Supplementary Table S 1: Summary of Included Studies

Study and Year	Country of Affiliation	Objective	LLMs	Medical Specialty	Key Findings	Evaluation
Scquizzato et al., 2024	Italy, UK	To assess the accuracy, reliability, comprehensiveness, and overall perceived value of answers provided by ChatGPT to a list of frequently asked questions in the fields of cardiac arrest and CPR.	ChatGPT	Cardiology	ChatGPT provided largely accurate, relevant, and comprehensive answers to questions about cardiac arrest. CPR-related answers received the lowest scores. Laypeople rated the answers higher than professionals.	14 Professionals (clinicians and researchers), 16 laypeople (survivors, lay rescuers and relatives)
Lee et al., 2024	Republic of Korea	To develop and validate LLMs for automatic clinical diagnoses by identifying medical symptoms and patient histories from simulated clinician-patient conversations in emergency triage rooms, and to perform XAI on the developed model and evaluate its reliability.	BERT, KLUE-RoBERTa, KLUE-BERT, KorBERT, KoBERT	Emergency Medicine	Fine-tuned LLMs, particularly KLUE-RoBERTa, can effectively classify medical symptoms and histories from clinician-patient conversations in Korean. XAI using SHAP provides reliable explanations, comparable to those of medical workers, as confirmed by a Turing test.	10 Emergency medicine specialists and 31 medical workers (nurses, medical residents, emergency medical technicians)
Chen et al., 2021	Taiwan	To create a model to extract concept embeddings from EHRs for disease pattern retrieval and further classification tasks.	BERT	Emergency Medicine	Disease concepts can be embedded meaningfully using contrastive learning methods. The disease concept model (EDisease) is suitable as a pretrained model for subsequent prediction tasks and disease retrieval, enhancing clinical practice capabilities.	1 seasoned doctor
Nimmi et al., 2022	India	To examine call content and classify calls based on emotion, and to develop an ensemble deep learning model for emotion detection using COVID-19 related emergency response support system calls and tweets.	BERT, DistilBERT, RoBERTa	Emergency Medicine	The proposed Average Voting Ensemble Deep Learning (AVEDL) model, combining BERT, DistilBERT, and RoBERTa, outperforms standard deep learning and machine learning models for emotion detection in emergency response	Pre-established ground truth (one PhD scholar and two postgraduate students)

					calls and tweets related to COVID-19.	
Gan et al., 2024	Spain, Cyprus, Malaysia	To evaluate and compare the performance of ChatGPT, Google Bard, and medical students in performing START triage during mass casualty incidents.	ChatGPT-3.5, Google Bard	Emergency Medicine	Google Bard was superior to ChatGPT in performing mass casualty incident triage using the START method. The difference in accuracy was statistically significant.	Pre-established ground truth (Emergency Medical Services Medical Directors)
Zaboli et al., 2024	Italy	To assess the effectiveness of AI (Chat-GPT) in patient stratification according to the Manchester Triage System (MTS) compared to risk stratification by healthcare professionals.	Chat-GPT 3.5	Emergency Medicine	The current reliability of Chat-GPT is insufficient to replace triage nurses in prioritizing emergency department patients. Chat-GPT performs worse than healthcare professionals in terms of clinical outcomes and has a higher tendency to over-triage.	Pre-established ground truth (Two emergency medicine nurses and one physician)
Bushuven et al., 2023	Germany	To test the performance and safety of ChatGPT and GPT-4 in identifying emergencies, providing correct diagnoses, and advising on further actions to lay rescuers in pediatric and basic life support case vignettes.	ChatGPT, GPT-4	Emergency Medicine	ChatGPT and GPT-4 showed high diagnostic accuracy in pediatric emergency scenarios but were inconsistent in advising emergency calls and providing correct first aid. Their reliability and safety as emergency support tools for lay rescuers are questionable.	5 physicians (Critical care medicine, emergency medicine, and anesthesiology)
Campbell et al., 2024	USA	To evaluate the quality of ChatGPT-generated responses for patient education on thyroid nodules.	GPT-3.5, GPT-4	Endocrinology	ChatGPT provides appropriate answers to most questions on thyroid nodules regardless of prompting. Responses are generated at grade levels well-above accepted recommendations for patients. Significant rates of AI hallucination may preclude recommending ChatGPT as an educational tool.	Otolaryngology resident physicians, experienced thyroid and parathyroid head-and-neck-attending surgeon)
Bellinger et al., 2023	USA	To quantitatively compare online	ChatGPT	ENT	Information on ChatGPT regarding	Pre-established ground truth (2 experts

		patient education materials found using traditional search engines (Google) versus conversational Artificial Intelligence (AI) models (ChatGPT) for benign paroxysmal positional vertigo (BPPV).			BPPV is more difficult to read, of lower quality, and more difficult to comprehend compared to information obtained from Google searches.	otolaryngologists with subspecialty certification in Neurotology, and a second-year fellow in Otology/Neurotology))
Abhilash et al., 2023	India	To present a novel approach for uncovering interesting insights in large datasets using ontologies and BERT models, specifically in the healthcare sector during the COVID-19 pandemic.	BERT, ClinicalBERT, BlueBERT, CovidBERT, BioClinicalBERT	Epidemiology	An ontology-based approach combined with BERT models can effectively extract semantically rich and interesting rules from COVID-19 data. BioClinicalBERT provided more detailed and semantically richer information compared to CovidBERT. The identified rules show high correlation with CDC listed symptoms and domain expert evaluations.	Domain experts
Zhang et al., 2022	China, UK, USA	To propose an intelligent COVID-19 early warning system using Twitter data with novel machine learning methods, specifically fine-tuning BERT for Twitter classification and a Twitter-based linear regression model for forecasting.	BERT	Epidemiology	Fine-tuning BERT outperforms other text classification methods for COVID-19 related tweets. A Twitter-based linear regression model can effectively forecast COVID-19 cases and provide early warning signals.	Pre-established ground truth (2 authors of the study)
Alkouz et al., 2022	United Arab Emirates, USA	To develop a deep learning model (Deepluenza) to accurately identify influenza reporting tweets in multiple languages (English and Arabic) and to use it for early detection of influenza outbreaks.	BERT base multilingual	Epidemiology	Deepluenza, based on the BERT base multilingual model, achieved high accuracy and F1-score for identifying influenza-reporting tweets. Combining English and Arabic tweets improved the correlation between tweet counts and hospital visits, suggesting its potential for early	2 Authors of the study

					influenza outbreak detection.	
Lu et al., 2021	USA	To develop a natural language processing method for answering factoid questions related to COVID-19 using published articles as knowledge sources.	BERT	Epidemiology	The proposed method successfully extracted and aggregated opinions from the literature to answer COVID-19 related questions. The trends in opinions and popular words reflect known characteristics and challenges of COVID-19, such as higher risk for seniors and those with pre-existing conditions, high contagion, and the need for screening and testing.	N/A
Klein et al., 2021	USA	To develop, evaluate, and deploy an automatic natural language processing pipeline to collect user-generated Twitter data as a complementary resource for identifying potential cases of COVID-19 in the United States that are not based on testing and, thus, may not have been reported to the Centers for Disease Control and Prevention.	BERT-Base-Uncased, COVID-Twitter-BERT	Epidemiology	An NLP pipeline using COVID-Twitter-BERT achieved an F1-score of 0.76 for detecting tweets self-reporting potential COVID-19 cases. The pipeline identified 13,714 tweets across the US, potentially unreported to the CDC.	Pre-established ground truth (2 Researchers)
Fu et al., 2022	USA	To develop and evaluate a hybrid model, combining a context-aware language model (BERT) with post-hoc heuristic rules, to identify fall events from EHR free text, and compare its performance to conventional rule-based and deep learning models.	BERT, ClinicalBERT, XLNet, XLMs, Albert	Geriatrics	A context-aware language model (BERT) effectively identified challenging fall events in EHR text. A hybrid model combining BERT with post-hoc rules further improved performance, achieving the highest f1-scores across sentence, document, and patient levels.	Pre-established ground truth (Two experts (geriatrician and palliative care physician) and two trained abstractors)
Zhang et al., 2022	China	To develop a feature engineering algorithm (HLAB) for predicting class I	ProtBert	Immunology	The proposed HLAB framework, utilizing ProtBert and BiLSTM for feature	Automated matrics

		HLA-binding peptides using ProtBert-encoded features and BiLSTM.			extraction, outperforms existing methods for predicting HLA-I binding peptides. Feature selection is crucial for optimal performance.	
Homburg et al., 2023	Netherlands	To develop and validate a BERT model for detecting COVID-19 consultations in general practice EHRs in the Netherlands.	BERT	Infectious Diseases	A BERT model was developed and validated to accurately identify COVID-19 consultations in Dutch general practice EHRs. The model demonstrated high performance on test and external validation sets and showed a significant correlation with COVID-19 hospitalizations, even predicting cases before the first confirmed case in the Netherlands.	Pre-established ground truth (2 researchers: A general practitioner and a microbiologist)
Maillard et al., 2024	France, Italy	To assess the quality and safety of the management suggested by ChatGPT-4 in real-life practice for patients with positive blood cultures, compared to infectious diseases consultants.	ChatGPT-4	Infectious Diseases	ChatGPT-4 provided detailed responses, but its management plans for bloodstream infections were often suboptimal and sometimes harmful. The use of ChatGPT-4 without consultant input remains hazardous in 2023, especially for severe infectious diseases.	Two senior ID physicians not involved in patient management
Wang et al., 2023	China	To evaluate the accuracy of COVID-19 information generated by ChatGPT 3.5 and 4.0, assessing its potential as a supplementary public information source during the pandemic.	ChatGPT 3.5, ChatGPT 4.0	Infectious Diseases	ChatGPT 3.5 and 4.0 can generate relatively accurate and relevant COVID-19 information. ChatGPT 4.0 significantly outperformed ChatGPT 3.5 in all evaluation dimensions. However, gaps and deficiencies exist compared to official WHO responses, particularly in areas requiring up-to-date knowledge or ethical considerations.	2 clinicians

Laison et al., 2023	Canada, Egypt, USA	To analyze global Twitter data to understand its potential and limitations as a tool for Lyme disease surveillance and propose a transformer-based classification system to identify potential Lyme disease cases using self-reported tweets.	DistilBERT, ALBERT, BERTweet	Infectious Diseases	BERTweet was the best classifier for identifying potential Lyme disease cases from self-reported tweets. Including emojis as sentiment words improved the performance of all transformer-based classifiers, particularly increasing recall.	Automated matrics
Wang et al., 2023	China	To propose a new model (BioDGW-CMI) for predicting circRNA-miRNA interaction (CMI) by combining biological text mining and wavelet diffusion-based sparse network structure embedding.	BERT, BioBERT	Medical Genetics and Genomics.	BioDGW-CMI, combining biological text mining and wavelet diffusion-based sparse network structure embedding, achieves high prediction performance for circRNA-miRNA interactions across multiple datasets. The model outperforms existing methods, demonstrating the effectiveness of BioBERT and GraphWave for this task.	clinicians and non-clinicians
Haze et al., 2023	Japan	To examine the accuracy and consistency of ChatGPT (GPT-3.5 and GPT-4) on the Japanese National Medical Examination and to assess the relationship between the accuracy in different medical fields and the amount of information published in those fields.	GPT-3.5, GPT-4	Multi-specialty	GPT-4 significantly outperformed GPT-3.5 on the Japanese National Medical Examination. Accuracy varied by medical field and was correlated with the number of publications in that field. Inconsistency in answers was associated with a higher likelihood of incorrect answers.	Automated matrics
Kim et al., 2023	Republic of Korea	To propose a medical specialty prediction model from patient-side medical question text based on pre-trained bidirectional encoder representations from transformers (BERT).	KM-BERT, KR-BERT, M-BERT	Multi-specialty	The proposed KM-BERT model, pre-trained on a Korean medical corpus, outperformed other BERT-based and conventional deep learning models in predicting medical specialties from patient-generated	Automated matrics

					question text. The model demonstrated higher accuracy, precision, recall, and F1-scores.	
Qiao et al., 2022	China	To develop a novel predictor (BERT-Kcr) for protein lysine crotonylation (Kcr) sites prediction using a transfer learning method with pre-trained BERT models.	BERT	Multi-specialty	BERT-Kcr, a transfer learning method using pre-trained BERT models, outperforms existing methods for predicting lysine crotonylation sites. The direct use of sequence information with advanced pre-trained NLP models is effective for identifying PTM sites.	Automated matrics
Liu et al., 2023	USA	To determine if ChatGPT can generate useful suggestions for improving clinical decision support (CDS) logic and to assess noninferiority compared to human-generated suggestions.	ChatGPT, GPT-3.5	Multi-specialty	ChatGPT can generate suggestions for improving clinical decision support (CDS) alert logic. AI-generated suggestions were understandable and relevant but less useful and acceptable than human-generated suggestions. AI suggestions can complement traditional CDS optimization.	5 Clinicians experts in clinical decision support (Trained in internal medicine, anesthesiology, pharmacy, emergency medicine, and pediatrics)
Wang et al., 2023	China, USA	To explore the performance of ChatGPT for primary screening of mild cognitive impairment (MCI) and standardize the design steps and components of the prompts.	GPT-4, GPT-3.5	Multi-specialty	ChatGPT was effective in the primary screening of participants with possible MCI. Improved standardization of prompts by clinicians improved model performance. GPT-4 outperformed GPT-3.5.	Automated matrics
Karthikeyan et al., 2022	UK	To develop and evaluate a deep neural network-based OCR post-correction methodology using RoBERTa to improve the accuracy of text extracted from medical reports.	RoBERTa	Multi-specialty	The proposed OCR post-correction method using RoBERTa significantly reduced word and character error rates in both NHS medical reports and a publicly available dataset (MiBio).	Pre-established ground truth (Pharmacists)

Guo et al., 2022	China	To build a retrieval-based medical question answering system using large language models and knowledge graphs, and to evaluate its performance on Chinese and English datasets.	BERT	Multi-specialty	A new model based on PLM and Siamese structure improves MAP and MRR in answer selection tasks. Integrating knowledge graphs and named entity recognition enhances fine-grained ranking in a medical question answering system. The proposed system outperforms baseline models in both quantification and qualification evaluations.	3 Professional medical doctors
Mulyar et al., 2021	USA	To develop and evaluate a single deep learning model (MT-Clinical BERT) that simultaneously performs 8 clinical tasks spanning entity extraction, personal health information identification, language entailment, and similarity, by sharing representations among tasks.	BERT, Clinical BERT, MT-Clinical BERT, Bio+Clinical BERT	Multi-specialty	MT-Clinical BERT, a single deep learning model, concurrently addresses 8 clinical tasks spanning named entity recognition, entailment, and semantic similarity, achieving an 8-fold computational speed-up at inference compared to traditional per-task models. There is a slight but consistent performance degradation in MT-Clinical BERT relative to sequential fine-tuning.	Automated matrics
Li et al., 2021	USA	To systematically assess and compare four state-of-the-art text generation algorithms in the medical domain by applying them to the task of generating History and Present Illness (HPI) sections in discharge summaries, and to evaluate the utility of synthetic clinical notes generated by the best-performing algorithm for clinical named entity recognition (NER) tasks.	CharRNN, SegGAN, GPT-2, CTRL	Multi-specialty	Synthetic clinical notes generated by GPT-2 can be used to train NER models with performance comparable to models trained on natural clinical notes. Combining synthetic and natural data for training further improves NER model performance.	Pre-established ground truth (Experienced annotators)

Liu et al., 2023	Australia, China, Taiwan, Japan, United Arab Emirates	To develop a hybrid deidentification pipeline for Australian EHR text notes using rules and transformers, and to investigate the impact of pretrained word embedding and transformer-based language models.	BioBERT, Clinical BioBERT, Discharge Summary BioBERT	Multi-specialty	The OpenDeID pipeline, combining rules and transformers, achieved high performance (F1-score of 0.9659) in deidentifying SHI entities in Australian pathology reports. Fine-tuning pretrained BERT-base models outperformed neural networks with word embeddings. Cascaded preprocessing rules significantly improved performance.	Automated matrics
Mihalache et al., 2024	Canada	To assess the performance of ChatGPT-4 at responding to USMLE Step 1, Step 2CK, and Step 3 practice questions.	ChatGPT-4	Multi-specialty	ChatGPT-4 answered a remarkably high proportion of practice questions correctly for USMLE examinations, significantly outperforming previous models. The mean length of ChatGPT-4's correct responses was significantly shorter than incorrect responses.	2 Authors of the study
Liu et al., 2024	USA	To evaluate the capability of using generative artificial intelligence (AI) in summarizing alert comments and to determine if the AI-generated summary could be used to improve clinical decision support (CDS) alerts.	GPT-4	Multi-specialty	GPT-4 is capable of summarizing user comments on clinical decision support (CDS) alerts with high clarity, accuracy, and completeness. AI-generated summaries were equivalent and, in some aspects (completeness), superior to human-generated summaries, suggesting their utility in improving CDS alert systems.	Pre-established ground truth (2 physicians (1 internal medicine physician and 1 psychiatrist))
Wang et al., 2023	China	To evaluate ChatGPT's ability to understand medical knowledge in Chinese, and its potential as an electronic health	GPT-3.5, GPT-4	Multi-specialty	GPT-4 significantly outperforms GPT-3.5 in Chinese medical examinations and shows improved capabilities in discharge	3 doctors not involved in the preliminary study

		infrastructure for medical development, by assessing its performance in medical examinations, records, and education.			summarization and group learning. GPT-4 demonstrates high verbal fluency and reduces hallucinations compared to GPT-3.5. AI systems face challenges of hallucinations, legal risks, and ethical issues.	
Tsoutsanis et al., 2024	UK, Germany	To assess the capacity of four LLMs (Llama 2, Google Bard, Bing Chat, and ChatGPT-3.5) in answering practice questions for the Clinical Problem Solving (CPS) paper of the Multi-Specialty Recruitment Assessment (MSRA) exam and compare their performance to that of medical doctors preparing for the exam.	Llama 2, Google Bard, Bing Chat, ChatGPT-3.5	Multi-specialty	Bing Chat outperformed other LLMs and human users in answering MSRA exam questions. Llama 2 performed worse than humans. Google Bard and ChatGPT-3.5 performed similarly to human users. Performance varied by medical subject.	Medical doctors
Caruccio et al., 2024	Italy, India	To conduct a comparative study between traditional machine learning approaches, Natural Language Processing models, Google BARD, and ChatGPT for intelligent diagnosis, simulating a scenario of interaction with people.	ChatGPT, GPT-Neo, MedBot, Google BARD, Claude, Luminous	Multi-specialty	Google BARD and ChatGPT's text-davinci-003 engine achieved better accuracy compared to other models. Google BARD often outperformed ChatGPT in disease classification tasks. ChatGPT and similar AI models require careful interpretation and analysis of their responses. Traditional ML models outperformed LLMs in some cases.	Automated matrics
Walker et al., 2023	UK, Switzerland, Saudi Arabia	To assess the reliability of medical information provided by ChatGPT on hepato-pancreatico-biliary (HPB) conditions.	ChatGPT-4	Multi-specialty	ChatGPT provides medical information of comparable quality to static internet information, but with limitations in quality. Agreement with guideline recommendations was moderate.	2 authors of the study
Ji et al., 2021	Finland	To investigate the effectiveness of	BERT, BlueBERT,	Multi-specialty	Pretraining in mixed domains improves	Automated matrics

		various pretrained language models, trained in different domains, for medical code assignment from clinical notes, and to compare their performance against classical CNN models.	BioBERT, BioRedditBERT, PubMedBERT, SapBERT, ClinicalBERT		performance over general domains. Hierarchical fine-tuning architecture improves performance for long clinical notes. A retrained classical CNN outperforms attention-based models on the MIMIC-III top-50 dataset, achieving new state-of-the-art results.	
Zhou et al., 2022	USA	To develop and evaluate cancer domain-pretrained CancerBERT models for extracting breast cancer phenotypes from clinical texts and investigate the effect of customized cancer-related vocabulary on model performance.	CancerBERT, BERT, BioBERT, BlueBERT, CharBERT	Multi-specialty	Cancer domain-specific BERT models (CancerBERT) were developed and outperformed existing BERT-based models in extracting breast cancer phenotypes. Using customized, cancer-specific vocabularies further improved the performance of CancerBERT models.	Pre-established ground truth (2 graduate students with clinical or pharmacy background)
Wang et al., 2023	China	To compare the knowledge and interpretation ability of ChatGPT with those of medical students in China by administering the Chinese National Medical Licensing Examination (NMLE) to both.	ChatGPT	Multi-specialty	ChatGPT's performance on the Chinese National Medical Licensing Examination (NMLE) was lower than that of medical students. Its performance varied across different years and units of the exam.	automated matrics
de Oliveira et al., 2024	Brazil	To propose a model for classifying free-text clinical notes in Portuguese into sentences following the SOAP standard, using domain-specific pre-trained language models.	mBERT, DistilBERT, BioBERT, BioBERTpt, BioBERTptRT	Multi-specialty	Domain-specific pre-trained language models, particularly BioBERTptRT, achieved the best performance in classifying sentences from Portuguese clinical notes into SOAP categories. Specialization in the Portuguese language, clinical terminology, and the medical group's domain improved performance. The Assessment class	3 annotators

					had the lowest performance.	
Datta et al., 2024	USA	To build a generalizable information extraction system leveraging large language models to extract granular eligibility criteria information for diverse diseases from free text clinical trial protocol documents, and to investigate the model's capability to extract criteria entities along with contextual attributes.	GPT-4, GPT-3, Davinci-003	Multi-specialty	AutoCriteria, a GPT-4 based system, demonstrates strong potential for extracting granular eligibility criteria information from clinical trial documents across diverse diseases without manual annotations. The system shows high generalizability and handles complex scenarios, although some limitations exist in identifying specific details and handling complex criteria conditions.	Pre-established ground truth (Knowledge experts)
Wilhelm et al., 2023	Germany, Sweden	To evaluate the performance of four prominent LLMs (Claude-instant-v1.0, GPT-3.5-Turbo, Command-xlarge-nightly, and Bloomz) in generating medical content spanning the clinical specialties of ophthalmology, orthopedics, and dermatology.	Claude-instant-v1.0, GPT-3.5-Turbo, Command-xlarge-nightly, Bloomz, GPT-4	Multi-specialty	LLMs showed notable capability in generating medical content, but there were evident lapses in content quality and potential harm. Claude-instant-v1.0 had the highest mDISCERN score, but also the highest falseness and harmfulness. GPT-3.5-Turbo was the safest. All models lacked in detailing risks, effects on quality of life, and additional information sources.	3 physicians (ophthalmology, orthopedics, and dermatology)
Kaplar et al., 2022	Serbia	To assess the performance of state-of-the-art NER methods on clinical narratives in the Serbian language.	BERT Multilingual, XLM RoBERTa, RoBERTa	Nephrology	State-of-the-art models can provide viable results for clinical NER in Serbian. The best performance was achieved by a majority voting ensemble of CRF, LSTM, and transformer models. Pretraining transformer models on a biomedical corpus improves performance.	Automated matrics
Hartman et al., 2023	USA	To develop and evaluate an automated method	BERT, BART	Neurology	An automated method for generating discharge	2 board-certified physicians (one internist and one neurologist)

		for summarizing the hospital course section of discharge summaries for neurology patients using encoder-decoder sequence-to-sequence transformer models.			summary hospital course sections using encoder-decoder transformer models achieved a quality level approaching that of physician-written summaries. 62% of automated summaries met the standard of care, as rated by board-certified physicians.	
Xie et al., 2022	USA	To automatically extract seizure frequency and seizure freedom information from unstructured text in clinical notes.	BERT, Bio_ClinicalBERT, RoBERTa	Neurology	Finetuned models (BERT, Bio_ClinicalBERT, and RoBERTa) achieved near-human or human-level performance in classifying patients as seizure-free and extracting seizure frequency and date of last seizure from clinical notes. The best combination was Bio_ClinicalBERT for classification and RoBERTa for text extraction. Most gains in performance required roughly 70 annotated notes.	Pre-established ground truth (7 clinicians (epilepsy nurse practitioners, neurology residents, epilepsy fellows, and attending physicians) and 8 non-clinicians (graduate students or epilepsy-related research personnel)
Vithanage et al., 2024	Australia	To develop and evaluate a deep learning and transfer learning-based named entity recognition (NER) model for extracting symptoms of agitation in dementia from nursing notes.	Clinical BioBERT	Neurology	The proposed NER model, combining Clinical BioBERT, BiLSTM, and CRF, achieved satisfactory performance in extracting symptoms of agitation in dementia from nursing notes, outperforming a baseline model using only Clinical BioBERT.	Automated matrics
Ilias and Askounis, 2023	Greece	To introduce deep neural networks for dementia detection from spontaneous speech, capturing inter- and intra-modal interactions between audio and text, and addressing model calibration.	BERT, RoBERTa, ELECTRA, YAMNet, VGGish, Word2Vec, GloVe, Sent2Vec, FastText, Whisper	Neurology	Deep neural networks trained end-to-end capture inter- and intra-modal interactions for dementia detection. Context-aware self-attention, optimal transport, and label smoothing improve	Automated matrics

					performance and calibration. The proposed approaches achieve competitive results on ADReSS and ADReSSo datasets.	
Fonseca et al., 2024	Portugal	To explore the reliability of large language models in neurology by comparing the performance of an AI chatbot with neurologists in diagnostic accuracy and decision-making.	GPT-3.5	Neurology	The AI chatbot (GPT-3.5) performed at a similar level to neurology peers in answering multiple-choice questions based on clinical vignettes, with a mean success rate of 71.3% versus 69.2%. The chatbot demonstrated high accuracy in providing justifications for its correct answers.	Neurology peers (neurologists/neurology residents)
Ali et al., 2023	USA	To assess the performance of ChatGPT and GPT-4 on a 500-question mock neurosurgical written board examination.	ChatGPT, GPT-4	Neurology	Both ChatGPT and GPT-4 achieved passing scores on a mock neurosurgical written board examination. GPT-4 significantly outperformed ChatGPT and human users. Increased word count and higher-order problem-solving were associated with lower accuracy for ChatGPT, but not for GPT-4.	automated matrics
Karakas et al., 2023	USA	To investigate the potential and limitations of using ChatGPT in an outpatient pediatric neurology clinic, focusing on its capabilities in enhancing clinical efficiency, productivity, and patient education.	ChatGPT	Neurology	ChatGPT can improve efficiency in pediatric neurology clinics by assisting with administrative tasks and patient education. However, it has limitations in complex clinical tasks and raises ethical concerns.	Automated matrics
El Haj et al., 2023	France	To assess whether ChatGPT can provide an accurate interpretation of neuropsychological screening data, specifically in a patient with mild Alzheimer's Disease.	ChatGPT	Neurology	ChatGPT can accurately interpret individual neuropsychological test scores in a patient with mild Alzheimer's Disease, but it does not use standardized scores or identify the	Pre-established ground truth (Two neuropsychologists)

					primary impaired cognitive domain, unlike human neuropsychologists.	
Mitchell et al., 2022	USA	To develop a BERT-based system (caBERTnet) to automatically extract detailed tumor site and histology information from free-text oncological pathology reports and predict ICD-O-3 codes.	BERT, BioBERT, ClinicalBERT, CancerBERT	Oncology	A novel NLP system (caBERTnet) was developed, outperforming existing algorithms in predicting ICD-O-3 codes for a wide range of tumor sites and histologies from free-text pathology reports. The system uses a network of three BERT-based models, including a new pathology-tuned BERT model (CancerBERT).	Pre-established ground truth (Certified tumor registrars)
Percha et al., 2022	USA	To evaluate the feasibility of natural language inference (NLI) as a scalable solution for registry curation.	ALBERT, BART, RoBERTa, XLNet, ELECTRA	Oncology	Pre-trained NLI models can curate structured clinical registries from unstructured text with varying degrees of success. ALBERT generally outperformed other models. Common errors involved misinterpreting historical findings and confusion with clinical terms and abbreviations.	2 Reviewers
Wang et al., 2022	USA	To develop deep learning models to recognize ophthalmic examination components from clinical notes in electronic health records (EHR) using a weak supervision approach.	DistilBert, BioBert, BlueBert, ClinicalBert	Ophthalmology	Transformer-based models achieved high precision in recognizing ophthalmic examination components from clinical notes using a weak supervision approach. BERT models outperformed a baseline regular expression model. Performance varied on an independent dataset, with high precision but lower recall for the best-performing BERT model.	Clinician with > 8 years of experience in the ophthalmic field.
Truhn et al., 2024	Germany, UK	To demonstrate the ability of a state-of-the-art LLM (GPT-4) to extract structured	GPT-4	Pathology	GPT-4 can extract structured data from unstructured pathology reports	Pre-established ground truth (experts)

		information from histopathological reports written in plain English using a zero-shot approach.			with high accuracy, rivaling human annotation but at a fraction of the cost. GPT-4 can also propose structured reporting templates and extract quantitative information. This capability can significantly reduce the time and effort required for manual annotation and enable large-scale analyses.	
Oon et al., 2024	Singapore	To benchmark the performance of ChatGPT against pathologists in diagnostic histopathology, and to evaluate the collaborative potential between pathologists and ChatGPT to deliver more accurate diagnoses.	ChatGPT4	Pathology	ChatGPT4 performed worse than pathologists in reaching the correct histopathology diagnosis. Consultation with ChatGPT4 provided limited help, and the information generated is dependent on the prompts provided and is not always correct. Participants rated the diagnostic accuracy of ChatGPT4 poorly but found it useful as an advanced search engine.	Pre-established ground truth (2 pathologists)
Shang et al., 2023	China	To develop PhaTYP, a tool to improve the accuracy of bacteriophage lifestyle prediction (virulent vs. temperate) on short contigs, and to overcome difficulties in lifestyle prediction due to limited annotations and fragmented sequences.	BERT	Pathology	PhaTYP, a BERT-based model, outperforms existing methods for predicting phage lifestyle (virulent or temperate), especially on short contigs. Self-supervised pre-training improves performance. PhaTYP can be applied to metagenomic data to analyze phage lifestyle composition.	Automated matrics
Lu et al., 2021	USA, Republic of Korea	To evaluate the validity of using different NLP/ML methods (Word2vec/ML and BERT) to	BERT, Word2vec, BioBERT, BlueBERT, Clinical BERT	Pediatrics	The BERT method performed better than Word2vec/SVM and Word2vec/XGBoost in characterizing	2 Patient-Reported Outcome Content Experts

		characterize two common symptom domains (pain interference and fatigue) in child and adolescent survivors of cancer, using expert-labeled interview data as the gold standard.			unstructured patient-reported outcomes related to pain interference and fatigue in pediatric cancer survivors. BERT demonstrated higher accuracy, AUROC, and AUPRC, particularly for cognitive attributes.	
Otsuka et al., 2023	Japan	To develop an accurate diagnostic model that covers all psychiatric disorders, using a large-scale linguistic model to predict diagnoses from the history of present illness (HPI) and compare its performance with that of psychiatrists.	UTH-BERT	Psychiatry	The UTH-BERT model achieved a match rate comparable to clinical psychiatrists in predicting psychiatric diagnoses from the history of present illness (HPI). The model's performance varied across different diagnostic categories.	6 psychiatrists (3 Residents and 3 semi-Designated)
Kim et al., 2023	Republic of Korea	To analyse the distinctive linguistic characteristics in users posts on specific mental disorder subreddits and validate their distinctiveness externally by comparing them with posts of subreddits not related to mental illness. Also, to confirm that these differences in linguistic formulations can be learned through a machine learning process.	BERT	Psychiatry	Significant linguistic differences exist between Reddit users posting in different mental health subreddits, and between users in mental health subreddits versus general subreddits. These differences can be detected through LIWC analysis and machine learning clustering techniques.	N/A
Li et al., 2024	Taiwan	To compare the performance of GPT-4, Bard, and Llama-2 on the Taiwan Psychiatric Licensing Examination (TPLE) and on clinical scenario questions for psychiatric differential diagnosis, and to compare their performance with	GPT-4, Bard, Llama-2	Psychiatry	GPT-4 outperformed Bard and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis questions. GPT-4's ability for differential diagnosis closely approached that of experienced psychiatrists.	24 Psychiatrists and senior psychiatry residents

		that of experienced psychiatrists.				
Acharya et al., 2024	USA	To investigate the potential of leveraging language models (LMs) to incorporate supplementary domain knowledge for improving the performance of various EHR-based risk prediction tasks, specifically for SUD, OUD, and Diabetes.	LLaMA2, LLaMA2-7B-chat	Psychiatry	LM-based methods, especially LLaMA2-EHR, significantly outperformed other methods in EHR-based risk prediction tasks. Text-based representation of EHR data improved performance. However, LLM performance can be sensitive to prompt changes.	Automated matrics
Hristidis et al., 2023	USA	To compare the quality of the results of ChatGPT and Google for a collection of queries related to people living with dementia or other cognitive decline (PLWD).	ChatGPT	Psychiatry	Google and ChatGPT have complementary strengths. Google provides sources and dates more often. ChatGPT provides more relevant and objective responses. Both have low readability.	3 domain experts
Parker and Spoelma, 2024	Australia	To assess the capabilities of ChatGPT in generating informative content related to bipolar disorders, including symptoms, classification, causes, and management, and to explore its creativity in generating topic-related songs.	ChatGPT3	Psychiatry	ChatGPT3 can offer superficial information on bipolar disorder and demonstrate creativity. However, its inability to provide accurate and up-to-date references limits its utility for creating a comprehensive review article for scientific journals.	psychiatric researchers/Authors
Butler et al., 2024	USA, The Netherlands	To evaluate the effectiveness of an AI-LLM at improving the readability of knee radiology reports and to assess the medical accuracy of each AI-LLM-generated report.	GPT-3.5	Radiology	AI-LLM (GPT-3.5) significantly improved the readability of knee radiology reports across X-ray, CT, and MRI modalities, as measured by FKRL and FRE scores. However, the AI-LLM generated reports had low, but non-zero, hallucination rates, and increased report length.	2 senior orthopaedic surgeons
Liu et al., 2021	China	To apply a deep learning model and rule-based natural language processing (NLP) method to	BERT	Radiology	A BERT-BiLSTM-CRF model achieved high performance in extracting APHE and PDPH features from	Pre-established ground truth (2 radiologists with >5 years' experience, with a third expert radiologist

		identify evidence for liver cancer diagnosis automatically.			Chinese radiology reports. Combining these features with rule-based NLP features improved liver cancer diagnosis prediction.	resolving inconsistencies)
Cai et al., 2023	China, USA	To develop and evaluate a novel framework of abstractive summarization methods to automatically generate the "impression" section of chest radiology reports, addressing domain-specific terminology challenges.	ChestXRayBERT, BERT, GPT, BioBERT, ClinicalBERT, BART	Radiology	ChestXRayBERT outperforms other neural network-based abstractive summarization models, including those using general-domain pre-trained models like BERT and BART. Pre-training on a radiology-specific corpus is crucial for performance. The model shows good performance even with limited training data.	7 volunteers, including 3 radiology researchers and a clinical radiologist.
Datta et al., 2022	USA	To propose a two-turn question answering (QA) method based on a transformer language model, BERT, for extracting detailed spatial information from radiology reports and to demonstrate the advantage that a multi-turn QA framework provides over sequence-based methods.	BERT	Radiology	Framing fine-grained spatial information extraction as a two-turn QA task using BERT outperforms traditional sequence labeling methods. Incorporating domain knowledge into the query improves results for some spatial and descriptive frame elements.	Automated matrics
Hu et al., 2024	China	To explore whether ChatGPT can extract information from radiological reports in a zero-shot manner.	ChatGPT	Radiology	ChatGPT can achieve competitive information extraction performance from radiological reports in a zero-shot manner, comparable to baseline methods for some tasks. Adding prior medical knowledge to prompts can improve performance for some tasks but may worsen it for others, particularly more complex ones.	Pre-established ground truth (Two engineers with a medical informatics background, and a clinician)

Bressem et al., 2024	Germany, USA, Netherlands	To build a comprehensive German clinical language model (medBERT.de) trained on a diverse set of medical texts and evaluate its performance on various medical benchmarks, including radiology reports, clinical practice guidelines, and electronic health records.	GottBERT, BioGottBERT, Multilingual BERT, German-MedBERT, medBERT.de	Radiology	medBERT.de, a German medical BERT model, outperforms general and other medical models on various medical benchmarks. Domain-specific models are particularly useful for longer texts. Data deduplication does not consistently improve performance.	Board-certified radiologists and medical students
Shea Ying Cong Tan et al., 2023	Singapore	To evaluate the relative performance of newer large language models against previous NLP methods for cancer disease response classification; explore the effects of data augmentation techniques and prompt-based fine-tuning on model performance; and test the best-performing model on a set of RECIST-based classifications to explore its potential utility in clinical trials.	GatorTron, BioMegatron, BioBERT, RoBERTa, RadBERT, XLNet, PubMedGPT, DeBERTa, BioClinicalBERT, BERT	Radiology	Large clinical language models, particularly GatorTron, can accurately infer cancer disease response from radiology reports. Data augmentation improves performance, and prompt-based fine-tuning reduces the required training data size while maintaining good performance.	Pre-established ground truth (Medical research staff, medical oncologists, curator medical oncologist)
Chizhikova et al., 2023	Spain	To present CARES, a new corpus of radiology medical reports written in Spanish and labeled with ICD-10 codes, and to test its validity for automatic text classification using several machine learning algorithms based on transformer language models.	BioBERT, RoBERTa, BETO	Radiology	A new corpus of Spanish radiology reports (CARES) labeled with ICD-10 codes is presented. The RoBERTa biomedical-clinical model achieved the best performance for classifying reports into ICD-10 chapters.	Pre-established ground truth (4 expert radiologists)
Ayoub et al., 2023	USA	To assess ChatGPT's capabilities as a source of medical knowledge, using Google Search as a comparison.	ChatGPT	Surgery	ChatGPT performed better than Google Search for general medical knowledge questions but worse when providing medical recommendations. Both generally described urgency, diagnoses,	2 Blinded Reviewers

					diagnostics, and management appropriately.	
Cuthbert et al., 2023	UK	To assess whether ChatGPT could pass Section 1 of the Fellowship of the Royal College of Surgeons (FRCS) examination in Trauma and Orthopaedic Surgery, using the UKITE 2022 examination as a surrogate.	ChatGPT	Surgery	ChatGPT is currently unable to pass the FRCS (Trauma & Orthopaedics) examination. The model fails to recognize its own limitations and provides incorrect answers with explanations.	UKITE committee
Mishra et al., 2023	USA	To assess the quality of neurosurgical information provided by ChatGPT and compare it to the American Association of Neurological Surgeons (AANS) 'For Patients' webpages.	ChatGPT	Surgery	ChatGPT is an adaptive resource for neurosurgical information but has shortcomings including poor readability, lack of references, and failure to fully describe treatment options. The quality of its responses is limited.	4 Reviewers
Mika et al., 2023	USA	To determine whether ChatGPT could appropriately answer frequently asked questions regarding total hip arthroplasty (THA).	ChatGPT	Surgery	ChatGPT effectively provided evidence-based responses to common patient questions about total hip arthroplasty (THA). Most responses were accurate, requiring only minimal or moderate clarification. The chatbot may be a valuable tool for patient education.	Two authors (MDs)
Bombieri et al., 2023	Italy, Germany	To extract procedural surgical knowledge from intervention procedure textbooks and papers, framed as a Semantic Role Labeling (SRL) task, and to evaluate the performance of different Transformer-based information extraction methods, including a new surgical language model (SurgicBERTa).	RoBERTa, BioMedRoBERTa, SurgicBERTa	Surgery	Fine-tuning a pre-trained domain-specific language model (SurgicBERTa) achieves the highest performance on all dataset splits and sub-tasks (predicate disambiguation, argument disambiguation, and predicate-argument disambiguation). Domain-specific pre-training and fine-tuning significantly improve performance	Automated matrics

					compared to zero-shot learning with general-English or biomedical models.	
Yun et al., 2023	Republic of Korea	To evaluate the answers provided by ChatGPT during hypothetical breast augmentation consultations across various categories and depths, using validated tools and comparing scores between plastic surgeons and laypersons.	ChatGPT	Surgery	Existing health information evaluation tools may not be entirely suitable for evaluating the quality of individual responses generated by ChatGPT. The development and implementation of appropriate evaluation tools to assess the appropriateness and quality of AI consultations are necessary.	10 (5 plastic surgeons and 5 laypersons)
Ghanem et al., 2024	USA	To evaluate the content and quality of AI-generated medical information on acute appendicitis.	ChatGPT-3.5, ChatGPT-4, Bard, Claude-2	Surgery	AI-generated information on appendicitis had generally favourable quality scores, but Claude-2 performed significantly worse. Bard was the only LLM to provide verifiable sources. Readability levels were far above the recommended 6th-grade level for all LLMs.	Pre-established ground truth (3 medical/surgical)
Xue et al., 2024	China	To evaluate ChatGPT-4's responses against those of professional health care providers in telemedicine, assessing artificial intelligence's capability to support the surge in remote medical consultations and its impact on health care delivery.	ChatGPT-4	Surgery	ChatGPT-4 matches the expertise in DingXiangYuan forums' paid consultations, particularly excelling in scientific education. It is a promising alternative for remote health advice.	3 blindfolded, experienced orthopedic surgeons
Kaarre et al., 2023	USA, Sweden	To investigate the feasibility of utilizing LLMs in orthopaedics by posing to ChatGPT questions relevant to anterior cruciate ligament (ACL) surgery and evaluating its responses, and to	GPT-4	Surgery	ChatGPT generated correct responses in approximately 65% of cases related to ACL surgery. It can adapt to different target audiences but cannot replace orthopaedic surgeons due to limited	Pre-established ground truth (Two board-certified orthopaedic sports medicine surgeons and two expert orthopaedic sports medicine surgeons)

		evaluate the depth of the LLMÇƒÙs knowledge and investigate its adaptability to different user groups.			understanding and potential for errors.	
Massey et al., 2023	USA	To compare the performance of orthopaedic residents against ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations, and to perform a subgroup analysis comparing performance on questions with and without image interpretation.	ChatGPT-3.5, GPT-4	Surgery	Orthopaedic residents outperformed both ChatGPT-3.5 and GPT-4 on orthopaedic assessment examinations. GPT-4 was superior to ChatGPT-3.5. Both LLMs performed significantly better on text-only questions compared to questions requiring image interpretation.	Automated matrics
Song et al., 2023	China	To evaluate the effectiveness of four large language models (LLMs) (Claude, Bard, ChatGPT4, and New Bing) in the context of medical consultation and patient education in urolithiasis.	Claude, Bard, ChatGPT4, New Bing	Urology	Claude demonstrated superior performance compared to Bard, ChatGPT4, and New Bing in urolithiasis consultation and education. LLMs show potential in medical health consultations but require professional review and further evaluation.	3 associate chief physicians with expertise in urolithiasis.