PANDORA: An AI model for the automatic extraction of clinical unstructured data and clinical risk score implementation

1st Daniel Jimenez
Arkangel AI
Bogota, Colombia
daniel.jimenez@arkangel.ai

2nd Natalia Castano-Villegas
Arkangel AI
Bogota, Colombia
natalia@arkangel.ai

3rd Isabella Llano
Arkangel AI
Bogota, Colombia
isabella.llano@arkangel.ai

4th Julian Martinez
Arkangel AI
Bogota, Colombia
julian.martinez@arkangel.ai

5th Laura Ortiz
Arkangel AI
Bogota, Colombia
laura.ortiz@arkangel.ai

6th Laura Velasquez
Arkangel AI
Bogota, Colombia
laura@arkangel.ai

7th Jose Zea
Arkangel AI
Bogota, Colombia
jose@arkangel.ai

2025 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3315-2400-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CAI64502.2025.00280

***Abstract*—Introduction: Medical records and physician notes contain valuable non-tabular information that requires significant manual effort to extract and structure. Large Language Models (LLMs) have demonstrated the ability to understand, reason, and retrieve information from such sources, transforming it into accessible information for clinical or research purposes. Objective: To present and asses the capabilities of PANDORA, our AI system, comprised of two LLMs that can extract data and use it in validated calculator and prediction models to provide recommendations. Methods: The study evaluates the model's ability to extract clinical features from discharge notes from the MIMIC database and synthetically generated outpatient charts. We use the PUMA calculator for Chronic Obstructive Pulmonary Disease (COPD) case finding, which interacts with the model to calculate a score based on seven criteria and determine which patients should undergo further spirometry testing. Results: The model's exhibited excellent extraction accuracy, achieving 100% for MIMIC and 99% for synthetic cases. Interaction with the PUMA scale produced accurate scores, with an accuracy of 94% for both databases. The final recommendation on COPD risk was based on the PUMA scale, classified as positive if score $\geq$ 5. Sensitivity was 86% for MIMIC and 100% for synthetic cases. Conclusion: LLMs have been successful in extracting information and generating recommendations. However, this is the first model that successfully extracts information based on existing, validated clinical scores from plain, non-tabular data and provides a recommendation mixing all these capabilities. It leverages existing knowledge, making it available to be explored in light of the highest-quality evidence in several medical fields.**

***Index Terms*—Artificial Intelligence, Information Extraction, Unstructured Data, COPD, LLM evaluation**

I. Introduction

One of the most important sources of bias in observational research is the quality of the secondary information sources, often clinical registries [1], [2]. As much as 80% of clinical data is unstructured [3], meaning that lots of valuable information is buried in text formats or lost in low-quality databases [4]. Advances in technology and computational sciences have led to the ability to collect, organise, operate, analyse, and interpret vast amounts of data (i.e., Big Data) [5], [6] and to program computers to reproduce tasks only performed by humans in past decades (Machine Learning or ML) [7]–[9]. When we analyze the impact of unstructured data, understanding that it is one of the most prominent sources of information for Artificial Intelligence (AI), poor structuring is not only harmful on an individual level but could also introduce bias in worldwide used algorithms that could affect millions of patients and have a significant impact on the economy [10], [11]. PANDORA was born from the advances made in Natural Language Processing (NLP). With it, we propose an effective and precise AI solution that could actively help healthcare personnel find the information they need from unstructured data. It was built as a robust algorithm framework that retrieves data from plain text and makes it accessible to understand disease patterns and make high-impact decisions. PANDORA can also apply scores and clinical practice guidelines to the information retrieved. In the following sections, we explain how PANDORA came to be and give an initial scope of what could be achieved with its implementation in healthcare scenarios.

II. Methodology

A. General Description

PANDORA is a modular, two-piece algorithm. One extracts information from Electronic Health Records (EHRs) constituting the extraction phase. The other uses the extracted information and applies it to clinical guidelines and validated scores (scoring or punctuation phase) to issue recommendations (recommendation phase). For this study we used the PUMA scale for opportunistic case finding of Chronic Obstructive Pulmonary Disease (COPD) [12]. The two agents work synchronically. The features extracted are defined according to the guideline or score (e.g., PUMA). From this, a knowledge

979-8-3315-2400-5/25/$31.00 ©2025 IEEE
DOI 10.1109/CAI64502.2025.00280

base is constructed and used to predict the outcome of interest (e.g., COPD risk). Figure 1 depicts PANDORA's workflow.

B. General Sources of Data

We created two types of validations to obtain cases resembling real-life clinical charts and extract specific information from EHRs. The first used the Medical Information Mart for Intensive Care (MIMIC) database. This database contains de-identified free-text clinical discharge notes from data collected in Intensive Care Unit (ICU) patients hospitalized at the Beth Israel Deaconess Medical Center in Boston, USA, from 2002 to 2019 (https://mimic.mit.edu/) [13]–[15]. The second was an AI-generated database with cases created using GPT, following a standard form designed by the medical team to simulate the structure of a clinical record made by a general physician at an outpatient consultation in compliance with the Ministry of Health in Colombia [16]. Regarding the Recommendation Phase, we used COPD as the pathological entity of interest. The decision to choose this disorder was based on its high prevalence in the general population [17]. We defined the presence or absence of risk for diagnosis of COPD as our primary outcome. This decision was based on its high sub-diagnosis (89%) [18] and the measurability of the binary outcome. Therefore, we used the standard clinical guideline for COPD, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2024 Report, available at https://goldcopd.org/, and the PUMA COPD opportunistic case finding tool [12], [19], including their latest validation in 2022 [20]. The synchronism of both algorithms or the Scoring Phase did not require training, as it used the capabilities embedded into the PANDORA Large Language Model (LLM) structure. The scoring was based on the PUMA scale (Supplementary Material 1), which rates the COPD risk by employing seven features. Risk is defined as a score ≥ 5.

C. Materials

We used Python 3.12.2, Microsoft Office 365, the Arkangel App (https://www.arkangel.ai/), AI translation and writing assistant Deepl (https://www.deepl.com).

D. Algorithmic Framework

PANDORA uses statistical and NLP algorithms to extract and analyse data from EHRs. Its capabilities include:

1) Natural Language Processing (NLP) techniques and processes to extract relevant disease-related factors from unstructured text within EHRs. This includes using models that understand medical terminology and context for accurate information extraction.
2) Chain of Thought Strategy (CoT): This strategy ensures a sequence of reasoning when extracting and analysing data. PANDORA can accurately map or associate patient data with disease factors by following a logical progression. CoT consists of steps to guide LLMs through multi-step reasoning problems, which improves LLM reasoning [21].
3) Non-Relational Database Algorithms: To manage the knowledge base, non-relational database algorithms efficiently store and retrieve patient-specific features, allowing quick access during the recommendation process.
4) Clinical algorithms: Used to construct recommendations. We employed the PUMA calculator for COPD risk screening.

E. System Development

Using the previous framework, we developed the system's three main components:

1) **EHR Data Extraction**: This process extracts critical disease factors from EHRs using advanced NLP techniques. It leverages the CoT strategy to ensure relevant data is accurately identified and captured from unstructured text.
2) **Knowledge Base Construction**: The extracted information is used to build a non-relational knowledge base to store disease features according to clinical scores (e.g., PUMA), which serves as a structured repository that supports the next step.
3) **Recommendation System**: PANDORA employs a recommendation mechanism instead of directly inferring from guidelines/scores. The model takes information stored in the knowledge base and suggests whether a patient is at risk for a disorder (COPD) by analysing the extracted factors and their alignment with established medical criteria.

F. Algorithm Evaluation

1) Evaluation of the Extraction Phase: We used the EHR-DS-QA dataset (https://physionet.org/content/ehr-ds-qa/1.0.0/) to evaluate the model's ability to extract information from EHRs. It was created using the LLM Meta Llama 2 to generate clinical question-answer (QA) pairs based on the 21.466 medical summaries from MIMIC-IV. It produced 156.599 QA-pairs, 506 of which were evaluated by physicians. This subset constituted our reference standard (Figure 2.). An example of a question in EHR-DS-QA is: Does the patient have any known allergies or adverse drug reactions? If extraction was correct, it should match the exact statement in EHR-DS-QA [21]. No preprocessing was applied to input data since the goal was to handle complex, unstructured data. We also emphasised handling open-ended responses so the algorithms could extract relevant factors. Then, we employed semantic metrics to assess text summarization against the EHR-DS-QA benchmark. The BERTscore, SimilarityScore, SemanticScore, and RelevanceScore are explained in Supplementary Material 2.

Additionally, we used the Judge Alignment Metric (JAM) to assess extraction. This strategy was developed as an automated, more scalable alternative to human evaluation [22]. It employs state-of-the-art LLMs to judge PANDORA's responses. Table I presents the model's semantic scores and operative characteristics according to JAM. The LLM used as JAM was Claude from the IA company Anthropic

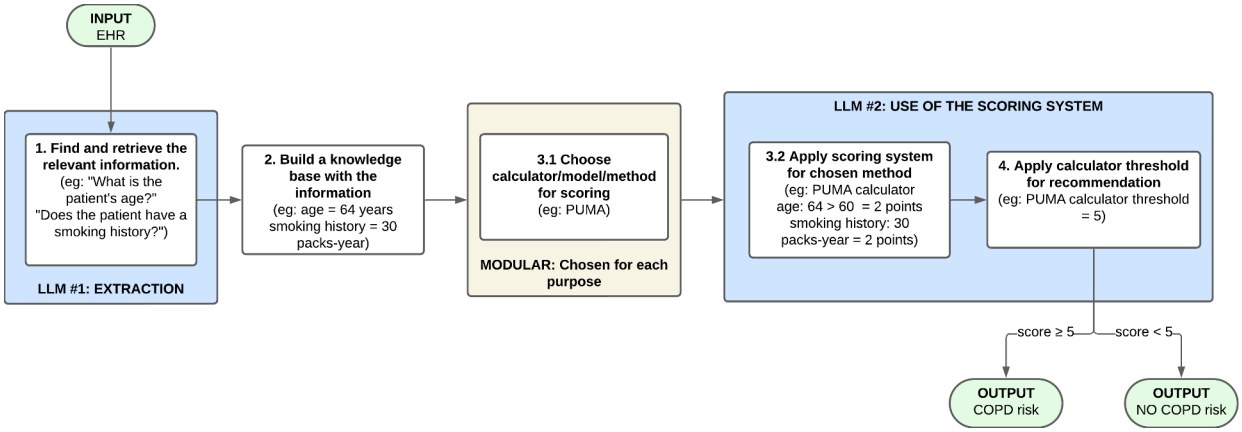

Fig. 1. Workflow structure of PANDORA

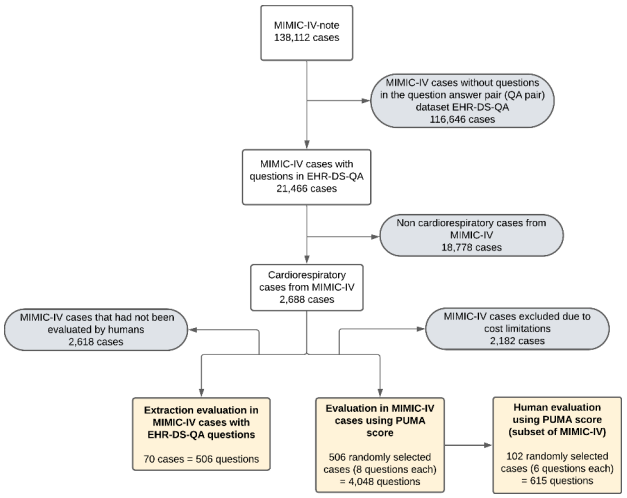

Fig. 2. Flow chart of the MIMIC-IV-Note database and sample size

(https://www.anthropic.com/claude). It was given the same information as PANDORA, including the EHRs and PUMA scale.

2) Recommendation Phase Evaluation: **First Strategy:** A two-step evaluation was designed to determine the model's specific extraction capability when using PUMA. First, we randomly selected 506 QA-pairs from EHR-DS-QA (different from the 506 human-revised QA-pairs) to evaluate extraction in equal sample sizes. Also, assessing all 2688 QA-pairs was extremely costly. We performed a human eval for a subset of 102 clinical cases (20%) to confirm that retrieved information was consistent with the information registered in EHR-DS-QA and to evaluate its accuracy. Table II. indicates outcomes for this stage. Supplementary Material 3. depicts an example of the human evaluation process. Second, we assessed how the model made its recommendations using the retrieved information. The PUMA scale classifies COPD risk on a 9 point scale (Supplementary Material 1.) According to PUMA scale validations in Latin America [12], [19], [23] and Asia [20], a threshold of five points was optimal for detecting more subclinical cases. For our study, we used the same threshold. To evaluate the recommendation capability per se, PANDORA's predictions were compared to true values in MIMIC-IV: 1 (COPD risk) or 0 (no COPD risk). The punctuation was evaluated as correct or incorrect and described using relative and absolute frequencies (Table III.) **Second Strategy:** For further evaluation, we also created synthetic clinical cases using the GPT family based on the guide designed by our team's medical lead simulating a Colombian-based outpatient consultation record (Supplementary Material 4). We employed nine COPD differential diagnoses (Supplementary Material 5) to test the model's accuracy in extracting and applying PUMA when respiratory symptoms were similar among cases. We created one clinical case for each and gave them as examples to the GPT framework to elaborate 100 synthetic clinical cases (Supplementary Material 6). We evaluated extraction, recommendation, and punctuation capabilities using confusion matrices and the model's operating characteristics. Several EHRs (>50%) stated a personal history of COPD; therefore, we instructed PANDORA to extract this feature and automatically classify them as presenting COPD risk, independent of their PUMA score (Table III). This strategy was subsequently applied to the evaluations in MIMIC.

III. Results

A. Semantic Scores

Optimal performance evidences PANDORA's high-quality text summarization and demonstrates its ability to produce a coherent answer with an understanding of the case provided (Table I.). Semantic scores will be further discussed in the next section.

TABLE I
SEMANTIC SCORES FOR EXTRACTION EVALUATION USING THE EHR-DS-QA DATASET.

Metric		Score
BERTScore		0.911
SemanticScore		0.925
RelevanceScore		0.901
Judge Alignment Metrics	Accuracy	0.838
	Recall/Sensitivity	0.838
	F1 Score	0.912

B. Extraction, scoring and recommendation with MIMIC-IV database as standard

The human evaluation of extraction and scoring capabilities using PUMA was performed in a subset of 102 QA from the human-evaluated portion of the EHR-DS-QA (Table II.). Age and smoking history were excluded; they were unavailable secondary to de-identification processes in the initial MIMIC-IV database. Adequate extraction was demonstrated in 100% of the remaining 615 QA-pairs, while 581 (94.47%) were classified as presenting accurate scoring. Most of the 34 mistakes occurred when the model did not recognize previous COPD diagnoses. Table III. summarises performance metrics for recommendation capability when instructing PANDORA to determine COPD risk if a prior diagnosis was present, disregarding the PUMA score. Sensitivity increased by 66% with this measure. Specificity decreased by 22.5%. Supplementary Materials 7 and 8 show the confusion matrix for each approach.

TABLE II
HUMAN EVALUATION OF EXTRACTION AND SCORING CAPABILITIES

Experiment	Extraction		Scoring	
	Correct answers (n)	Accuracy	Correct answers (n)	Accuracy
MIMIC-IV (n = 615)	615	1.0	581	0.945
Synthetic Cases (n = 700)	697	0.996	659	0.941

The results in Table II present the accuracy of the test as the proportion of correct answers on each capability, using the MIMIC-IV database and synthetic cases as standard and based on the PUMA scale.

C. Extraction, scoring and recommendation with synthetic cases as standard

The human evaluation of 100 synthetic cases revealed optimal extraction and scoring capabilities. PANDORA accurately extracted the information in 99.6% of cases (697/700) and assigned the correct score in 94% of cases (658/700) (Table II.) Two of the three extraction mistakes are associated with the model's assumption that a history of COPD means the patient's expectoration is chronic. The other mistake occurred when PANDORA failed to recognize the phrase "worsening of the symptoms over the past few months", signifying chronicity. Of the 41 scoring mistakes, 28 correspond to the model's misunderstanding of threshold values for smoking history categories in PUMA. Of the remaining 13 scoring errors, 3 were incorrect age classifications. The remaining errors (10) are definitions; the model makes mistakes without specific timeframes to define symptoms' chronic or acute nature. PANDORA's outcome is "COPD" (PUMA score $\geq$ 5) or "other" ($<$ 5). Table III. shows the performance metrics for recommendations in the synthetic cases. PANDORA reached a sensitivity of 100% and specificity of 20% since it classified 64 synthetic cases as false positives. Supplementary Material 10. presents the distribution of the PUMA features in these cases.

TABLE III
PERFORMANCE METRICS FOR RECOMMENDATION CAPABILITIES OF PANDORA WHEN USING THE MIMIC-IV DATABASE AND SYNTHETIC CASES AS STANDARD, ACCORDING TO COPD'S PREVIOUS DIAGNOSIS.

Metric	MIMIC-IV		Synthetic Cases
	Considering the history of COPD	Not considering the history of COPD	
Sensitivity	0.855	0.194	1.0
Specificity	0.700	0.925	0.200
Precision	0.815	0.800	0.238
Accuracy	0.794	0.480	0.360
F1 score	0.835	0.312	0.385
Cohen's Kappa	0.790	0.470	0.347

IV. DISCUSSION

LLMs have proven their ability to extract information from text sources, such as EHRs [24]–[27]. They offer the opportunity to recover unstructured data for research and clinical decisions without manually organizing information into databases. Available models offer separate retrieval and recommendation capabilities [28]–[31]. Others are ML algorithms focused on diagnosis and risk prediction from tabular data [27], [31]–[33]. Yu-Tzu Lee explored the integration of the extraction capability and recommendation in their thesis paper (https://arxiv.org/abs/2407.10453). He enhanced medication recommendations using LLMs to extract information from free-text notes. The study used the MIMIC-III datasets and tested seven LLMs. The G-BERT model reached AUPRC of 77.6% [33]. Another approach was proposed by Ozan Unlu et al. [34], who developed a model that would retrieve information from EHRs according to predefined selection criteria to select appropriate candidates for the clinical study COPILOT-HF (ClinicalTrials.govNCT05734690). Their assistant extracted information according to inclusion criteria and used exclusion criteria to recommend final candidates. Compared to human evaluation, the model's sample selection had 92% sensitivity and 94% specificity. The current literature does not describe a modular LLM capable of producing information retrieval and implementing human-made clinical algorithms using that recovered information. Therefore, this study is our initial approach to validating a model for several purposes: information retrieval, integration with clinical guidelines, and recommendations based on them. PANDORA demonstrated

high-quality text summarization, generating relevant responses that maintain meaning, context, and alignment with input queries. We conclude that the model uses pertinent data in its recommendations. A Judge Alignment Metric (JAM) was also applied; the good marks imply that our model performs well semantically in the presence of state-of-the-art LLMs and their capabilities for evaluating input and output. State-of-the-art (SoTA) refer to LLMs with reported highest accuracies. The current SoTA performance for GPT-4 is 90.2% and 85.4% for Med-PaLM 2. Accuracies are calculated from reference standard datasets of QA pairs from medical board exam questions or telemedicine interactions [35], [36], which provide revised answers for quantitative analyses [26], [29], [37]–[40]. The outcome is a binary classification of correct or incorrect, according to the nature of the dataset, and results are usually presented using accuracies [21], [41], [42]. Although no scientific consensus exists on a Gold Standard in LLM evaluation [25], human expert revisors are the desirable assessment [43], [44]. The JAM strategy's rationale was that human eval, although ideal, is often unfeasible, given the databases' sizes, making it costly, time-consuming and exhausting for the professionals involved. The MIMIC-IV database is widely used as a benchmark for training and evaluating LLMs in healthcare [13]–[15]. We used it as the reference for the qualitative assessment of PANDORA, ensuring the model's exposure to raw, intricate data typical in clinical notes. PANDORA adequately extracts data from unstructured sources, such as medical records. To validate this, we performed a Human Evaluation using a subset of the MIMIC-IV database. Our model demonstrated perfect extraction capability (100%); additionally, we explored its ability to interact with a validated risk calculator (scoring capability), the PUMA scale for COPD risk assessment, which revealed that our model understands the rationale behind the scoring rules. Regarding the recommendation capability, it could point out the risk for COPD in all synthetic cases and 89% of the MIMIC-IV notes. Still, the specificity for this last capability was 20% for synthetic cases and 70% using MIMIC-IV. These results might be due to using a sensitive COPD case-finding tool such as PUMA. The PUMA Study [45] was validated in adult, heavy-smoker Latin American population. Its threshold for COPD risk ($\geq$ 5) was set accordingly. This raises the question of the tool's applicability in a population with different baseline characteristics: could a different threshold be used? The scale's precision is extensively described elsewhere [19], [20], [23], [46]. However, it is plausible that setting a higher threshold could enable PUMA to be used in a broader population base. Supplementary Material 11. depicts the operative characteristics of the PUMA scale in our population sample, evaluating thresholds from 1 to 9. Still, both sources of clinical cases (MIMIC and synthetic) were tested using the same scale, which does not explain the 50% difference in specificity (70% for MIMIC, 20% for synthetic cases). The explanation lies in the intrinsic characteristics of the sources. The MIMIC-IV was a real-world dataset with patients from an ICU. Their profile is that of a sicker patient. Unfortunately, we could not evaluate smoking habits or age, as they were part of the erased information in the de-identification process, mandatory for public access to sensitive information. This fact makes the score results not comparable to those from the synthetic cases, which had all the information. The synthetic cases were created to present the scale with cases that showed similar symptoms to COPD. The subject was a healthier outpatient without severe symptoms. Here, PANDORA misclassified 64 cases as COPD, which shows the adequate extraction and scoring of the model but the high sensitivity and low specificity of PUMA. To elaborate on this, take the example of a 62-year-old man (3 points) presenting with dyspnea (1 point) due to heart failure, with prior spirometry (1 point); this patient would be classified as COPD if he is only evaluated using PUMA. We found 62 synthetic cases with a personal history of COPD. We used this criterion for subsequent analysis on MIMIC, adding an item to the extraction phase prompted by "Does the patient have a smoking history?". This approach improved sensitivity by 66%. A remarkable contribution of the present research is the application of the "Self-Thought Evaluator" through the development and use of a JAM [47]. Our experience was remarkable since it allows for assessing any number of queries expediting model analyses, comparisons, and improvements. The downside is that AI judges could make assessment mistakes by amplifying possible biases introduced during their creation. We could not assess the complete PUMA questions in the MIMIC-IV since age and smoking history were missing in all cases. Consequently, we performed a human evaluation of each case to ensure every other capability was conserved. Human evaluation is the desired standard, but it is a cumbersome task. We followed the experts' recommendation of using 100 cases for the evaluation [43]. Additionally, synthetic cases may not fully represent the variability of real-world clinical data, and EHR structure may vary by country or institution. To control this, we based our clinical cases on the recommendations of medical professionals and government guidelines. We explicitly instructed the LLM to consider all races, sexes, occupations, nationalities, and social backgrounds for egalitarian inclusion. Still, continuous human monitoring and re-evaluation are necessary to ensure and supervise PANDORA's and other LLMs' outcomes. In conclusion, implementing PANDORA will allow research and clinical tasks without structured databases. Early or missed diagnoses of several diseases could be achieved by combining the vast capabilities developed in NLP with traditional, validated clinical scores and updated clinical guidelines containing the best available evidence and experts' consensuses worldwide.

REFERENCES

[1] Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. BMC Med Res Methodol. 2019 Mar 11;19(1):53.

[2] Terris DD, Litaker DG, Koroukian SM. Health state information derived from secondary databases is affected by multiple sources of bias. J Clin Epidemiol. 2007 Jul 1;60(7):734–41.

[3] Kong HJ. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res. 2019 Jan;25(1):1–2.

[4] Fisher C, Lauria E, Chengalur-Smith S. Introduction to Information Quality. AuthorHouse; 2012. 277 p.
[5] Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 2014 Apr 1;1(1):2053951714528481.
[6] Provost F, Fawcett T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data. 2013 Mar;1(1):51–9.
[7] Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. Clin Pharmacol Ther. 2020;107(4):871–85.
[8] Yao Q, Wang M, Chen Y, Dai W, Hu YQ, Li YF, et al. Taking the Human out of Learning Applications: A Survey on Automated Machine Learning.
[9] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015 Jul 17;349(6245):255–60.
[10] Chen Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanit Soc Sci Commun. 2023 Sep 13;10(1):1–12.
[11] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health. 2023 Jun 22;2(6):e0000278.
[12] López Varela MV, Montes de Oca M, Rey A, Casas A, Stirbulov R, Di Boscio V, et al. Development of a simple screening tool for opportunistic COPD case finding in primary care in Latin America: The PUMA study. Respirol Carlton Vic. 2016 Oct;21(7):1227–34.
[13] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023 Jan 3;10(1):1.
[14] Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: Deidentified free-text clinical notes [Internet]. PhysioNet; [cited 2024 Aug 21]. Available from: https://physionet.org/content/mimic-iv-note/2.2/
[15] Johnson AEW, Pollard TJ, Shen L, Lehman L wei H, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3(1):160035.
[16] Ministerio de Salud y Protección Social. Interoperabilidad de Datos de la Historia Clínica en Colombia Términos y siglas [Internet]. 2019 [cited 2024 Sep 11]. Available from: https://www.minsalud.gov.co/ihc/Documentos%20compartidos/ABC-IHC.pdf
[17] Ciapponi A, Alison ,Lee, Agustina ,Mazzoni, Demián ,Glujovsky, Silvana ,Cesaroni, and Edgardo S. The Epidemiology and Burden of COPD in Latin America and the Caribbean: Systematic Review and Meta-Analysis. COPD J Chronic Obstr Pulm Dis. 2014 Jun 1;11(3):339–50.
[18] Caballero A, Torres-Duque CA, Jaramillo C, Bolívar F, Sanabria F, Osorio P, et al. Prevalence of COPD in five Colombian cities situated at low, medium, and high altitude (PREPOCOL study). Chest. 2008 Feb;133(2):343–9.
[19] Bastidas G. AR, Estupiñán B. MF, Arias B. JS, Estrada H. M, López O. J, Mateus M. SL, et al. Validación externa y reproducibilidad del cuestionario PUMA para el diagnóstico de EPOC en una población latinoamericana: Validación externa del cuestionario PUMA. Rev Chil Enfermedades Respir. 2022 Mar;38(1):11–9.
[20] PUMA screening tool to detect COPD in high-risk patients in Chinese primary care-A validation study - PubMed [Internet]. [cited 2024 Aug 26]. Available from: https://pubmed.ncbi.nlm.nih.gov/36084011/
[21] Kotschenreuther K. EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Discharge Summaries for Enhanced Medical Information Retrieval Systems [Internet]. PhysioNet; [cited 2024 Aug 28]. Available from: https://physionet.org/content/ehr-ds-qa/1.0.0/
[22] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena [Internet]. arXiv; 2023 [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2306.05685
[23] Herrera AC, Oca MM de, Varela MVL, Aguirre C, Schiavi E, Jardim JR, et al. COPD Underdiagnosis and Misdiagnosis in a High-Risk Primary Care Population in Four Latin American Countries. A Key to Enhance Disease Diagnosis: The PUMA Study. PLOS ONE. 2016 Apr 13;11(4):e0152266.
[24] Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing [Internet]. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022 [cited 2024 Sep 12]. p. 1998–2022. Available from: https://aclanthology.org/2022.emnlp-main.130
[25] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–40.
[26] Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns [Internet]. 2024 Mar 8 [cited 2024 Jul 15];5(3). Available from: https://www.cell.com/patterns/abstract/S2666-3899(24)00042-4
[27] Wang B, Lai J, Cao H, Jin F, Li Q, Tang M, et al. Enhancing Real-World Data Extraction in Clinical Research: Evaluating the Impact of the Implementation of Large Language Models in Hospital Settings [Internet]. 2023 [cited 2024 Aug 29]. Available from: https://www.researchsquare.com/article/rs-3644810/v2
[28] BioBERT: a pre-trained biomedical language representation model for biomedical text mining — Bioinformatics — Oxford Academic [Internet]. [cited 2024 Aug 13]. Available from: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506
[29] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems [Internet]. arXiv; 2023 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2303.13375
[30] Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: Language Models for Dialog Applications [Internet]. arXiv; 2022 [cited 2024 Aug 27]. Available from: http://arxiv.org/abs/2201.08239
[31] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways [Internet]. arXiv; 2022 [cited 2024 Aug 27]. Available from: http://arxiv.org/abs/2204.02311
[32] Wiest IC, Wolf F, Leßmann ME, Treeck M van, Ferber D, Zhu J, et al. LLM-AIx: An open source pipeline for Information Extraction from unstructured medical text based on privacy preserving Large Language Models [Internet]. medRxiv; 2024 [cited 2024 Sep 12]. p. 2024.09.02.24312917. Available from: https://www.medrxiv.org/content/10.1101/2024.09.02.24312917v1
[33] Lee YT. Enhancing Medication Recommendation with LLM Text Representation [Internet]. arXiv; 2024 [cited 2024 Sep 12]. Available from: http://arxiv.org/abs/2407.10453
[34] Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-Augmented Generation–Enabled GPT-4 for Clinical Trial Screening. NEJM AI. 2024 Jun 27;1(7):AIoa2400181.
[35] Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In: Proceedings of the Conference on Health, Inference, and Learning [Internet]. PMLR; 2022 [cited 2024 Aug 13]. p. 248–60. Available from: https://proceedings.mlr.press/v174/pal22a.html
[36] Labrak Y, Bazoge A, Dufour R, Rouvier M, Morin E, Daille B, et al. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain [Internet]. arXiv; 2023 [cited 2024 Jul 28]. Available from: http://arxiv.org/abs/2304.04280
[37] Mukherjee S, Gamble P, Ausin MS, Kant N, Aggarwal K, Manjunath N, et al. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare [Internet]. arXiv; 2024 [cited 2024 May 24]. Available from: http://arxiv.org/abs/2403.13313
[38] Rahimi H, Hoover JL, Mimno D, Naacke H, Constantin C, Amann B. Contextualized Topic Coherence Metrics [Internet]. arXiv; 2023 [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2305.14587
[39] Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc JAMIA. 2018 Jul 11;25(9):1248–58.
[40] Papers with Code - Best practices for the human evaluation of automatically generated text [Internet]. [cited 2024 Aug 12]. Available from: https://paperswithcode.com/paper/best-practices-for-the-human-evaluation-of
[41] Suri H, Zhang Q, Huo W, Liu Y, Guan C. MeDiaQA: A Question Answering Dataset on Medical Dialogues [Internet]. arXiv; 2021 [cited 2024 Jul 28]. Available from: http://arxiv.org/abs/2108.08074
[42] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering [Internet]. arXiv; 2019 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/1909.06146
[43] van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. In: van Deemter K, Lin C, Takamura H, editors. Proceedings of the 12th International Conference on Natural Language Generation [Internet]. Tokyo, Japan: Association for Computational Linguistics; 2019 [cited

2024 Aug 10]. p. 355–68. Available from: https://aclanthology.org/W19-8643
[44] Abeysinghe B, Circi R. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches [Internet]. arXiv; 2024 [cited 2024 Aug 10]. Available from: http://arxiv.org/abs/2406.03339
[45] Schiavi E, Stirbulov R, Hernández Vecino R, Mercurio S, Di Boscio V. COPD Screening in Primary Care in Four Latin American Countries: Methodology of the PUMA Study. Arch Bronconeumol Engl Ed. 2014 Nov 1;50(11):469–74.
[46] Validation of the PUMA score for detecting COPD in a primary care population at the Hospital Maciel, Montevideo — European Respiratory Society [Internet]. [cited 2024 Aug 29]. Available from: https://erj.ersjournals.com/content/50/suppl_61/PA1198
[47] Wang T, Kulikov I, Golovneva O, Yu P, Yuan W, Dwivedi-Yu J, et al. Self-Taught Evaluators [Internet]. arXiv; 2024 [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2408.02666