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110.
92. Weissenow, K., Heinzinger, M., and Rost, B. (2022). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure *30*, 1169–1177.e4.
93. Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics *37*, 2112–2120.
94. Yamada, K., and Hamada, M. (2022). Prediction of RNA–protein interactions using a nucleotide language model. Bioinform. Adv. *2*, vbac023.
95. Mock, F., Kretschmer, F., Kriese, A., Böcker, S., and Marz, M. (2022). Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. Proc. Natl. Acad. Sci. USA *119*. e2122636119.
96. Heinzinger, M., Weissenow, K., Sanchez, J.G., Henkel, A., Steinegger, M., and Rost, B. (2023). ProstT5: Bilingual language model for protein sequence and structure. Preprint at bioRxiv. https://doi.org/10.1101/2023.07.23.550085.
97. Danilov, G., Kotik, K., Shevchenko, E., Usachev, D., Shifrin, M., Strunina, Y., Tsukanova, T., Ishankulov, T., Lukshin, V., and Potapov, A. (2022). Predicting the

length of stay in neurosurgery with RuGPT-3 language model. Stud. Health Technol. Inform. *295*, 555–558.

98. Wang, M., Chen, J., and Lin, S. (2021). Medication recommendation based on a knowledge-enhanced pre-training model, pp. 290–294.
99. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., and Yu, L. (2022). Multi-granularity cross-modal alignment for generalized medical visual representation learning. Adv. Neural Inf. Process. Syst. *35*, 33536–33549.
100. Kaur, N., and Mittal, A. (2022). RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings. J. Biomed. Inform. *135*, 104220.
101. Zhang, L., Fan, H., Peng, C., Rao, G., and Cong, Q. (2020). Sentiment Analysis Methods for HPV Vaccines Related Tweets Based on Transfer Learning, *3* (MDPI), p. 307.
102. Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., and Kim, J. (2021). Bioalbert: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition (IEEE), pp. 1–7.
103. Jain, S., and Wallace, B.C. (2019). Attention is not explanation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1902.10186.
104. Wiegreffe, S., and Pinter, Y. (2019). Attention is not not explanation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.04626.
105. Hao, Y., Dong, L., Wei, F., and Xu, K. (2021). Self-attention attribution: Interpreting information interactions inside transformer, *35*, pp. 12963–12971.
106. Córdova Sáenz, C.A., and Becker, K. (2021). Assessing the Use of Attention Weights to Interpret BERT-Based Stance Classification, pp. 194–201.
107. Shi, T., Zhang, X., Wang, P., and Reddy, C.K. (2021). Corpus-level and concept-based explanations for interpretable document classification. ACM Trans. Knowl. Discov. Data *16*, 1–17. Article 48. https://doi.org/10.1145/3477539.
108. Chrysostomou, G., and Aletras, N. (2021). Improving the faithfulness of attention-based explanations with task-specific information for text classification. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.02657.
109. Bacco, L., Cimino, A., Dell'Orletta, F., and Merone, M. (2021). Explainable sentiment analysis: a hierarchical transformer-based extractive summarization approach. Electronics *10*, 2195.
110. Niu, S., Yin, Q., Song, Y., Guo, Y., and Yang, X. (2021). Label dependent attention model for disease risk prediction using multimodal electronic health records, pp. 449–458.
111. Tutek, M., and Šnajder, J. (2022). Toward practical usage of the attention mechanism as a tool for interpretability. IEEE Access *10*, 47011–47030. https://doi.org/10.1109/ACCESS.2022.3169772.
112. Liu, D., Greene, D., and Dong, R. (2022). A novel perspective to look at attention: bi-level attention-based explainable topic modeling for news classification. Preprint at arXiv. https://doi.org/10.18653/v1/2022.findings-acl.178.
113. Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., and Scotton, P. (2021). Attention-based Interpretability with Concept Transformers.
114. Attanasio, G., Nozza, D., Pastor, E., and Hovy, D. (2022). Benchmarking Post-hoc Interpretability Approaches for Transformer-Based Misogyny Detection (Association for Computational Linguistics).
115. Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2015). Visualizing and understanding neural models in NLP. Preprint at arXiv. https://doi.org/10.48550/arXiv.1506.01066.
116. Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.0473.
117. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.05695.
118. Xie, Q., Ma, X., Dai, Z., and Hovy, E. (2017). An interpretable knowledge transfer model for knowledge base completion. Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.05908.
119. Ding, S., and Koehn, P. (2021). Evaluating saliency methods for neural language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2104.05824.
120. Malkiel, I., Ginzburg, D., Barkan, O., Caciularu, A., Weill, J., and Koenigstein, N. (2022). Interpreting BERT-based Text Similarity via Activation and Saliency Maps. In Proceedings of the ACM Web Conference 2022 (Association for Computing Machinery).
121. Rajani, N.F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.02361.
122. Abujabal, A., Roy, R.S., Yahya, M., and Weikum, G. (2017). Quint: Interpretable Question Answering over Knowledge Bases, pp. 61–66.
123. Brand, E., Roitero, K., Soprano, M., Rahimi, A., and Demartini, G. (2022). A neural model to jointly predict and explain truthfulness of statements. J. Data Inf. Qual. *15*, 1–19. Article 4. https://doi.org/10.1145/3546917.
124. Sammani, F., Mukherjee, T., and Deligiannis, N. (2022). NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks, pp. 8312–8322.
125. Dunn, A., Inkpen, D., and Andonie, R. (2021). Context-Sensitive Visualization of Deep Learning Natural Language Processing Models (IEEE), pp. 170–175.
126. Li, Z., Wang, X., Yang, W., Wu, J., Zhang, Z., Liu, Z., Sun, M., Zhang, H., and Liu, S. (2022). A unified understanding of deep NLP models for text classification. IEEE Trans. Vis. Comput. Graph. *28*, 4980–4994. https://doi.org/10.1109/TVCG.2022.3184186.
127. Aflalo, E., Du, M., Tseng, S.Y., Liu, Y., Wu, C., Duan, N., and Lal, V. (2022). VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers, pp. 21374–21383.
128. Yan, X., Jian, F., and Sun, B. (2021). SAKG-BERT: enabling language representation with knowledge graphs for chinese sentiment analysis. IEEE Access *9*, 101695–101701. https://doi.org/10.1109/ACCESS.2021.3098180.
129. Islam, S.M., and Bhattacharya, S. (2022). AR-BERT: aspect-relation enhanced aspect-level sentiment classification with multi-modal explanations. In Proceedings of the ACM Web Conference 2022 (Association for Computing Machinery). https://doi.org/10.48550/arXiv.2108.11656.
130. Feder, A., Oved, N., Shalit, U., and Reichart, R. (2021). Causalm: Causal model explanation through counterfactual language models. Comput. Ling. *47*, 333–386.
131. Taylor, N., Sha, L., Joyce, D.W., Lukasiewicz, T., Nevado-Holgado, A., and Kormilitzin, A. (2021). Rationale production to support clinical decision-making. Preprint at arXiv. https://doi.org/10.48550/arXiv.2111.07611.
132. Li, D., Hu, B., Chen, Q., Xu, T., Tao, J., and Zhang, Y. (2022). Unifying model explainability and robustness for joint text classification and rationale extraction, *36*, pp. 10947–10955.
133. Creswell, A., Shanahan, M., and Higgins, I. (2022). Selection-inference: exploiting large language models for interpretable logical reasoning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.09712.
134. Poerner, N., Roth, B., and Schütze, H. (2018). Evaluating neural network explanation methods using hybrid documents and morphological agreement. Preprint at arXiv. https://doi.org/10.18653/v1/P18-1032.
135. Croce, D., Rossini, D., and Basili, R. (2018). Explaining Non-linear Classifier Decisions within Kernel-Based Deep Architectures, pp. 16–24.
136. Aken, B.v., Winter, B., Löser, A., and Gers, F.A. (2019). How does BERT answer questions? a layer-wise analysis of transformer representations. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (Association for Computing Machinery). https://doi.org/10.1145/3357384.3358028.
137. Aken, B.v., Winter, B., Löser, A., and Gers, F.A. (2020). VisBERT: hidden-state visualizations for transformers. In Companion Proceedings of the Web Conference 2020 (Association for Computing Machinery). https://doi.org/10.48550/arXiv.2011.04507.
138. Sevastjanova, R., Kalouli, A.-L., Beck, C., Schäfer, H., and El-Assady, M. (2021). Explaining Contextualization in Language Models Using Visual Analytics, pp. 464–476.
139. Janizek, J.D., Sturmfels, P., and Lee, S.-I. (2021). Explaining explanations: axiomatic feature interactions for deep networks. J. Mach. Learn. Res. *22*. Article 104.
140. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One *10*, e0130140.
141. Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: learning important features through propagating activation differences. Preprint at arXiv. https://doi.org/10.48550/arXiv.1605.01713.
142. Feng, S., Wallace, E., Grissom, A., II, Iyyer, M., Rodriguez, P., and Boyd-Graber, J. (2018). Pathologies of neural models make interpretations difficult. Preprint at arXiv. https://doi.org/10.18653/v1/D18-1407.
143. Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile, *33*, pp. 3681–3688.
144. Martins, A., and Astudillo, R. (2016). From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification (PMLR), pp. 1614–1623.

145. Kaushik, D., Hovy, E., and Lipton, Z.C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.12434.

146. Abraham, E.D., D'Oosterlinck, K., Feder, A., Gat, Y., Geiger, A., Potts, C., Reichart, R., and Wu, Z. (2022). CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. Adv. Neural Inf. Process. Syst. *35*, 17582–17596.

147. Basu, S., Pope, P., and Feizi, S. (2020). Influence functions in deep learning are fragile. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.14651.

148. Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. Trans. Assoc. Comput. Ling. *9*, 160–175.

149. Wallace, E., Gardner, M., and Singh, S. (2020). Interpreting Predictions of NLP Models, pp. 20–23.

150. De Cao, N., Schlichtkrull, M., Aziz, W., and Titov, I. (2020). How do decisions emerge across layers in neural models? interpretation with differentiable masking. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.14651.

151. Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods, pp. 180–186.

152. Kovalerchuk, B., Ahmad, M.A., and Teredesai, A. (2021). Survey of Explainable Machine Learning with Visual and Granular Methods beyond Quasi-Explanations. Interpretable Artificial Intelligence: A Perspective of Granular Computing, pp. 217–267.

153. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., and Wallace, B.C. (2019). ERASER: A benchmark to evaluate rationalized NLP models. Preprint at arXiv. https://doi.org/10.48550/arXiv.1911.03429.

154. Jacovi, A., and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.03685.

155. Weerts, H.J., van Ipenburg, W., and Pechenizkiy, M. (2019). A human-grounded evaluation of shap for alert processing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.03324.

156. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., and Eckersley, P. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery). https://doi.org/10.48550/arXiv.1909.06342.

157. Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., and Müller, H. (2023). AI for life: Trends in artificial intelligence for biotechnology. N. Biotechnol. *74*, 16–24.

158. Muller, H., Mayrhofer, M.T., Van Veen, E.-B., and Holzinger, A. (2021). The ten commandments of ethical medical AI. Computer *54*, 119–123.

159. Kargl, M., Plass, M., and Müller, H. (2022). A literature review on ethics for AI in biomedical research and biobanking. Yearb. Med. Inform. *31*, 152–160.

160. Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., and Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. N. Biotechnol. *70*, 67–72.

161. Zhou, J., Müller, H., Holzinger, A., and Chen, F. (2023). Ethical ChatGPT: concerns, challenges, and commandments. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.10646.

162. Mozannar, H., and Sontag, D. (2020). Consistent Estimators for Learning to Defer to an Expert (PMLR), pp. 7076–7087.

163. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., and Kasirzadeh, A. (2021). Ethical and social risks of harm from language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.04359.