Supplementary Table S2: *Evaluation Metrics*

Study / Year	LLMs	Medical Specialty	Task / Application	Evaluation Context	Key Findings	Reported Performance	Evaluated Compared to
Liu, et al, 2021.	BERT, BiLSTM-CRF	Radiology	Evidence extraction for liver cancer diagnosis	Pre-established ground truth (2 radiologists with >5 years experience, with a third expert radiologist resolving inconsistencies)	A BERT-BiLSTM-CRF model achieved high performance in extracting APHE and PDPH features from Chinese radiology reports. Combining these features with rule-based NLP features improved liver cancer diagnosis prediction.	APHE/PDPH F1: 98.40%/90.67% (Report level); Diagnosis F1: 88.55% (Combined features)	Baseline models (CRF, BiLSTM-CRF)
Hartman, et al, 2023.	BERT, BART	Neurology	Automate discharge summary hospital course generation	2 board-certified physicians (one internist and one neurologist)	An automated method for generating discharge summary hospital course sections for neurology patients 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	Physician Quality Rating (Mean Likert 1-10): 6.52/10; 62% met standard of care (≥7/10)	Human physicians

					rated by board-certified physicians.		
Mitchell, et al, 2022.	BERT, BioBERT, ClinicalBERT, CancerBERT	Oncology, Pathology	ICD-O-3 code prediction from pathology reports	Pre-established ground truth (Certified tumor registrars)	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).	Accuracy (Top-5, >5 samples): Site 93.28%, Histology 96.27%; Accuracy (Group-level): Site 93.53%, Histology 97.6%	Existing algorithms
Campbell, et al, 2024.	GPT-3.5, GPT-4	Endocrinology	Patient education on thyroid nodules	Otolaryngology resident physicians, experienced thyroid and parathyroid head-and-neck-attending surgeon)	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.	Correctness: 69.2% (at least correct); FKGL (8th-grade prompt): 13.43 (collegiate)	N/A

					Significant rates of AI hallucination may preclude recommending ChatGPT as an educational tool.		
Bellinger, et al, 2023.	ChatGPT	ENT	Quality/readability of BPPV patient information	Pre-established ground truth (2 experts otolaryngologists with subspecialty certification in Neurotology, and a second-year fellow in Otology/Neurotology))	Information on ChatGPT regarding BPPV is more difficult to read, of lower quality, and more difficult to comprehend compared to information obtained from Google searches.	FKGL: 13.9 (vs 10.7 Google); DISCERN Part 2: 17.5/25.0 (vs 25.4 Google Search)	Google Search results
Alkouz, et al, 2022.	Deepluenza (BERT multilingual)	Epidemiology	Influenza detection from Tweets	Pre-established ground truth (2 authors of the study)	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.	F1 (Positive Class): 0.98; Accuracy: 0.99; NCC with hospital visits: 0.93 (Combined lang.)	KNN, SVM, DPCNN

Bombieri, et al, 2023.	RoBERTa, BioMedRoBERTa, SurgicBERTa	Surgery	Surgical action extraction from procedure textbooks	SRL metrics (P, R, F1) on manually annotated dataset (RSPF)	Fine-tuning domain-specific SurgicBERTa achieves highest performance for surgical SRL. Full fine-tuning > Zero-shot. Predicate-argument eval needed.	F1 (Arg Disambig): 0.773; F1 (Pred-Arg Disambig): 0.732 (SurgicBERTa FFT, BAL split)	RoBERTa, BioMedRoBERTa (Zero-shot & FFT)
Cai, et al, 2021.	ChestXRayBERT, TransAbs, BART, Seq2Seq, PGN	Radiology	Abstractive summarization (Findings → Impression)	ROUGE, BLEU, Human Eval (Readability, Correctness, Informativeness, Redundancy)	Custom domain-specific ChestXRayBERT outperformed general/biomedical models (BERT, BioBERT, ClinicalBERT) and non-pretrained models.	ROUGE-L: 47.05 (vs BART 45.97, TransAbs 44.07)	BERT, BioBERT, ClinicalBERT, TransAbs, BART, Seq2Seq, PGN
CB, et al, 2023.	BERT, BioClinicalBERT, CovidBERT	Bioinformatics	Ontology-based semantic data interestingness mining	Expert evaluation, Chi-Square	Ontology combined with BERT models improves semantic richness and interestingness of association rules from COVID-19 data. BioClinicalBERT > CovidBERT.	Qualitative: High correlation with expert ratings	Traditional ARM

Chen, et al, 2021.	BERT, Transformer	Emergency Medicine	Disease concept embedding and retrieval from ED EHRs	t-SNE visualization, AUC-ROC	Proposed EDisease model (Transformer + self-supervised methods) performs well in disease concept embedding, retrieval, and downstream outcome prediction.	AUC-ROC (Outcome Prediction): 0.876 (NTUH), 0.88 (NHAMCS)	Other DNN models
Chizhikova, et al, 2023.	BETO, RoBERTa-BNE, Bio-BERT-Spanish, RoBERTa biomedical	Radiology	ICD-10 chapter classification from Spanish radiology reports	Micro/Macro F1, Precision, Recall	Biomedical/clinical domain RoBERTa model achieved best performance. General domain models showed cross-domain potential.	F1 Macro: 0.8328; F1 Micro: 0.8676 (Best model)	N/A
Cuthbert, et al, 2023.	ChatGPT	Orthopedics	Performance on UKITE 2022 exam (FRCS surrogate)	Accuracy	Failed to pass. Performed better on basic science, worse on trauma.	Accuracy: 35.8% (vs 65.8% pass mark)	N/A
Datta, et al, 2022.	BERT (Cased, Uncased, MIMIC+Cased)	Radiology	Fine-grained spatial information extraction from radiology reports	Average F1 (10-fold CV)	Two-turn QA approach outperforms sequence labeling for spatial IE. Domain knowledge in query helps.	F1 (Triggers): 90.07; F1 (Figure): 78.13; F1 (Ground): 83.77	Sequence labeling baseline

Datta, et al, 2023.	GPT-4	Multi-specialty	Eligibility criteria extraction from ClinicalTrials.gov documents	F1 Overall, Accuracy	AutoCriteria (GPT-4 based) performance evaluated across 9 diseases.	F1 Overall (Entities): 89.42%; Accuracy (Attr+Val+Type+Temp+Mod): 78.95%	N/A
Fu, et al, 2022.	BERT, CNN, Bi-LSTM, Rule-based	Geriatrics	Fall occurrence identification from EHR progress notes	F1 (Sentence, Document, Patient levels)	Hybrid model (BERT + post-hoc rules) achieved highest performance, outperforming BERT alone, rules, CNN, and Bi-LSTM.	F1 Sentence: 0.971; F1 Document: 0.985; F1 Patient: 0.954	Rule-based, CNN, Bi-LSTM
Guo, et al, 2022.	BERT, Siamese-BERT, Elasticsearch	Multi-Specialty	Medical QA system with KG enhancement	ACC@1, MAP, MRR	Proposed Siamese-BERT model improves ranking. Entity matching (NER, KG) enhances performance. System outperforms baseline QA system.	ACC@1 (cMedQA): 0.777 (vs 0.758 BERT); ACC (In-house): 63.2% (vs 47.4% baseline)	Elasticsearch QA, BERT, MP-CNN
Ilias, et al, 2023.	BERT, DeiT, Optimal Transport	Neurology	Dementia recognition from speech/transcripts	Accuracy, F1, Precision, Recall, Specificity, ECE, ACE	Context-aware attention, OT domain adaptation, label smoothing improve performance and calibration for multimodal dementia detection.	Accuracy: 91.25%; F1: 91.06% (Best model, ADReSS)	Unimodal/Multimodal baselines

Ji, et al, 2021.	BERT variants (BERT-base, BlueBERT, etc.), CNNs	Multi-Specialty	ICD code assignment from MIMIC-III discharge summaries	AUC-ROC, F1 (Macro/Micro), P@k	CNN outperformed BERT variants on Top-50 dataset. Pre-training domain matters. Performance drops for infrequent codes.	F1 Micro (Top-50): Retrained CNN 0.671 vs Best BERT 0.681	Baseline CNNs, Other BERT variants
Kaplar, et al, 2022.	CRF, LSTM-CRF, BERT Multilingual, XLM-RoBERTa	Nephrology	Clinical NER for Serbian EHR discharge summaries	F1 (Strict/Token), Precision, Recall	Ensemble model achieved best performance. Pretraining helps deep models on small datasets. State-of-the-art methods applicable to complex non-English languages.	F1 Strict: 0.892; F1 Token: 0.899 (Best Ensemble)	N/A
Karthikeyan, et al, 2022.	RoBERTa	Multi-Specialty	OCR post-correction for scanned medical reports	WER, CER	RoBERTa-based post-processing significantly reduces OCR errors on NHS and MiBio datasets.	WER/CER Reduction (NHS): 1.39/0.83; WER/CER Reduction (MiBio): 5.709/2.07	Tesseract OCR output
Kim, et al, 2023. under	BERT	Psychiatry	Understanding mental health issues from Reddit posts	Qualitative (LIWC), Clustering (k-means, UMAP, HDBSCAN)	Distinct linguistic features identified across mental health subreddits using LIWC and BERT embeddings. Users active in multiple subreddits show	Clustering Accuracy (External): 96.0% (Sentence-BERT), 89.1% (k-means), 99.8%/99.9% (UMAP)	N/A

					varied patterns. Mental health posts distinguishable from general posts.		
Klein, et al, 2021.	BERT-Base-Uncased, COVID-Twitter-BERT	Epidemiology	Identifying potential COVID-19 cases from Twitter	F1, Precision, Recall	COVID-Twitter-BERT outperformed BERT-Base. Pipeline identified 13,714 potential case tweets with US state geolocation.	F1 (Potential Case): 0.76 (COVID-Twitter-BERT) vs 0.70 (BERT-Base)	N/A
Li, et al, 2021.	BiLSTM-CRF	Clinical NLP	NER on synthetic vs natural clinical notes	F1 (Strict/Relaxed), Precision, Recall	Synthetic notes generated by GPT-2 are useful for training NER models, performing comparably or slightly better than models trained on natural notes when tested externally. Augmenting natural data with synthetic data improves performance.	F1 Strict Overall (Synthetic for external_1/2): 0.709 / 0.748 vs (Natural for external_1/2): 0.706 / 0.737	Models trained on natural data
Liu, et al, 2023. using	BERT variants (BioBERT, etc.),	Multi-Specialty	Deidentification (SHI extraction) from Australian	F1 (Strict/Relaxed), Precision, Recall	Discharge Summary BioBERT fine-tuned model with cascaded preprocessing rules achieved the best performance. BERT	F1 Strict: 0.9659; F1 Relaxed: 0.9687 (Best model)	LSTM+Embeddings

	LSTM+Embeddings		pathology reports		models > LSTM+Embeddings.		
Liu, et al, 2023.	ChatGPT	Clinical NLP	Using AI suggestions to optimize Clinical Decision Support (CDS)	Expert Likert ratings	AI suggestions rated lower on usefulness/acceptance than human but provided unique perspectives. Can complement traditional CDS optimization.	Usefulness (Likert 1-5): AI 2.7 vs Human 3.5 ($p<0.001$)	Human-generated suggestions
Lu, et al, 2021. natu	BERT, Word2vec+SVM, Word2vec+XGBoost	Pediatrics, Oncology	Characterizing unstructured PROs (Pain/Fatigue) from interviews	Accuracy, F1, AUROC, AUPRC	BERT performed best in identifying symptom attributes (physical, cognitive, social) compared to Word2vec/ML methods. High specificity, low sensitivity.	Accuracy (Cognitive/Pain): 0.931; AUROC (Cognitive/Fatigue): 0.948; AUPRC (Cognitive/Fatigue): 0.855 (BERT)	Word2vec+SVM, Word2vec+XGBoost
Lu, et al, 2021. Revea	BERT, BTM, k-means	Multi-Specialty	QA and Opinion Aggregation from CORD-19 literature	Qualitative: Opinion trend analysis	Pipeline successfully retrieved answers, aggregated opinions, and revealed longitudinal trends in coronavirus research.	N/A	N/A
Mika, et al, 2023.	ChatGPT	Orthopedics	Answering patient questions	Custom rating (Excellent, Satisfactory+/-	ChatGPT provided mostly satisfactory answers (requiring	7/10 Satisfactory, 2/10 Excellent,	N/A

			about Total Hip Arthroplasty (THA)	, Unsatisfactory)	clarification) but often outdated or lacking detail. Potential as education adjunct but needs caution.	1/10 Unsatisfactory	
Mishra, et al, 2023.	ChatGPT	Neurosurgery	Quality/Readability of neurosurgical patient information	DISCERN score, Flesch-Kincaid	ChatGPT provided "fair" quality info, better than some media but worse than AANS website. High reading difficulty. Significant reference inaccuracies (hallucinations).	DISCERN Score: 44.2 ("Fair"); Flesch-Kincaid Score: 32.4 ("Difficult")	AANS Website
Mulyar, et al, 2021.	MT-Clinical BERT (Bio+Clinical BERT based)	Clinical NLP	Multitask Learning (NER, Entailment, STS) across 8 clinical datasets	Pearson Rho, Accuracy, Micro-F1	Single multitask model performs competitively with multiple task-specific models, offering computational benefits. Slight performance degradation observed, especially on orthogonal tasks (language/objective) .	F1 (Micro NER avg): 87.4 - 91.9% (task-dependent)	Optimized Clinical BERT (single-task)

Otsuka, et al, 2023.	UTH-BERT	Psychiatry	Diagnosing psychiatric disorders from HPI in discharge summaries	Accuracy, Precision, Recall, F1	BERT model performance comparable to clinical psychiatrists (semi-Designated and Residents) in predicting ICD-10 diagnosis from HPI.	Accuracy: 74.3% (vs 71.5% semi-Designated, 69.4% Residents)	Clinical psychiatrists
Percha, et al, 2021.	ALBERT, BART, RoBERTa, XLNet, ELECTRA	Oncology	NLI for clinical registry curation (breast oncology) from EHR notes	Accuracy (MES-T > MES-F)	ALBERT outperformed other NLI models. Performance varied significantly across registry fields. Temporality and term ambiguity were major error sources.	Accuracy Range: 11.9% (PR/Right) to 94.4% (Age) depending on field and model. ALBERT best overall.	N/A
Qiao, et al, 2021.	BERT-Kcr (BERT + BiLSTM)	Bioinformatics	Lysine crotonylation site prediction	AUC-ROC, Sn, Sp, ACC, MCC	Proposed BERT-Kcr model outperformed state-of-the-art methods on independent test set. NLP features effective.	AUROC: 0.905 (vs 0.859 Deep-Kcr)	Deep-Kcr, Position_weight, CKSAAP_CrotSite, LightGBM-CrotSite
Shang, et al, 2023.	PhaTYP (BERT based), DeePhage, PHACTS, etc.	Medical Genetics and Genomics	Phage lifestyle prediction (virulent/temperate) from contigs	AUC-ROC, Sn, Sp, Acc	PhaTYP (BERT-based with self-supervised + fine-tuning) outperforms other methods, especially on short contigs.	AUC-ROC (Complete Genomes): 0.99 (vs 0.96 DeePhage); AUC-ROC (100-	DeePhage, PHACTS, BACPHLIP, PhagePred

						400bp): 0.92 (vs 0.82 DeePhage)	
Vithanage, et al, 2024.	BioBERT, BiLSTM-CRF	Dementia	NER (Symptoms of agitation) from nursing notes	F1, Accuracy	Hybrid deep learning model outperformed baseline BioBERT.	F1: 75%; Accuracy: 78% (vs 68% / 70% for BioBERT only)	BioBERT only
Wang, et al, 2022.	CancerBERT (BlueBERT based), BERT-large, etc.	Oncology	Cancer phenotype NER from EHRs (pathology/clinical notes)	F1 (Strict/Lenient, Macro/Micro)	CancerBERT models (especially with custom vocabulary) outperform other BERT variants and BiLSTM-CRF baseline. Custom vocab improves performance.	F1 Macro Strict: 0.876; F1 Micro Strict: 0.909 (Best CancerBERT)	Other BERT models, BiLSTM-CRF
Wang, et al, 2023. text	BERT, BiLSTM-CRF	Multi-Specialty	EHR text deidentification using Transformers	F1 (Strict/Relaxed), Precision, Recall	Discharge Summary BioBERT fine-tuned model with cascaded preprocessing rules achieved best performance. BERT models > LSTM+Embeddings.	F1 Strict: 0.9659; F1 Relaxed: 0.9687 (Best model)	LSTM+Embeddings
Wang, et al, 2023. effici	BioDGW-CMI (BioBERT+GraphWave+DAE), etc.	Bioinformatics	circRNA-miRNA interaction prediction	AUC, AUPR	BioDGW-CMI outperformed existing CMI prediction models significantly on all datasets.	AUC: 0.9026 (CMI-9905), 0.9476 (CMI-9589), 0.7821 (CMI-753)	KGDCMI, WSCD, SGCNCMI, JSNDCMI, CMIVGSD, GCNCMI, NECMA, IIMCCMA

Xie, et al, 2022.	BERT, Bio_Clinical BERT, RoBERTa	Neurology	Extracting seizure frequency/date from epilepsy clinic notes	Accuracy (Classif.), F1 (Extraction)	Finetuned models achieve human-level performance on text extraction and near-human on classification. Small annotation set effective.	Accuracy (Classif.): >80% (Bio_ClinicalBERTFT); F1 (Extraction): >0.80 (All models)	Human annotator performance
Zhou, et al, 2022.	CancerBERT (BlueBERT based), BERT-large, etc.	Oncology	Cancer phenotype NER from EHRs (pathology/clinical notes)	F1 (Strict/Lenient, Macro/Micro)	CancerBERT models (especially with custom vocabulary) outperform other BERT variants and BiLSTM-CRF baseline. Custom vocab improves performance.	F1 Macro Strict: 0.876; F1 Micro Strict: 0.909 (Best CancerBERT)	Other BERT models, BiLSTM-CRF