npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01233-2

Privacy-preserving large language models for structured medical information retrieval

Isabella Catharina Wiest[1,2], Dyke Ferber[2,3], Jiefu Zhu[2], Marko van Treeck[2], Sonja K. Meyer[4], Radhika Juglan[2], Zunamys I. Carrero[2], Daniel Paech[5,6], Jens Kleesiek[7,8,9], Matthias P. Ebert[1,10,11], Daniel Truhn[12] & Jakob Nikolas Kather[2,3,13]

Most clinical information is encoded as free text, not accessible for quantitative analysis. This study presents an open-source pipeline using the local large language model (LLM) "Llama 2" to extract quantitative information from clinical text and evaluates its performance in identifying features of decompensated liver cirrhosis. The LLM identified five key clinical features in a zero- and one-shot manner from 500 patient medical histories in the MIMIC IV dataset. We compared LLMs of three sizes and various prompt engineering approaches, with predictions compared against ground truth from three blinded medical experts. Our pipeline achieved high accuracy, detecting liver cirrhosis with 100% sensitivity and 96% specificity. High sensitivities and specificities were also yielded for detecting ascites (95%, 95%), confusion (76%, 94%), abdominal pain (84%, 97%), and shortness of breath (87%, 97%) using the 70 billion parameter model, which outperformed smaller versions. Our study successfully demonstrates the capability of locally deployed LLMs to extract clinical information from free text with low hardware requirements.

It is estimated that 80% of clinical data exists in an unstructured format[1]. Unstructured data includes data in non-tabular formats, such as images, video, and text, that are not accessible for quantitative analysis. This "dark matter" of healthcare data is currently unusable for quantitative computational analysis. While deep learning methods have made structured data from Electronic Health Records (EHRs) usable for individual risk prediction[2], can make diagnoses and extract biomarkers from radiology or histopathology images[3,4], natural language has not been widely used as a source to extract structured information. Making an unstructured data resource readable for downstream tasks has a variety of benefits, such as improvements in individual healthcare outcomes[5], the possibility to obtain scientific insights[6], and improvements in billing processes and quality control[7].

In natural language processing (NLP), computational methods are applied to unstructured text. Medical applications of NLP have been explored for decades[8,9], but real-world applications are still very rare. However, real-world data analysis is increasingly being recognized and implemented for timely evidence generation, making the need to extract real-world data from text even more pressing[10]. Several hurdles have been discussed for NLP in healthcare, among them the lack of annotated datasets and user-centered design as well as hand-crafted over-engineered software pipelines which lack scalability[11,12]. Large language models (LLMs) have impacted this field: they are transformer neural networks which are trained on large bodies of unstructured text data with self-supervised learning (SSL)[13–16]. LLMs are foundation models which can be applied to a broad range of tasks without having been explicitly trained for these tasks. This

[1]Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany. [2]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany. [3]Department of Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg University Hospital, Heidelberg, Germany. [4]Department of Surgery I, University Hospital Würzburg, Würzburg, Germany. [5]German Cancer Research Center, Division of Radiology, Heidelberg, Germany. [6]University Hospital Bonn, Clinic for Neuroradiology, Bonn, Germany. [7]Institut für KI in der Medizin (IKIM), Universitätsmedizin Essen, Girardetstr. 2, 45131 Essen, Germany. [8]Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen (WTZ), 45122 Essen, Germany. [9]TU Dortmund University, Department of Physics, Otto-Hahn-Straße 4, 44227 Dortmund, Germany. [10]DKFZ Hector Cancer Institute at the University Medical Center, Mannheim, Germany. [11]Molecular Medicine Partnership Unit, EMBL, Heidelberg, Germany. [12]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany. [13]Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, 01307 Dresden, Germany.
e-mail: jakob_nikolas.kather@tu-dresden.de

"zero-shot" application, where LLMs are tasked with a potentially unseen problem, changes the conventional wisdom in medical artificial intelligence by which a model for a certain task needs to be trained on a large dataset representing this specific task[17]. In particular, the LLM Generative Pre-trained Transformer (GPT) and its user interface ChatGPT, have demonstrated remarkable proficiency in structuring text and extracting relevant information in a quantitative way[18]. Their capabilities could revolutionize the way we comprehend and process vast quantities of healthcare data[19–21]. For example, GPT-4 has been used to extract structured clinical information from free text reports in radiology[18], pathology and medicine[22].

However, these LLMs run as cloud services and using them requires the transfer of privileged information to remote servers. This brings along immense legal and ethical challenges, especially in the European Union (EU), where the export of personal health data is not legally permitted[23,24]. Ideally, LLMs should run on-premise of healthcare institutions, potentially even at the point of care[25,26]. However, this requires software pipelines using lightweight LLMs such as quantized LLMs, which are currently not validated for medical tasks. Quantized models have lower numerical precision of the model parameters and have lower graphics processing unit (GPU) memory consumption than unquantized models, allowing for easier integration with existing hospital hardware. Here, we therefore aimed to build and validate a fully automated pipeline for end-to-end processing of clinical text data which uses locally deployable LLMs and can potentially be used at the point of care. We investigated the capabilities of our new pipeline with a task of high clinical importance: the extraction of specific clinical features from medical free text, using the example of features that help detect decompensated liver cirrhosis. Approximately 1% of the population in the EU has liver cirrhosis[27] and decompensation is one of the most common emergencies faced by these patients[28]. Decompensation is often overlooked initially, but can be a turning point in the prognosis of cirrhotic patients, thus early identification and management are crucial to improve patient outcomes[29]. Automatic detection of decompensated liver cirrhosis, enhanced with features extracted from free text, provides a more robust basis for future early warning systems. In addition, this approach could facilitate retrospective analysis of clinical data for scientific, quality control or billing purposes, and it could be applied to other areas of medicine too.

Results

Key medical features are unevenly represented in medical histories

Our analysis of the Llama 2 model's data extraction capabilities from text reports focused on five key medical features: liver cirrhosis, ascites, abdominal pain, shortness of breath, and confusion. We found that the frequency of these features varied significantly across the reports. Abdominal pain and shortness of breath were frequently documented in the data ("abdominal pain": $N = 209/500$ reports and "shortness of breath": $N = 130/500$ reports). However, liver cirrhosis and ascites were less prevalent ("liver cirrhosis": $N = 1/500$, since liver cirrhosis was sometimes explicitly mentioned in other combinations (e.g. "HCV cirrhosis"), we also performed a keyword search on the word stem "cirrhos": $N = 29/500$ reports, "ascites": $N = 20$), mentioned in only about 5% of cases, as detailed in Fig. 1.

While liver cirrhosis and ascites were explicitly mentioned when present (ascites was mentioned in 20 reports and also present in 20 reports), making their detection more straightforward, the documentation of abdominal pain, shortness of breath, and confusion often required more nuanced interpretation, as these symptoms were described in multiple ways by physicians. Abdominal pain, shortness of breath, and confusion were not always explicitly stated but could be inferred from contextual information. For example, abdominal pain might be indicated through a variety of descriptors or understood from the absence of certain findings, e.g., "pain in the RUQ" stands for "pain in the right upper quadrant of the abdomen" thus indicating the presence of abdominal pain.

Similarly, shortness of breath and confusion, while not always directly stated, could be inferred from contextual clues or specific medical terminology used in the reports. This implies that accurately identifying such implicit features demands a nuanced understanding of medical language and context, as well as some level of clinical expertise. For example, a statement like "10-point review of systems negative" implies the absence of

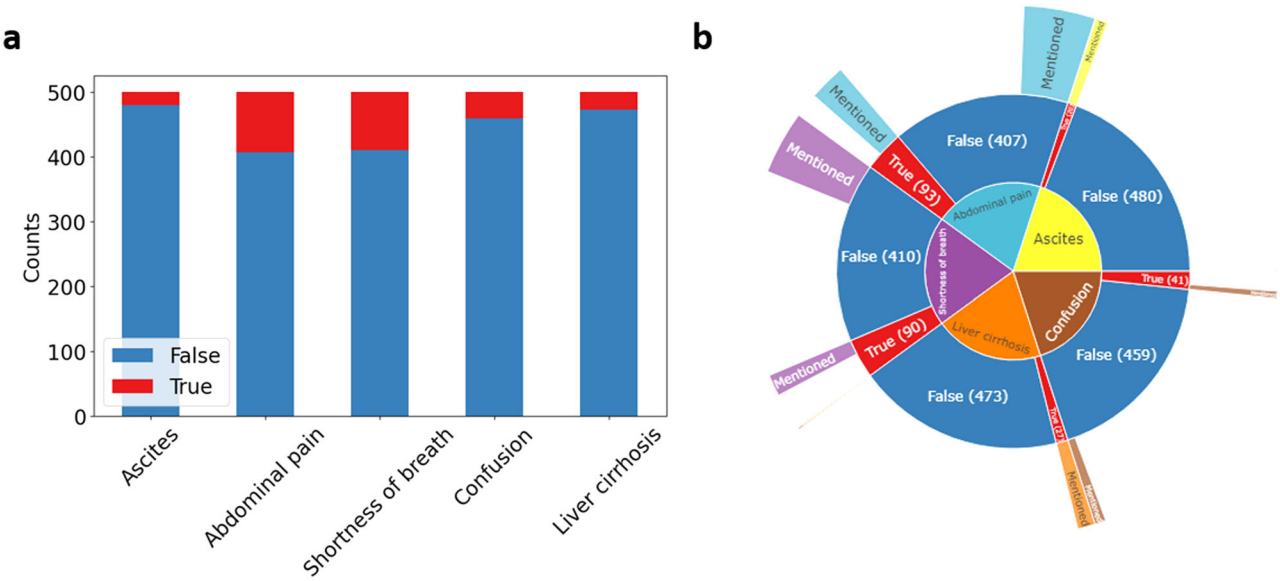

Fig. 1 | Feature distribution in 500 MIMIC present medical histories. a The bar chart visualizes data from 500 present medical history reports extracted from the MIMIC-IV database. It displays the counts for five extracted features, with "true" counts in red and "false" in blue. **b** The sunburst plot indicates the amount of reports, in which the features' term is explicitly mentioned as a share of false and true counts. Liver cirrhosis and ascites are the features with the highest share of explicitly mentioned features, with every mention aligning with a "true" classification in the ground truth evaluation. Abdominal pain and shortness of breath were most frequently mentioned over all reports. "Explicit features" are consistently described with identical terminology (e.g., ascites, cirrhosis), whereas "implicit features" vary in description (e.g., shortness of breath: "SOB," "difficulties in breathing," "dyspnea").

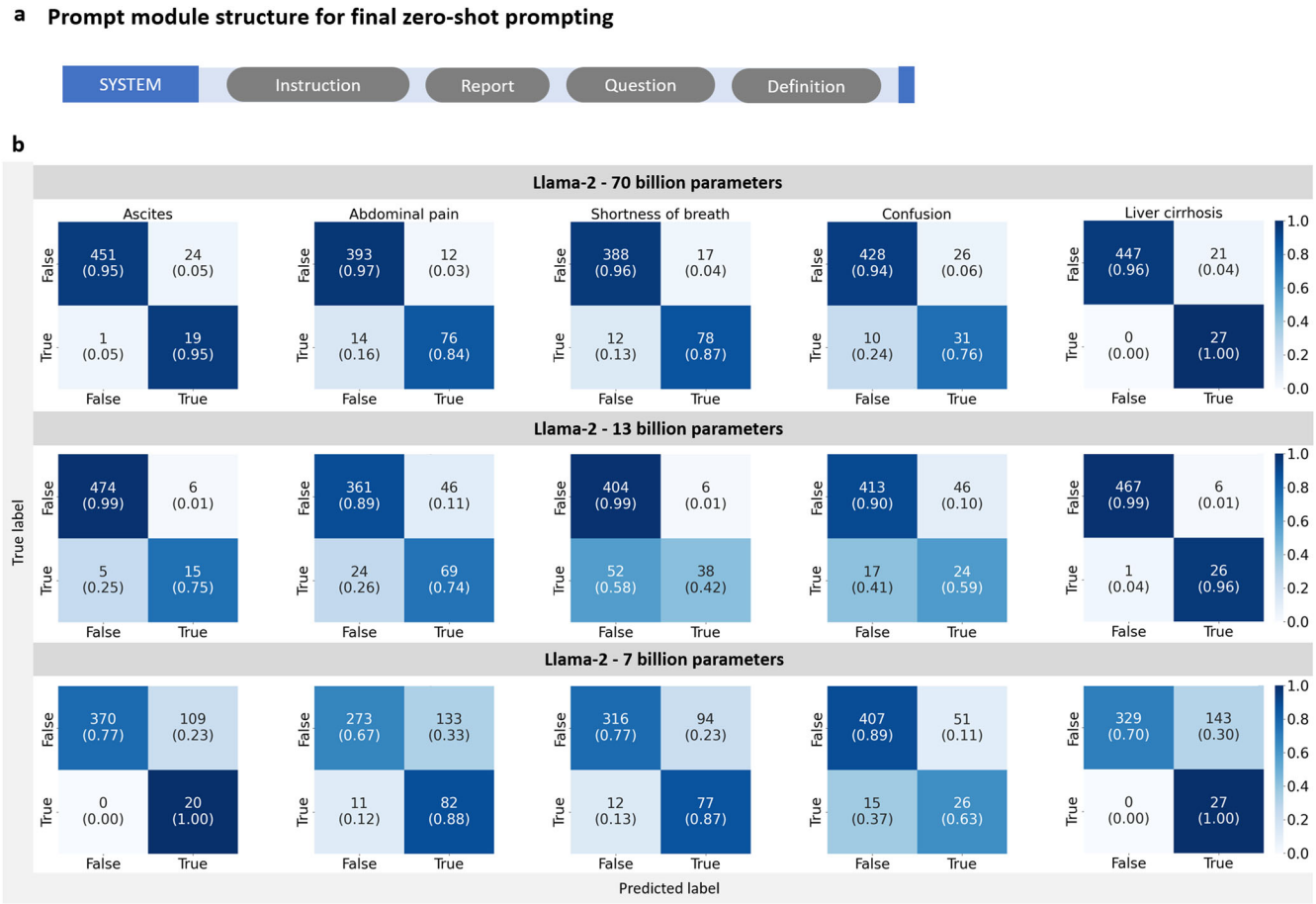

Fig. 2 | Confusion matrices for extracted features with zero-shot prompting. **a** shows the prompt modules used for zero shot prompting. The detailed instruction was included, followed by a report and the corresponding instruction formulated as a question. This was followed by a definition of the features to be extracted. **b** The confusion matrices visualize the performance of the Llama 2 models with 7 billion, 13 billion and 70 billion parameters in retrieving the presence or absence of the five features ascites, abdominal pain, shortness of breath, confusion and liver cirrhosis in all $n = 500$ medical histories from MIMIC IV. All matrices are divided into four quadrants with the two labels "true" or "false" in each axis. The x-axis depicts the predicted labels, the y-axis depicts the true labels. The confusion matrices are normalized to show proportions, where each cell represents the fraction of predictions within the actual class. Values along the diagonal indicate correct predictions (true positives and true negatives), while off-diagonal values represent misclassifications (false positives and false negatives). The sum of each row's fractions equals 1, indicating the proportion of predictions for each actual class. The "n" values represent the absolute number of observations in each category. In the top left matrix, the extraction of ascites with the 70b model is shown. The top left quadrant (true negatives) shows a high score of 0.95, indicating a high rate of correct predictions for non-cases of ascites. The top right quadrant (false positives) has a score of 0.05, suggesting few cases were incorrectly predicted as having ascites. The bottom left quadrant (false negatives) has a score of 0.05, indicating few cases were incorrectly identified as not having ascites. Finally, the bottom right quadrant (true positives) shows a high score of 0.95, which means a high rate of correct predictions for actual cases.

symptoms like shortness of breath, abdominal pain, and confusion, requiring the model to interpret these indirect clues effectively.

Llama 2 is able to extract relevant information from unstructured text

In our assessment, the 70b model displayed remarkable proficiency. Sensitivity of detecting liver cirrhosis and ascites was 100% and 95%, respectively. For abdominal pain and shortness of breath, sensitivities were lower with 84% and 87%, respectively. Confusion was the symptom that was most difficult to extract for the LLM with a sensitivity of only 76%. Specificity for liver cirrhosis was 96%, for ascites 95% and even higher for abdominal pain (97%), shortness of breath (96%) and confusion (94%). Confusion matrices are shown in Fig. 2.

One-shot prompting yielded slightly better results with higher sensitivities (ascites: 95%, abdominal pain: 92%, shortness of breath: 83%, confusion: 88% and liver cirrhosis 100%) and specificities (ascites: 99%, abdominal pain: 92%, shortness of breath: 96%, confusion: 94% and liver cirrhosis 97%) (Fig. 3 and Table 2).

The models with more parameters performed better, with the most substantial increase in accuracy from the Llama 2 7b to 13b model (Table 1 and Fig. 4). For implicit features, the 70b model yielded the highest accuracy. The 7b model faced challenges in accurately identifying false classifications. For example, in one case, the model stated "She had confusion present at admission," even though there was no information about confusion in the report. Similarly, the model interpreted the feature "ascites" as present, but the report only stated "(…) healthy female with incidental finding of right renal mass suspicious for RCC (…)". This hallucination was particularly present in smaller models such as Llama 2 7b. All models presented a high negative predictive value. Precision and specificity tended to improve most from 7b to 13b parameter model size. Recall was best in the explicitly mentioned features (Tables 1 and 2).

Prompt engineering enhances accuracy, especially in smaller sized models

In our initial test with the 7b model, we used a combination of a system prompt with general instructions and a user prompt containing the report and questions (prompting strategy details in Supplementary Figs. 2 and 3). Including a one-shot example in the prompt slightly enhanced the model's accuracy except for the feature abdominal pain (Supplementary Fig. 2). The

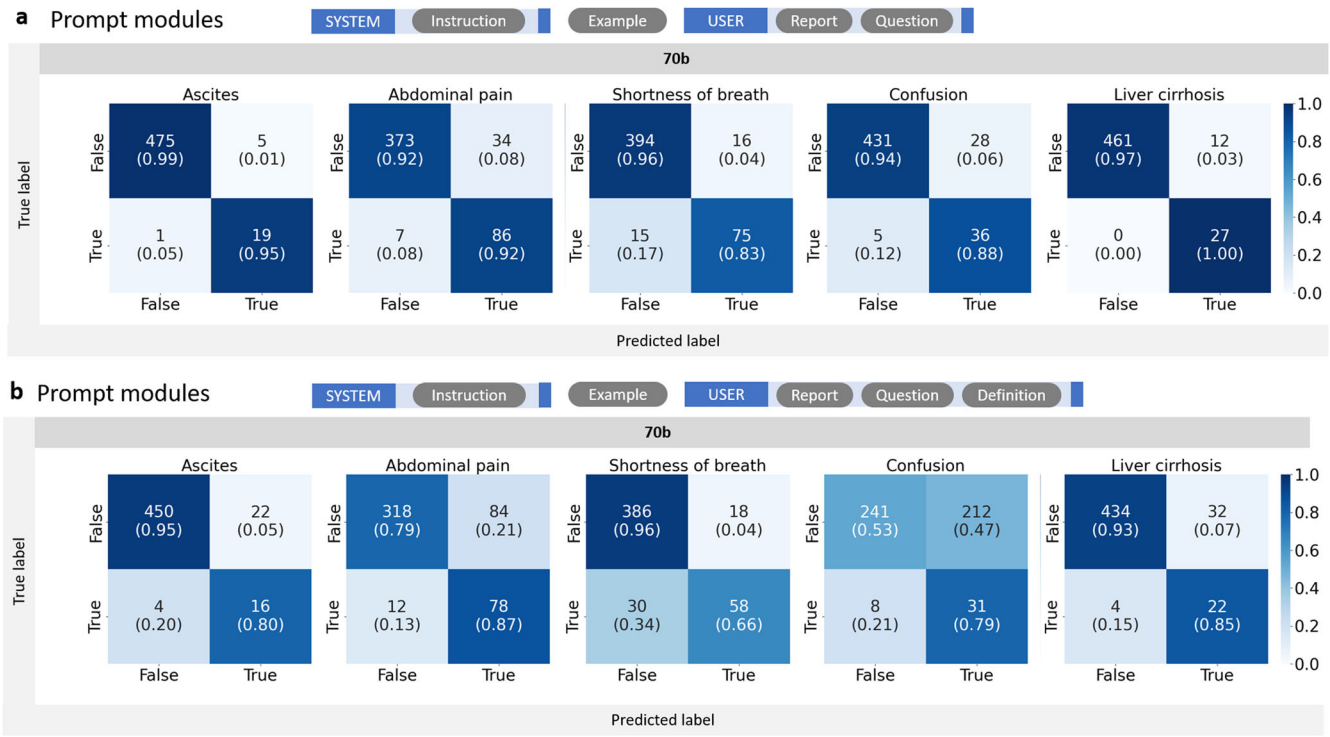

Fig. 3 | Confusion matrices for extracted features with one-shot prompting. The confusion matrices visualize the performance of the Llama 2 models with 70 billion parameters in retrieving the presence or absence of the five features ascites, abdominal pain, shortness of breath, confusion and liver cirrhosis in all $n = 500$ medical histories from MIMIC IV. All matrices are divided into four quadrants with the two labels "true" or "false" in each axis. The *x*-axis depicts the predicted labels, the *y*-axis depicts the true labels. The confusion matrices are normalized to show proportions, where each cell represents the fraction of predictions within the actual class. Values along the diagonal indicate correct predictions (true positives and true negatives), while off-diagonal values represent misclassifications (false positives and false negatives). The numbers indicate absolute counts, the figure in brackets indicate fractions. The sum of each row's fractions equals 1, indicating the proportion of predictions for each actual class. **a** shows the best one-shot prompt architecture and results. Whereas adding definitions, which improved performance with zero-shot prompting, deteriorated the results for one-shot prompting (**b**).

Table 1 | Model performance—zero-shot prompting with definitions

	Sensitivity			Specificity			Positive predictive value			Negative predictive value			Accuracy		
	7b	13b	70b	7b	13b	70b	7b	13b	70b	7b	13b	70b	7b	13b	70b
Ascites	1.00	0.75	0.95	0.77	0.99	0.95	0.16	0.71	0.44	1.00	0.99	1.00	0.78	0.98	0.95
Abdominal pain	0.88	0.74	0.84	0.67	0.89	0.97	0.38	0.60	0.86	0.96	0.94	0.97	0.71	0.86	0.95
Shortness of breath	0.87	0.42	0.87	0.77	0.99	0.96	0.45	0.86	0.82	0.96	0.89	0.97	0.79	0.88	0.94
Confusion	0.63	0.59	0.76	0.89	0.90	0.94	0.34	0.34	0.54	0.96	0.96	0.98	0.87	0.87	0.93
Liver cirrhosis	1.00	0.96	1.00	0.70	0.99	0.96	0.16	0.81	0.56	1.00	1.00	1.00	0.71	0.99	0.96

Comparing three versions of Llama 2, the largest (70b) model showed the highest performance whereas the smallest (7b) model performed worst. The 13b and 70b models show higher accuracy across all conditions when compared to the 7b model.

human instructions in the Llama prompt needed to be indicated within specific tags ([INST],[/INST]). Notably, the one-shot example needed to be excluded from the instruction section, otherwise the performance deteriorated substantially, because the model answered the questions with the example given. Requesting an excerpt from the text followed by a binary answer (Chain-of-thought prompting) did not yield improved results. We found deteriorated accuracy for the features ascites (−25 percentage points (ppts)), abdominal pain (−6 ppts) and confusion (−5 ppts). The features shortness of breath (+1 pp) and liver cirrhosis (+15 ppts) improved slightly (Supplementary Fig. 2). For explainability reasons, we nevertheless forced the model with the grammar (which is outlined in detail in the github repository) to provide, first, an excerpt, and only then the binary outcome and found that this did not adversely affect performance.

Providing definitions for all features only improved the extraction of the more implicitly mentioned features shortness of breath and abdominal pain, but deteriorated the extraction of explicitly mentioned features. Subsequent testing involved consolidating both the report and question components within the system prompt, instead of dividing them between system and user prompts. This change resulted in improved performance for the 7b model, whereas this trend was not consistently present for the 70b model. Whereas system prompting improved the accuracy of detecting ascites by 4 ppts, liver cirrhosis by 7 ppts, abdominal pain by 6 ppts, shortness of breath by 4 ppts and confusion by 2 ppts in the 7b model, the system prompting effect was less consistent in the 70b model, leading to improvement for ascites detection by 9 ppts, liver cirrhosis by 1 pp, abdominal pain by 1 pp and slight deterioration of accuracy for confusion and shortness of breath (1 pp) (All metrics are displayed in Supplementary Fig. 3). These results indicate a more effective prompt structure when integrated into the system prompt (Supplementary Methods). Finally, the most effective prompt structure for zero-shot prompting, as concluded from our experiments, was

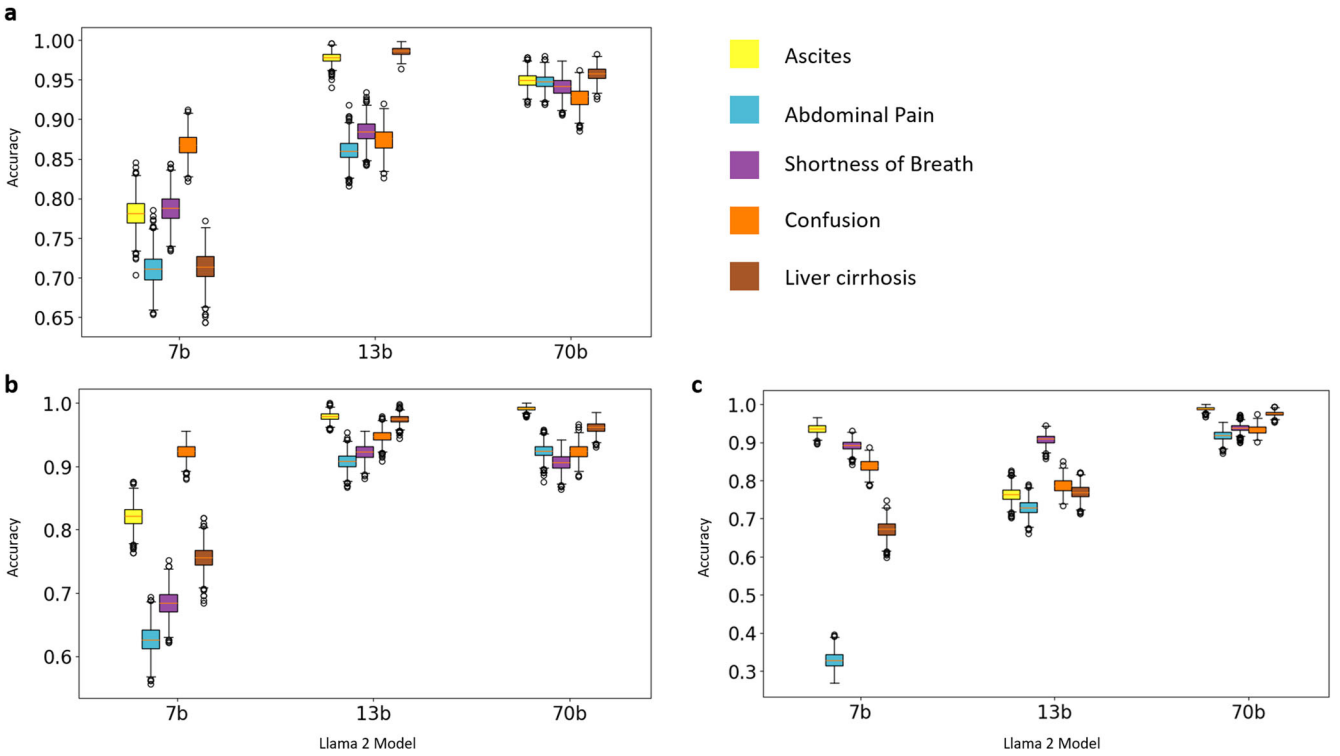

Fig. 4 | Accuracy for prediction of present features with different parameter size models. This graph compares the accuracy of different models (7b, 13b, and 70b) in extracting the five features Ascites, Abdominal pain, Shortness of breath, Confusion, Liver cirrhosis. **a** depicts the accuracy of the final zero-shot prompting, **b** with plain zero shot prompting without additional definition or example, **c** the accuracy of the best one-shot prompting example. Error bars represent the variability or confidence intervals, calculated with 1000-fold bootstrapping.

Table 2 | Model performance—one-shot prompting

	Sensitivity			Specificity			Positive predictive value			Negative predictive value			Accuracy		
	7b	13b	70b	7b	13b	70b	7b	13b	70b	7b	13b	70b	7b	13b	70b
Ascites	0.95	1.00	0.95	0.94	0.76	0.99	0.38	0.13	0.79	1.00	1.00	1.00	0.94	0.76	0.99
Abdominal pain	0.99	0.95	0.92	0.18	0.68	0.92	0.22	0.40	0.72	0.99	0.98	0.98	0.33	0.73	0.92
Shortness of breath	0.64	0.59	0.83	0.95	0.98	0.96	0.72	0.87	0.82	0.92	0.91	0.96	0.89	0.91	0.94
Confusion	0.71	0.85	0.88	0.85	0.78	0.94	0.30	0.25	0.56	0.97	0.98	0.99	0.84	0.79	0.93
Liver cirrhosis	1.00	1.00	1.00	0.65	0.76	0.97	0.14	0.18	0.69	1.00	1.00	1.00	0.67	0.77	0.98

Comparing three versions of Llama 2, the largest (70b) model showed the highest performance whereas the smallest (7b) model performed worst. The 13b and 70b models show higher accuracy across all conditions when compared to the 7b model.

to include all components within the system prompt. This encompassed providing a report, asking specific questions, giving definitions for implicit features, and enforcing a chain-of-thought response through grammatical structuring without a chain-of-thought questioning strategy. Nevertheless, the prompt experiments changed each feature differently. In principle, the least differences between the prompting techniques can be seen in the largest, 70b model. In summary, these data show that prompt engineering can help improve performance especially in the smallest model, whereas larger model sizes demonstrated greater robustness, with remarkably high performance of simple prompts, improving only marginally through prompt engineering.

Discussion

In this study, we present an open-source software pipeline which can use local LLMs to extract quantitative data from clinical free text and evaluate it on the detection of symptoms indicating decompensated liver cirrhosis, an important medical emergency. We demonstrate that the LLM "Llama 2" yields an excellent performance on this task, even in a zero-shot way without any task-specific fine-tuning. Specifically, the 70 billion parameter model was able to achieve 90% accuracy or more for both implicitly and explicitly mentioned features. Historically, rule-based or dictionary-based methods were used for information extraction[30], but these approaches struggle with the variability of medical texts and the scarcity of labeled training data[31]. Additionally, such rule-based hand-crafted methods cannot extract implicitly stated information in a zero-shot way. Therefore, we show that LLMs can fill the gap in information extraction and will be of utmost importance for versatile healthcare data processing.

The performance of LLMs is increasing massively[32] and we expect that future LLMs will further improve the performance. Many proof-of-concept studies for LLMs in medicine only show a semiquantitative analysis—in contrast, we employ a rigorous, quantitative, pre-specified analysis comparing the models' outputs to a ground truth obtained by three blinded observers. We posit that such a systematic analysis should be the gold standard in assessing the benefits and shortcomings of LLMs in medicine.

Not surprisingly, we find that clinical features that are explicitly mentioned in clinical texts are recalled more effectively by our model than those that are implied, indicating a limited grasp of contextual subtleties. The model particularly struggled with extracting "confusion" due to

inconsistent documentation and definition, which even required medical experts to consent about a definition (see Supplementary Tables 1 and 2 in the Supplementary Information for raters' agreement and feature consensus definition). Despite this, the Llama 2 70b model excels in identifying implicitly mentioned features, showing a superior understanding of context linked to its larger parameter size. Our prompt experiments' findings indicate that models with larger parameter size demonstrate enhanced robustness, and their performance remains largely unaffected by variations in prompt engineering, suggesting promising prospects for the development of even better and larger models in the future. Llama has been previously successful in tasks like DRG prediction and tested for ICD code extraction from clinical notes[33,34]. Our analysis reaffirms Llama 2's strong information extraction capabilities and secure processing of sensitive patient data. Nevertheless, Llama as a decoder-only model has proven to struggle more with unseen information types than encoder-decoder models[35], although decoder-only models with more extensive pre-training overcome this limitation. Continuous improvements to Llama and other LLMs, as seen with ChatGPT, could further boost their performance in complex tasks[36]. Several related studies have shown that the LLM GPT-4 excels at structured information extraction from medical text and is often superior to Llama 2. However, GPT-4 runs in the cloud and its architecture is unknown to the public[37], making it currently not suitable for processing personal healthcare data.

LLMs have some fundamental limitations that users must be aware of. In our analysis, we encountered some of these: For instance, our analysis revealed that when Llama 2 was asked to determine a patient's gender from medical history, it based its decision on the prevalence of certain symptoms in one gender over another, rather than using clear identifiers like personal pronouns, which prove the gender instead of suggesting it by probabilities (Supplementary Fig. 1, Supplementary Information). Addressing biases in LLMs is essential to ensure the accuracy and impartiality of the information they deliver. Continuous investigation and the development of advanced methods to assess these models' functioning are vital. This will enable us to rely on these models for information that reflects the actual content, rather than assumptions made by the model. Furthermore, we analyzed Llama's proficiency in evaluating English-language patient histories; its ability to handle data in other languages needs to be further elucidated, since 90% of Llama-2's training data was English language data[26].

Our analysis has the potential to form a basis for clinical decision support systems, aiding in identifying symptoms of conditions like decompensated liver cirrhosis and applicable in various medical fields. Further refinement and evaluation, potentially through fine-tuning, retrieval augmented generation approaches[38] and improved LLMs are necessary to obtain the necessary security in handling medical data, especially to overcome the tendency of LLMs to hallucinate[39], which has also been shown in examples of our experiments. Nevertheless, our research reveals substantial chances for broader medical settings: Enhanced information extraction from free text enables more effective quantitative analysis in research. Moreover, it can streamline quality control in hospital procedures and simplify billing encoding, thereby reducing labor-intensive information extraction tasks.

Methods

Ethics statement

We solely utilized anonymized patient data from the MIMIC IV database. The MIMIC IV dataset is a comprehensive and publically available collection of anonymized medical data from patients admitted to the emergency department or intensive care unit at Beth Israel Deaconess Medical Center in Boston Massachusetts, United States and enables text based research in healthcare and serves as a benchmark for medical AI studies[40]. The MIMIC IV database contains a broad spectrum of patient data collected from 2009 to 2019, thereby being representative of multiple clinical scenarios[41]. All research procedures were conducted in accordance with the Declaration of Helsinki.

Data preparation

We applied for access to the MIMIC-IV database available from physionet.org and obtained access to the comprehensive health-related data of patients treated in an emergency department or intensive care setting[40,42,43]. Central to our study was the early detection of decompensated liver cirrhosis in admission records, a critical task due to the condition's potential lethality and rapid progression to complications such as variceal bleeding, hepatic encephalopathy, or renal failure. Early and accurate identification is vital for initiating immediate treatment and guiding patient management. For this study, we selected the first 500 patient histories (0.15% of all MIMIC IV clinical notes), focusing on identifying signs of decompensation in liver cirrhosis. We utilized Llama 2 to extract three symptoms—shortness of breath, abdominal pain, and confusion—from the text, and to identify two explicitly stated conditions: liver cirrhosis and ascites. This approach aimed to demonstrate the model's effectiveness in discerning both implicit and explicit medical information crucial for patient care.

Model details and data processing

The study's goal was to assess the capability of the LLM "Llama 2", in extracting the mentioned information from the textual medical data. We employed the zero-shot method to run the model. In our approach, all three versions of Llama 2 were used, the 7 billion-, 13 billion-, and 70 billion parameter-sized model. Our aim was to retrieve information about the five predefined features from patients' present medical histories[40]. Initially, the model was prompted to give JavaScript Object Notation (JSON) formatted output, but the model's JSON output was inconsistent and defective. The model output missed relevant parenthesis displaying non-escaped characters that could not be parsed. Therefore, we utilized the llama.cpp version[44], a framework originally designed to run Llama 2 models on lower-resource hardware as well as support grammar-based output formatting. Thus, we enforced the JSON format generation using llama.cpp's grammar-based sampling, which dictates text generation through specific grammatical rules to ensure valid JSON. We then converted these JSON outputs into CSV format using Python's pandas library. The whole pipeline is depicted in Fig. 5.

Prompt engineering

We implemented a technique known as zero-shot chain-of-thought prompting, wherein the model is tasked with identifying relevant text passages without prior training specific to the task, which tests the model's ability to apply its pre-trained knowledge to new problems. By employing a specific grammar-based sampling approach, we enhanced the explainability of the model. Thus, it structured the output as follows: First, an explanation with excerpts from the original report were given, then the binary response indicating the presence or absence of a feature was determined (example output: \"abdominal pain\": {\"excerpt\": \"Patient reported mild right upper quadrant pain.\", \"present\": true}. This also implemented a "chain-of-thought" process, which allowed sequential reasoning where the LLM output transparently outlines its thought process, to verify the existence of a particular feature within the text. To enhance outcomes via prompt engineering, one-shot prompting was also employed[45], providing the model with an example report and corresponding JSON formatted output. Blinded medical raters established a consensus on precise definitions for the queried features during ground truth definition, which were subsequently provided to the model (definition prompting). Ultimately, single-shot and definition chain-of-thought prompting were combined. The standard Llama 2 prompt contains two modules, the "system" and the "user" part. The system prompt provides initial instructions or explanations to guide the interaction, while the user prompt includes the user's input or query, further shaping the response process. We experimented with different arrangements of system and user prompts in combination with definition, one-shot and chain-of-thought prompting and prompt modules containing general instructions, original report and questions. The MAIN ZERO SHOT PROMPT

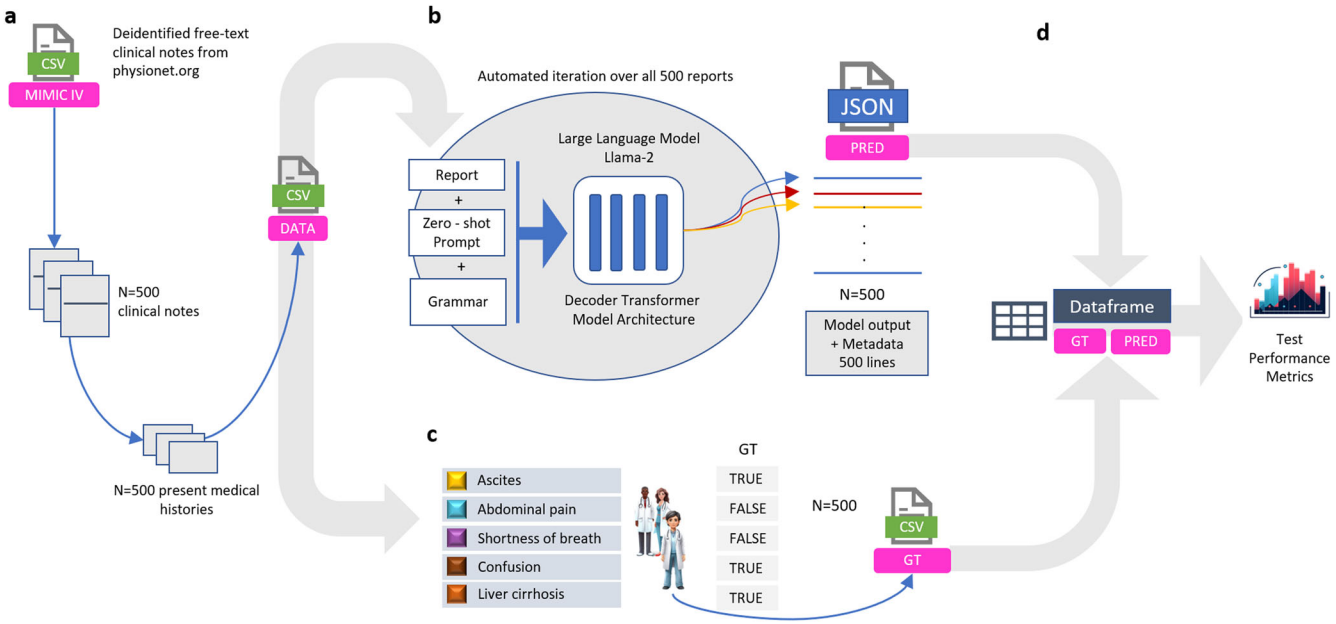

Fig. 5 | Experimental design and feature extraction pipeline. a We implemented an automated process to extract 500 free-text clinical notes from the MIMIC IV database, focusing specifically on the patients' present medical histories. These selected anamnesis reports were then systematically converted and stored in a CSV file for further processing. **b** Utilizing this CSV file, our custom-designed software algorithm selected one report at a time and combined it with a predetermined prompt and grammatical structures. This combination was then input into the advanced large language model, Llama 2. The primary function of Llama 2 in our study was to meticulously identify and extract specific, predefined clinical features (namely, Shortness of Breath, Abdominal Pain, Confusion, Ascites, and Liver Cirrhosis) from the clinical reports. The extracted data were subsequently formatted into a JavaScript Object Notation (JSON) file. To ensure a high degree of precision and structured output, we applied a grammar-based sampling technique. **c** To establish a benchmark, we engaged three medical experts who independently analyzed the same clinical reports. They extracted identical items as the Llama 2 model, thereby creating a reliable "ground truth" dataset. **d** This ground truth dataset served as a reference point for a quantitative comparison and analysis of the model's performance, assessing the accuracy and reliability of the information extracted by Llama 2. Icons are generated by the author with the AI generation tool Midjourney[46].

(Supplementary information) shows the final zero shot prompt, underlying the results in Figs. 2 and 4a.

Definition of the ground truth

For validation, the 500 reports were independently assessed by three human observers to establish a ground truth. In the event of disagreement, a consensus was always reached through discussion (Supplementary Tables 1 and 2). A comprehensive overview regarding consensus about the ground truth rating, as well as challenges and methodologies concerning ground truth definition, can be found in the Supplementary Information.

Evaluation of model results

Positive Predictive Value (Precision, PPV), Sensitivity (Recall), Specificity, Negative Predictive Value (NPV) and Accuracy were computed to assess the performance of the different model's outputs. To obtain reliable estimates, we employed bootstrapping, a statistical resampling technique, executing 1000 iterations. This method involves repeatedly sampling from the dataset with replacement to create many "bootstrap" samples. These samples are then used to estimate the variability and confidence of our statistical estimates, enhancing their robustness and credibility.

Data availability

All data used in the study were obtained from the MIMIC-IV database available from physionet.org and can be accessed as credentialed user, who has completed required training and signed the data use agreement for the project[40,42,43].

Code availability

All source codes are available at https://github.com/I2C9W/fromtexttotables/releases/tag/v0.5.0. All scripts are compatible with Python 3.8. All Python packages required are listed in the requirements.txt file. Execution advice guidance can be found in the README file.

Received: 9 January 2024; Accepted: 19 August 2024;
Published online: 20 September 2024

References

1. Kong, H.-J. Managing unstructured big data in healthcare system. *Healthc. Inform. Res.* **25**, 1–2 (2019).
2. Tomašev, N. et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.* **16**, 2765–2787 (2021).
3. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
4. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
5. Chiu, C.-C. et al. Integrating structured and unstructured EHR data for predicting mortality by machine learning and latent Dirichlet allocation method. *Int. J. Environ. Res. Public Health* **20**, 4340 (2023).
6. Price, S. J., Stapley, S. A., Shephard, E., Barraclough, K. & Hamilton, W. T. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case–control study. *BMJ Open* **6**, e011664 (2016).
7. Pivovarov, R., Coppleson, Y. J., Gorman, S. L., Vawdrey, D. K. & Elhadad, N. Can patient record summarization support quality metric abstraction? *AMIA Annu. Symp. Proc.* **2016**, 1020–1029 (2016).

8. Locke, S. et al. Natural language processing in medicine: a review. *Trends Anaesth. Crit. Care* **38**, 4–9 (2021).
9. Chary, M., Parikh, S., Manini, A. F., Boyer, E. W. & Radeos, M. A review of natural language processing in medical education. *West. J. Emerg. Med.* **20**, 78–86 (2019).
10. Castelo-Branco, L. et al. ESMO guidance for reporting oncology real-world evidence (GROW). *Ann. Oncol.* https://doi.org/10.1016/j.annonc.2023.10.001 (2023).
11. Chapman, W. W. et al. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J. Am. Med. Inform. Assoc.* **18**, 540–543 (2011).
12. Wang, Y. et al. Clinical information extraction applications: a literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
13. Paaß, G. & Giesselbach, S. *Foundation Models for Natural Language Processing: Pre-Trained Language Models Integrating Media* (Springer Nature, 2023).
14. Yang, X., Bian, J., Hogan, W. R. & Wu, Y. Clinical concept extraction using transformers. *J. Am. Med. Inform. Assoc.* **27**, 1935–1942 (2020).
15. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst*. **30**, 1–11 (2017).
16. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
17. Bommasani, R. et al. On the opportunities and risks of foundation models. *arXiv* https://doi.org/10.48550/arXiv.2108.07258 (2021).
18. Adams, L. C. et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).
19. Kleesiek, J. An Opinion on ChatGPT in Health Care-Written by Humans Only. *J. Nucl. Med.* **64**, 701–703 (2023).
20. Li, J., Dada, A., Kleesiek, J. & Egger, J. ChatGPT in healthcare: a taxonomy and systematic review. *bioRxiv* https://doi.org/10.1101/2023.03.30.23287899 (2023).
21. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat. Med*. https://doi.org/10.1038/s41591-023-02594-z (2023).
22. Simon Jones, N. J. et al. Evaluating ChatGPT in information extraction: a case study of extracting cognitive exam dates and scores. *medRxiv* https://doi.org/10.1101/2023.07.10.23292373 (2023).
23. Minssen, T., Vayena, E. & Cohen, I. G. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA* **330**, 315–316 (2023).
24. Weatherbed, J. OpenAI's regulatory troubles are only just beginning. *The Verge. Artificial Intelligence.* https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy (2023)
25. Raeini, M. Privacy-preserving large language models (PPLLMs). https://doi.org/10.2139/ssrn.4512071 (2023).
26. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. *arXiv* https://doi.org/10.48550/arXiv.2307.09288 (2023).
27. Huang, D. Q. et al. Global epidemiology of cirrhosis—aetiology, trends and predictions. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 388–398 (2023).
28. Volk, M. L., Tocco, R. S., Bazick, J., Rakoski, M. O. & Lok, A. S. Hospital readmissions among patients with decompensated cirrhosis. *Am. J. Gastroenterol.* **107**, 247–252 (2012).
29. Balcar, L. et al. Risk of further decompensation/mortality in patients with cirrhosis and ascites as the first single decompensation event. *JHEP Rep.* **4**, 100513 (2022).
30. Landolsi, M. Y., Hlaoua, L. & Ben Romdhane, L. Information extraction from electronic medical documents: state of the art and future research directions. *Knowl. Inf. Syst.* **65**, 463–516 (2023).
31. He, K. et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv* https://doi.org/10.48550/arXiv.2310.05694 (2023).
32. Open LLM Leaderboard. *Huggingface* https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (2023).
33. Wang, H. et al. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med*. **7**, 16 (2024).
34. Boyle, J. S. et al. Automated clinical coding using off-the-shelf large language models. *arXiv* https://doi.org/10.48550/arXiv.2310.06552 (2023).
35. Gao, J. et al. Benchmarking large language models with augmented instructions for fine-grained information extraction. *arXiv* https://doi.org/10.48550/arXiv.2310.05092 (2023).
36. OpenAI. GPT-4 technical report. *arXiv* https://doi.org/10.48550/arXiv.2303.08774(2023).
37. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Med.* **6**, 120 (2023).
38. Ferber, D. & Kather, J. N. Large language models in uro-oncology. *Eur. Urol. Oncol.* https://doi.org/10.1016/j.euo.2023.09.019 (2023).
39. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is inevitable: an innate limitation of large language models. *arXiv* https://doi.org/10.48550/arXiv.2401.11817 (2024).
40. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
41. Mark, R. The story of MIMIC. 2016 Sep 10. In *Secondary Analysis of Electronic Health Records* (ed. MIT Critical Data) (Springer Nature, 2016).
42. Johnson, A., Bulgarelli, L., Pollard, T. & Horng, S. MIMIC-IV—*PhysioNet* (2020).
43. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
44. Gerganov, G. llama.cpp. *GitHub* (2023).
45. White, J. et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv* https://doi.org/10.48550/arXiv.2302.11382 (2023).
46. Midjourney. Midjourney (V5) [Text-to-image model]. (2023).

Acknowledgements

J.N.K. is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

I.C.W. conceptualized the study and developed the methodology in close coordination with J.N.K., D.F., M.T. and J.Z. I.C.W. developed the scripts and I.C.W. and S.M. ran the experiments. I.C.W., D.F. and S.M. were determining the ground truth and evaluating the model results. I.C.W., D.F. and J.N.K. were writing the initial manuscript, reviewed by D.T., Z.I.C., R.J., S.M., J.K., D.P. and M.P.E. All authors contributed scientific advice and approved the final version of the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
J.N.K. declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Scailyte, Switzerland; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore he holds shares in StratifAI GmbH, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. D.T. has received honoraria for lectures for Bayer and holds shares in StratifAI GmbH, Dresden, Germany. I.C.W. received honoraria from AstraZeneca. The authors have no other financial or non-financial conflicts of interest to disclose. D.F., J.Z., M.T., S.M., R.J., Z.I.C., D.P., J.K. and M.P.E. have no competing interests to declare.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01233-2.

Correspondence and requests for materials should be addressed to Jakob Nikolas Kather.