Original Paper

Large Language Model Performance in Structured Data Extraction From Neurosurgical Electronic Health Records: Comparative Benchmark Study

Martin Černý[1*], MD, PhD; Martin Májovský[1*], MD, PhD; Sanju Lama[2], MD, PhD; Edward Guo[2,3], MD; Hana Hallak[4], MBBS; Rahul Singh[2], BSc; Homer Riva-Cambrin[2], BSc; Petr Šustek[5,6], LLD, PhD; Lucie Široká[6], LLM; Katarina Horčičáková[7], MD; Samuel Hužava[8], MD; Kateřina Sajfrídová[9], MD; Klára El-Haj Ahmadová[9], MD; Garnette Roy Sutherland[2], MD; David Netuka[1], MD, PhD

[1]Department of Neurosurgery and Neurooncology, 1st Faculty of Medicine, Charles University, Military University Hospital Prague, Prague, Czech Republic

[2]Department of Clinical Neurosciences, Project neuroArm, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

[3]Department of Neurosurgery, Neurosurgery Residency Program, University of Toronto, Toronto, ON, Canada

[4]Department of Neurosurgery, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia

[5]Department of Medical Law, Faculty of Law, Charles University, Prague, Czech Republic

[6]Department of Civil Law, Faculty of Law, Charles University, Prague, Czech Republic

[7]Jessenius Faculty of Medicine, Comenius University Bratislava, Martin, Bratislava Region, Slovakia

[8]2nd Faculty of Medicine, Charles University, Prague, Czech Republic

[9]1st Faculty of Medicine, Charles University, Prague, Czech Republic

*these authors contributed equally

Corresponding Author:

David Netuka, MD, PhD
Department of Neurosurgery and Neurooncology
1st Faculty of Medicine, Charles University, Military University Hospital Prague
U Vojenské nemocnice 1200
Prague 16902
Czech Republic
Phone: +420 973 202 963
Email: david.netuka@uvn.cz

Abstract

Background: Large language models (LLMs) are increasingly used to extract information from electronic health records (EHRs). Given the rapid pace of LLM development, robust scenario-specific benchmarks are essential to evaluate clinical usefulness and support safe deployment.

Objective: This study aimed to compare contemporary LLMs for structured data extraction from real neurosurgical EHRs written in Czech.

Methods: In a prospective single-center cohort, 172 hospitalized patients provided informed consent for the use of anonymized EHRs. For each patient, predefined records were collected and concatenated. Ground truth for 35 data points was established by dual extraction with consensus. A standardized prompt requesting JSON output was submitted to 19 LLMs. The primary outcome was overall accuracy; secondary outcomes were category-level accuracy and the proportion of complete machine-readable outputs.

Results: A total of 6264 documents were collected (median 33, IQR 25-41 per patient). Ground truth was established with 92.6% initial interrater agreement before consensus. Several models produced complete JSON outputs for 100% of cases (Claude 4.1 Opus, Grok 4, and Gemini 2.5 Flash); GPT-4.1 (DeepSearch) and GPT-5 completed 99.4%. The highest accuracy was achieved by GPT-4.1 (87.6%), followed by GPT-4.5 (85.6%), Claude 4.1 (84.8%), and Grok 4 (84.2%). Accuracy declined by data type: binary (up to 95%), numeric (~89%), multiple-choice (~75%), and short text (~78%).

Conclusions: Current LLMs can reliably extract structured clinical information from full, non-English EHRs, while older or smaller models show major limitations. A hybrid workflow—automated extraction with targeted validation—appears practical for research use.

JMIR Form Res 2026;10:e92871; doi: 10.2196/92871

Keywords: large language models; electronic health records; neurosurgery; data extraction; patient records; artificial intelligence; AI; data protection; prompt engineering

Introduction

Large language models (LLMs) have rapidly become a general-purpose technology for natural language processing (NLP) and text generation across a wide range of domains. In medicine, they are increasingly explored as tools for information retrieval, summarization, documentation support, and decision assistance. Multiple standardized tests have been proposed to measure LLM performance across many different domains, such as the Massive Multitask Language Understanding benchmark and its later versions [1-3]. Both the potential and the risks of using LLMs in clinical settings have been described, drawing attention to issues such as hallucinated outputs, bias, data security, and the need for consistent and transparent evaluation standards [4-8].

Within health care, LLMs, which are particularly well suited for processing unstructured clinical text, are now being applied to diverse tasks including querying medical questions, evidence summarization, biomedical literature mining, and patient communication. Systems such as Med-PaLM and related models have achieved or approached expert-level performance on medical exam–style benchmarks, while domain-specific evaluations have compared closed- and open-source models across multiple biomedical NLP tasks [9-11]. However, work in this area is characterized by substantial heterogeneity in task definitions, evaluation metrics, and reporting standards [6,12].

Benchmarking efforts tailored to medicine have started to move beyond exam questions toward more realistic day-to-day clinical scenarios. Multilingual benchmarks such as BRIDGE and biomedical NLP suites assess LLMs across question answering, information extraction, classification, and summarization using real-world clinical and biomedical sources [10,11,13]. Several recent studies have specifically evaluated LLMs for structured data extraction from electronic health records (EHRs), generally reporting encouraging performance but also notable variability between models, tasks, and clinical settings [14-16]. These studies consistently underscore that model choice, prompt design, and domain shift can substantially impact extraction accuracy.

However, the LLM landscape is evolving extremely quickly, with new proprietary and open-weight models released at short intervals and substantial improvements in reasoning, context length, and instruction following [17-34]. To expand the current evidence base, we designed a benchmark study specifically targeting the ability of contemporary LLMs (all prompts were submitted on December 8 and 9, 2025) to extract predefined structured clinical data points from real-world neurosurgical non-English (Czech) EHRs and to compare their performance in terms of both machine-readable completeness and extraction accuracy. The aim of this study was to provide practical actionable recommendations based on the performance of the compared models on various tasks.

Methods

Study Design

This was a prospective, single-center, cross-sectional benchmark study comparing 19 contemporary LLMs for structured data extraction from real-world neurosurgical EHRs. Models were evaluated against a consensus-derived human ground truth for completeness and accuracy of extraction for 35 predefined clinical data points. This study was reported in accordance with the TRIPOD-LLM (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model) reporting guideline [35].

Patients

Patients hospitalized at the Department of Neurosurgery and Neurooncology of the Military University Hospital Prague during a 1-month enrollment period (November 2023) were screened for inclusion in the study. Eligible participants were adults (aged ≥18 years) admitted during the enrollment window who had the capacity to provide informed consent and signed written informed consent to participate in the study. Patients who declined participation, were unable to provide written informed consent, or were not discharged during the enrollment period or within 7 days thereafter were excluded. Individuals with multiple qualifying hospitalizations during the enrollment period were included as separate participants.

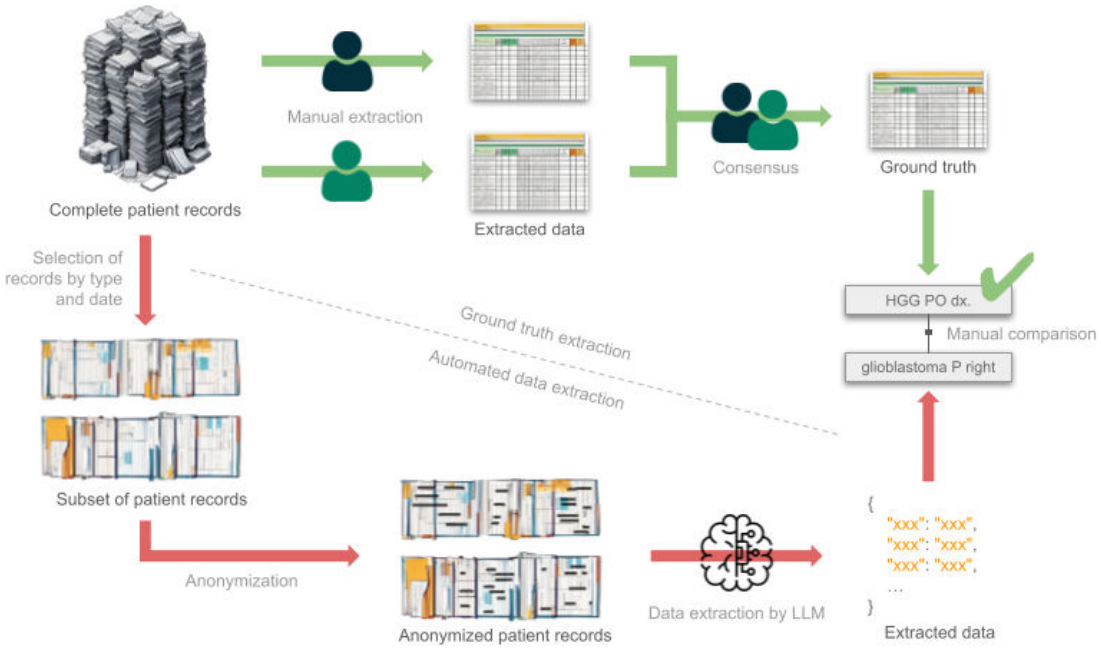
Ground Truth Extraction

A total of 35 predefined data points were extracted from the EHR at the time of discharge (Table 1). Data points were categorized as binary (yes or no; n=13), free-text (n=7), numeric (n=8), or multiple-choice variables (n=7). Two authors (KH and SH) independently reviewed the complete patient records and recorded each data point. Ground truth was established by comparing both extractions. In cases of disagreement or ambiguity, 2 senior authors (MČ and MM) jointly rereviewed the source documentation and resolved discrepancies by consensus. If unable to establish a consensus or if the information was not present, the data point was omitted for the respective case. Figure 1 presents a schematic overview of the experimental design.

Table 1. Predefined clinical data points extracted from patient records.

Category	Data points (type)
Hospitalization	Admission diagnosis (text), admission diagnosis *International Classification of Diseases* code (text), hospitalization purpose (therapeutic or diagnostic), surgery (text), further patient course (discharge or transfer), number of days in ICU (number), and revision surgery (text, if the patient underwent a revision)
Neurological state on admission	Glasgow Coma Scale (number), speech impairment (yes or no), motor impairment (no, upper extremity monoparesis, lower extremity monoparesis, hemiparesis, paraparesis, quadriparesis, or cranial nerve paresis), sensory impairment (yes or no), upper motor neuron signs (yes or no), pain intensity (number and visual analog scale), pain localization (text), and visual impairment (yes or no)
Neurological state at the first postoperative day	Same as on admission
Neurological state at discharge	Same as on admission
Therapy	Corticotherapy (yes or no), algotherapy (no, peroral, intravenous, or parenteral opioids), antibiotics (no, peroral, or intravenous, excluding perioperative prophylaxis), and antibiotics (number of days, excluding perioperative prophylaxis)

Figure 1. Schematic overview of the experimental design, illustrating the parallel processes of human ground truth extraction (green) and automated model–based extraction (red), converging on the final accuracy comparison. LLM: large language model.



List of the 35 structured clinical data points used for model evaluation, categorized by hospitalization course, neurological status, and therapeutic interventions. Variables are grouped by data type.

Dataset Collection and Anonymization

For each participant, a predefined set of EHR document types was collected (Table 2). All documents were exported automatically from the institutional hospital information system (AMIS*H version 2.0; ICZ sro).

Table 2. Types of electronic health record (EHR) documents collected for each participant.

EHR types	Description
EHR types collected for the hospitalization period	
Clinical summaries	A brief summary of the patient, usually 1 or 2 paragraphs, includes relevant patient history, presenting symptoms, performed procedures, and imaging exams; at least one at admission, with possibly multiple such summaries during the hospitalization
History of present illness	Usually 1 or 2 paragraphs, including the nature and onset of symptoms, their duration, severity, and progression, as well as factors that exacerbate or alleviate them
Neurological assessment on admission	An examination report detailing consciousness, orientation, speech, cranial nerve functions, motor functions, sensory system functions, reflexes, gait, cerebellar functions, and meningeal signs
Consultation notes	Reports from interdisciplinary consultations requested during the hospitalization
Progress notes	Daily notes reporting any changes in patient signs and symptoms and performed procedures
Prescription orders	Daily drug prescriptions

EHR types	Description
Therapeutic orders	Prescriptions for nonpharmacological therapy, including physiotherapy, ergotherapy, wound management, catheters, and dietary measures
Nursing assessment	Includes vital signs, physical and mental status assessment, fluid intakes and outputs, pain assessment, patient education, and communication and safety measures
EHR types collected for the 2 years before hospitalization	
Neurosurgery and neurology outpatient consult notes	Include the patient's medical history, neurological findings, diagnosis, treatment plan, and follow-up instructions relevant to their neurological condition
Emergency department visit summaries	Typically include the reason for the visit, initial assessment, interventions provided, results of diagnostic tests, final diagnosis, and any follow-up plans or referrals
Radiology reports	Detailed findings from imaging studies, including information about the technique used, area imaged, findings, and any potential abnormalities or diagnostic clues relevant to the patient's condition

Anonymization was performed manually by the authors (KS and KA). All direct identifiers (including patient names, personal identification numbers, and names of health care professionals or institutions) were removed. Temporal information was transformed into relative references (eg, "September 27th" to "second postoperative day"). Names and serial numbers of medical devices were removed from imaging reports. No information that could reasonably facilitate reidentification was retained. The order of documents in the resulting dataset was randomized.

The English terms were chosen to match terminology commonly used in Epic (Epic Systems Corporation). The right column provides a brief explanation of each EHR type.

LLM Prompting

For each patient, all collected and deidentified clinical records were concatenated into a single text document. A standardized prompt was appended to the end of this document, instructing the model to extract 35 predefined clinical data points and to return the results in machine-readable JSON format, using the exact field names specified in the study protocol.

Each concatenated input was submitted to all evaluated language models under identical prompting conditions. The following models were included in the experiment: OpenAI GPT-4 [17], OpenAI GPT-4 o [18], OpenAI GPT-3.5 [19], OpenAI GPT-4.1 (DeepSearch mode) [20], OpenAI GPT-4.5 [21], OpenAI GPT-5 [22], Google Bard (Legacy) [23], Google Gemini 2.5 Flash [24], Anthropic Claude 2 [25], Anthropic Claude 4 Sonnet [26], Anthropic Claude 4.1 Opus [27], xAI Grok 3 [28], xAI Grok 4 [29], DeepSeek V2 [30], neuroGPT-X [31], Cohere Command R+ [32], Meta LLaMA-2 [33], and Meta LLaMA-3 [34]. All prompts were submitted to the evaluated models on December 8 and 9, 2025.

All clinical source texts and prompts were written in Czech, reflecting the language of the institutional EHR and ensuring a consistent linguistic setting across models. The standardized prompt used in this study is provided in full in Multimedia Appendix 1 (original Czech version) and Multimedia Appendix 2 (English translation). The actual patient documentation is represented by a placeholder in the multimedia appendices.

Correctness Evaluation

Responses that did not contain syntactically valid JSON were classified as fully incorrect, and all data points within such responses were considered incorrect. Within valid JSON outputs, missing fields were treated as incorrect extractions.

For binary (yes or no) variables, model outputs were compared directly against the ground truth Boolean value. For numeric variables, exact numeric equivalence was required. For multiple-choice variables, the model output was required to match one of the predefined enumerated values.

For free-text variables (admission diagnosis, *International Classification of Diseases* code, surgical procedure name, revision surgery name, pain localization, and 2 additional text fields), exact string matching was not required. Instead, model outputs were evaluated manually by 2 authors (MČ and MM) against the ground truth, with semantic equivalence considered correct. Responses were accepted as correct if they conveyed the same clinical meaning as the ground truth, regardless of wording, abbreviation style, or level of detail, provided no clinically meaningful information was added or omitted. For example, "HGG PO dx." and "glioblastoma of the right parietal lobe" were both accepted as correct for the same ground truth value. Responses that introduced clinically incorrect information, omitted a key qualifier, or described a meaningfully different entity were marked as incorrect. In cases of disagreement between the two evaluating authors, a consensus decision was reached through discussion.

Outcomes

The primary outcome was the overall extraction accuracy, defined as the percentage of correctly extracted data points out of all evaluable ground truth values. Secondary outcomes were the percentage of correctly extracted data points within each question category (yes or no, text, numeric, and multiple choice), and the proportion of complete, machine-readable JSON replies (ie, syntactically valid JSON containing all 35 data points).

Statistical Analysis

Descriptive statistics were reported as means with 95% CI for continuous outcomes and as absolute counts with percentages for categorical outcomes. Interrater agreement for ground truth establishment was assessed using Cohen κ. For JSON completeness, pairwise comparisons between

the best-performing model and each remaining model were conducted using the McNemar test on paired binary outcomes (df=1 for all tests). For extraction accuracy, per-patient scores were computed by aggregating binary correct or incorrect outcomes across all applicable data points, yielding one continuous score per patient per model. Pairwise comparisons of per-patient scores against the best-performing model were conducted using paired-samples 2-tailed *t* tests (df=171 for all comparisons). Normality of pairwise differences was assessed using the Shapiro-Wilk test; where the assumption of normality was violated (Shapiro-Wilk, $P<.05$), Wilcoxon signed-rank tests were used instead. Effect sizes are reported as Cohen *d* for *t* tests and rank-biserial correlation for Wilcoxon tests (nonparametric test, df not applicable; n=172). To control for multiple comparisons, Bonferroni correction was applied within each outcome family (18 comparisons per outcome). A corrected *P* value below .05 was considered statistically significant. The detailed statistical outcomes are available in Multimedia Appendix 3.

Ethical Considerations

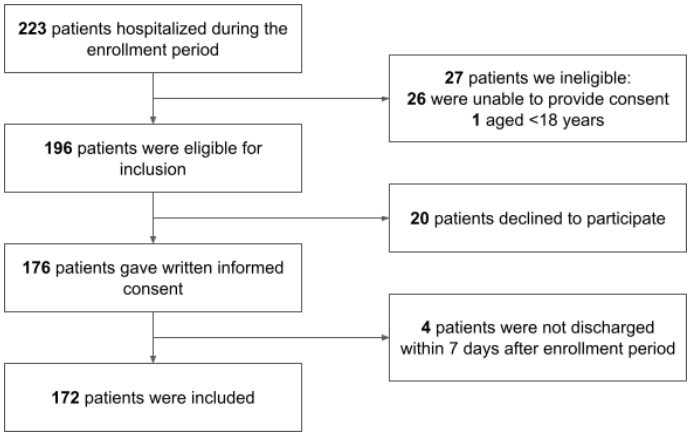
This study was approved by the Institutional Review Board of the Military University Hospital Prague (108/18-101/2023). Written informed consent was obtained from all participants, explicitly covering the use of anonymized EHRs by third-party systems and the publication of aggregated results. All direct patient identifiers, including names, personal identification numbers, and names of health care professionals and institutions, were removed prior to processing; temporal references were replaced with relative descriptors. The study was conducted in accordance with the ethical standards of the Declaration of Helsinki. No participant compensation was provided.

Results

Patients and Data

During the enrollment period, 223 patients were hospitalized at our institution. Of 196 (87.9%, 196/223) eligible and invited patients, 176 (89.8%, 176/196) agreed to participate. A total of 172 (97.7%, 172/176) patients were discharged during the enrollment period or within 1 week thereafter and were included in the study sample. The flow of patient inclusion is illustrated in Figure 2. The cohort consisted of 81 (47.1%, 81/172) male patients and 91 (52.9%, 91/172) female patients, with a mean age of 59.1 (SD 15.3) years at admission. The median length of hospitalization was 6 (IQR 5-7) days.

Figure 2. Flowchart of patient inclusion, from initial screening through enrollment in the study sample.



Most patients were treated for spinal conditions (n=88), followed by brain tumors (n=40), vascular disorders (n=24), functional disorders (n=9), surgical complications (n=9), and peripheral nerve conditions (n=2).

In total, 6264 individual patient record items were collected (Table 3), corresponding to a median of 33 items per patient (IQR 25-41). The mean length of the concatenated preoperative documentation was 906.0 (SD 158.7) words. Ground truth could be established for 91.9% of data points. The initial interrater agreement prior to consensus was 92.6% (κ=0.46).

Table 3. Total number and median per-patient counts of collected patient record types.

Patient record type	Per patient, median (IQR)
Clinical summaries (n=224)	1 (1-1)
History of present illness (n=172)	1 (1-1)
Neurological assessment on admission (n=172)	1 (1-1)
Consultation notes (n=67)	0 (0-1)
Progress notes (n=998)	6 (4-7)
Prescription orders (n=891)	5 (4-6)

Patient record type	Per patient, median (IQR)
Therapeutic orders (n=905)	5 (4-6)
Nursing assessment (n=2170)	10 (6-14)
Outpatient neurosurgery consults (n=314)	1 (1-2)
Outpatient neurology consults (n=54)	0 (0-0)
Emergency department visit summaries (n=33)	0 (0-0)
Radiology reports (n=264)	1 (0-2)
Total (n=6264)	33 (25-41)

The table summarizes the volume of documentation extracted for each record type across all enrolled patients, reported as the total number of documents and the median number per patient with IQR.

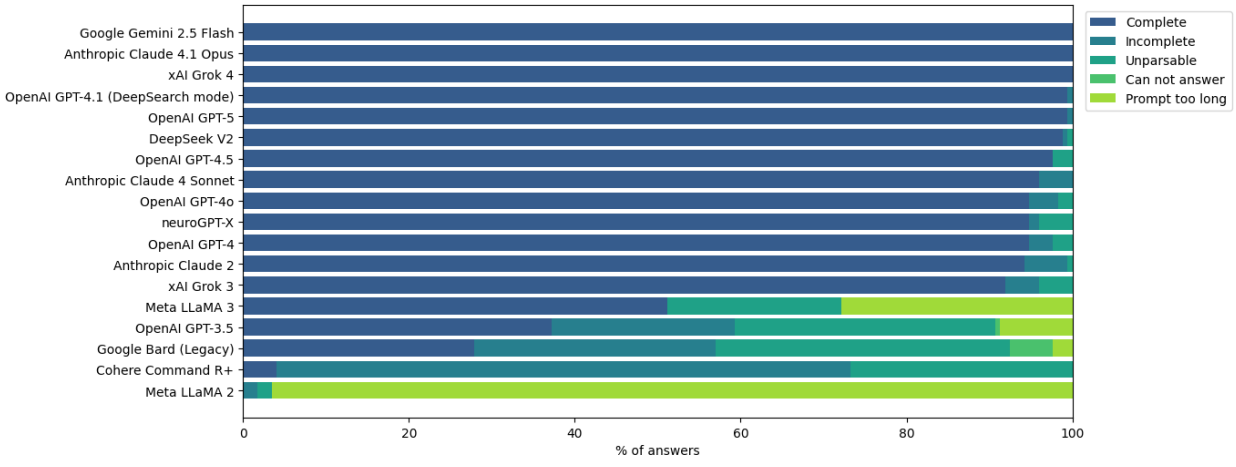
Machine Readability and Completeness of Responses

Claude Opus 4.1, Grok 4, and Gemini 2.5 Flash generated 100% complete JSON outputs with no parsing failures. Performance in the near-state-of-the-art tier remained stable, with models such as GPT-4.1 (DeepSearch mode), GPT-5, GPT-4, and Claude Sonnet 4 demonstrating minimal failure modes. The lowest performance was observed with LLaMA 2, for which almost all outputs were unusable due to excessively long prompts or malformed structure. For early or smaller models, the dominant failure mode was structural unreadability, in which the JSON response was unparsable, whereas for the strongest models, the rare failures were almost exclusively due to partial outputs in which one or more fields were omitted.

Pairwise McNemar tests showed that the reference model was significantly more complete than most older or smaller models. Newer top-tier models—GPT-4.1, GPT-5, GPT-4.5, DeepSeek V2, and Claude 4 Sonnet—performed comparably to the reference, with no meaningful difference between them, while GPT-4, GPT-4o, and neuroGPT-X showed lower completeness rates than the reference that did not reach statistical significance after Bonferroni correction (P=.07).

A visual comparison of the distribution of output types for each model is shown in Figure 3. Detailed machine-readability outcomes, including the classification of each response (complete, incomplete, unparsable, declined, or prompt too long) and pairwise McNemar test results against the reference model, are available in Multimedia Appendix 3, reported separately given the large number of pairwise statistical comparisons performed.

Figure 3. Distribution of machine-readable JSON outputs across models. Bars represent the percentage of outcomes relative to 172 test cases per model. Models are ordered by completeness rate.



Data Extraction Accuracy

Across all evaluated data points, the highest overall accuracy was achieved by OpenAI GPT-4.1 (DeepSearch mode), followed closely by OpenAI GPT-4.5, Anthropic Claude 4.1 Opus, and xAI Grok 4. A group of models, including DeepSeek V2, Anthropic Claude 4 Sonnet, and OpenAI GPT-5, reached comparable performance levels. By contrast, older or smaller-scale models showed substantially lower accuracy, with LLaMA 2 performing close to random output.

Paired 2-tailed t tests with Bonferroni correction confirmed that GPT-4.1 significantly outperformed all models except GPT-4.5, against which no significant difference was detected (t_{171}=–1.747; P=.99). Among the remaining top-tier models, Claude 4.1 Opus (P<.001), Grok 4 (P<.001), Claude 4 Sonnet (P<.001), DeepSeek V2 (P<.001), and GPT-5 (P<.001) all differed significantly from the reference despite their numerically close accuracy values.

When stratified by question type, all models performed best on binary (yes or no) queries. Accuracy decreased for

numeric extraction tasks, with top models achieving 89% to 90%, and declined further for multiple-choice classification, where performance for the best models ranged from 74% to 80%. The lowest performance was observed for open-ended textual output, indicating that free-text generation remained the most error-prone task in automated clinical data extraction.

A summary of the total accuracy and the performance across the 4 question categories for all evaluated models is provided in Table 4. Complete per-patient accuracy results, including pairwise statistical comparisons against the reference model and associated effect sizes, are presented in Multimedia Appendix 4.

Table 4. Overall and category-level accuracy of evaluated language models[a].

Models	Overall accuracy (%), mean (SD; 95% CI)	Yes or no questions (%), mean (SD; 95% CI)	Numeric questions (%), mean (SD; 95% CI)	Multiple-choice questions (%), mean (SD; 95% CI)	Text questions (%), mean (SD; 95% CI)
OpenAI GPT-4.1 (DeepSearch mode)	87.6 (0.330; 86.7-88.5)	95.2 (0.215; 94.3-96.1)	89.5 (0.306; 87.8-91.3)	80.4 (0.397; 78.1-82.6)	77.6 (0.417; 75.0-80.1)
OpenAI GPT-4.5	85.6 (0.352; 0.35; 84.6-86.5)	91.6 (0.278; 90.4-92.7)	87.2 (0.334; 85.3-89.1)	80.5 (0.397; 78.2-82.7)	76.7 (0.423; 74.1-79.3)
Anthropic Claude 4.1 Opus	84.8 (0.359; 83.8-85.7)	93.6 (0.244; 92.6-94.7)	85.0 (0.357; 83.0-87.1)	74.1 (0.438; 71.6-76.6)	77.9 (0.415; 75.3-80.4)
xAI Grok 4	84.2 (0.365; 83.2-85.1)	91.7 (0.277; 90.5-92.8)	88.1 (0.324; 86.2-89.9)	72.8 (0.445; 70.3-75.4)	76.9 (0.422; 74.3-79.5)
DeepSeek V2	83.9 (0.367; 82.9-84.9)	92.1 (0.270; 90.9-93.2)	86.8 (0.339; 84.9-88.7)	71.6 (0.451; 69.0-74.2)	77.4 (0.419; 74.8-79.9)
Anthropic Claude 4 Sonnet	83.9 (0.368; 82.9-84.8)	92.2 (0.269; 91.0-93.3)	86.2 (0.345; 84.2-88.2)	72.2 (0.448; 69.7-74.8)	77.0 (0.421; 74.4-79.6)
OpenAI GPT-5	83.0 (0.376; 82.0-84.0)	93.5 (0.246; 92.5-94.5)	79.6 (0.403; 77.3-81.9)	71.5 (0.451; 69.0-74.1)	77.7 (0.417; 75.1-80.2)
Google Gemini 2.5 Flash	81.4 (0.389; 80.4-82.4)	90.2 (0.298; 88.9-91.4)	81.1 (0.391; 78.9-83.4)	72.9 (0.445; 70.4-75.4)	72.8 (0.445; 70.1-75.6)
xAI Grok 3	78.8 (0.409; 77.7-79.9)	85.0 (0.357; 83.5-86.5)	82.3 (0.382; 80.1-84.5)	68.1 (0.466; 65.4-70.7)	74.0 (0.439; 71.3-76.7)
OpenAI GPT-4o	76.4 (0.425; 75.3-77.5)	90.1 (0.298; 88.9-91.4)	74.9 (0.434; 72.4-77.4)	57.6 (0.494; 54.8-60.4)	70.5 (0.456; 67.7-73.3)
OpenAI GPT-4	76.0 (0.427; 74.9-77.2)	88.6 (0.318; 87.2-89.9)	72.8 (0.445; 70.3-75.4)	67.7 (0.468; 65.0-70.4)	62.6 (0.484; 59.6-65.6)
neuroGPT-X	70.1 (0.458; 68.9-71.3)	89.1 (0.312; 87.8-90.4)	61.4 (0.487; 58.6-64.1)	58.4 (0.493; 55.6-61.2)	53.2 (0.499; 50.2-56.3)
Anthropic Claude 2	69.4 (0.461; 68.2-70.6)	85.5 (0.352; 84.1-87.0)	66.7 (0.472; 64.0-69.4)	54.1 (0.499; 51.3-57.0)	55.6 (0.497; 52.6-58.7)
Meta LLaMA 3	38.6 (0.487; 37.3-39.9)	45.2 (0.498; 43.1-47.3)	37.1 (0.483; 34.3-39.9)	34.0 (0.474; 31.3-36.7)	31.6 (0.465; 28.7-34.5)
OpenAI GPT-3.5	32.5 (0.469; 31.3-33.8)	36.7 (0.482; 34.6-38.7)	27.1 (0.445; 24.5-29.6)	33.6 (0.472; 30.9-36.3)	28.8 (0.453; 26.1-31.6)
Google Bard (Legacy)	32.0 (0.466; 30.7-33.2)	40.5 (0.491; 38.5-42.6)	33.8 (0.473; 31.0-36.5)	17.7 (0.382; 15.5-19.9)	28.1 (0.450; 25.4-30.9)
Cohere Command R+	5.1 (0.219; 4.5-5.6)	4.1 (0.199; 3.3-5.0)	6.3 (0.242; 4.9-7.6)	3.0 (0.170; 2.0-3.9)	8.1 (0.273; 6.4-9.7)
Meta LLaMA 2	0.3 (0.055; 0.2-0.5)	0.2 (0.043; 0.0-0.4)	0.5 (0.072; 0.1-0.9)	0.3 (0.050; 0.0-0.5)	0.4 (0.063; 0.0-0.8)

[a]Values represent mean per-patient accuracy (the proportion of correctly extracted data points per patient, averaged across all 172 patients), with 95% CIs calculated via the *t* distribution; this is the same unit of analysis used for the paired-samples comparisons in Multimedia Appendix 4. Columns represent the overall accuracy across all data points and stratified performance in 4 question categories: binary (yes or no), numeric extraction, multiple-choice classification, and free-text interpretation. The complete results for individual data points, including pairwise statistical comparisons, and effect sizes are presented in Multimedia Appendix 4.

Discussion

Principal Findings

In this study, we evaluated 19 contemporary LLMs on structured data extraction from real neurosurgical EHRs written in Czech. The results demonstrated that current frontier models could achieve near-perfect machine-readability and high extraction accuracy under real-world conditions, with GPT-4.1 achieving the highest overall accuracy (87.6%), while older or smaller systems showed substantial limitations. GPT-4.1 significantly outperformed all models except GPT-4.5, against which no significant difference was detected.

Interpretation and Comparison With Prior Work

Several state-of-the-art models (Claude 4.1 Opus, Grok 4, and Gemini 2.5 Flash) produced 100% valid JSON outputs across all cases, and models such as GPT-4.1, GPT-5, GPT-4.5, and DeepSeek V2 performed similarly, failing only on isolated prompts. This indicates that syntactic correctness is now effectively solved for top-tier systems, at least for structured extraction tasks. In contrast, earlier models—GPT-3.5, Bard, Command R+, and LLaMA-2—often failed at the basic level of producing complete JSON outputs, illustrating that model choice is a primary determinant of feasibility in automated data extraction.

Beyond structure, the strongest models reached 85% to 88% correctness, even though they operated on long,

heterogeneous documents spanning multiple note types without curation. Accuracy followed a consistent gradient: highest for binary variables, slightly lower for numerical values, and lowest for multiple-choice and open-text fields. This pattern aligns with prior work showing that constrained outputs (eg, yes or no and numbers) are easier for LLMs to reliably extract than unconstrained text interpretation from EHRs [36-40]. In practical terms, this suggests that operationalizing clinical variables into structured formats—where possible—is key to robust automated extraction.

Most previous benchmarks have evaluated LLMs on English-language exam questions or short, synthetic tasks rather than full EHRs [1-3]. A smaller body of work addresses information extraction from clinical notes, but typically relies on retrospective datasets without explicit consent, with narrower task definitions and more controlled inputs [39,40]. Our design expands on this by using prospectively collected, consent-based EHRs in a non-English language and by enforcing both machine-readability and strict field completeness, a requirement that reflects real needs in registry building and retrospective chart review [41].

The strong performance observed in our dataset supports emerging evidence that frontier LLMs now provide generalist medical competence that extends beyond English and beyond exam-style testing [36,37]. Notably, a neurosurgery-focused, context-enriched model (neuroGPT-X [31]) did not outperform the newest general models despite domain adaptation. This aligns with recent findings showing that baseline model capacity and alignment now outweigh narrow domain tuning for many extraction tasks [37,42]. Domain-specialized approaches may remain useful for rare conditions or specific reasoning tasks, but for structured data mining, starting from the strongest base model appears more important than specialty fine-tuning.

For clinical research workflows such as retrospective chart review, disease registries, or quality audits, our findings support a hybrid approach: automated extraction of all structured variables using a frontier LLM, particularly binary and numeric fields, combined with targeted manual abstraction in a small subset of records to validate accuracy and detect systematic errors. This approach mirrors established practice in medical informatics, where automated pipelines are supplemented by manual gold-standard validation to maintain reliability while minimizing human workload [43-45]. A strict output schema (eg, JSON) and automatic parsing also ensure that single-field failures do not silently contaminate downstream datasets.

In addition, performance differences across question types suggest that prompt design matters. Clear field names, explicit answer formats, and reduction of inferential steps likely contribute to accuracy. While frontier models demonstrated robust multilingual performance on Czech EHRs, maintaining the original documentation language is important: translation introduces ambiguity and may degrade extraction [46]. For institutions with non-English records, native-language prompting appears preferable.

We deliberately used fully anonymized data with written informed consent, aligning with core data protection principles such as the General Data Protection Regulation definition of anonymized data falling outside the scope of personal data regulation [47,48]. This design represents a conservative but transparent approach. Future work may explore whether high-performing models on secure on-premise hardware offer a viable balance between performance and data protection when cloud-hosted models are not acceptable [49]. While regulatory questions surrounding LLM-assisted EHR analysis are evolving, our study demonstrates that meaningful benchmarking is already possible within an ethically straightforward framework.

Limitations

This was a single-center neurosurgical cohort with specific documentation patterns, which may limit generalizability. We evaluated a fixed set of 35 variables; performance for more complex phenotypes may differ. Although we used a consensus human standard, ground truth extraction is not perfect; initial interrater agreement, while high in absolute terms (92.6%), corresponded to only moderate chance-corrected agreement (κ=0.46), reflecting both kappa's sensitivity to skewed category distributions and genuine ambiguity in some source documentation. Only a single standardized prompt was used across all models; the impact of alternative prompting strategies, including chain-of-thought reasoning or few-shot examples, was not evaluated. Finally, model capabilities change rapidly; performance and rankings may evolve as new versions are released. These factors underscore the need for continuous domain-specific reevaluation rather than relying on static benchmarks.

Conclusions

This study demonstrates the significant potential of LLMs for extracting structured data from patient records. This has direct relevance for international clinical registries, multicenter research, and health care systems where manual chart review has historically been the only viable approach.

The consistent performance gradient across question types—binary > numeric > multiple-choice > free text—provides actionable guidance for study design: structuring data collection instruments to favor binary and numeric variables will maximize the reliability of automated extraction pipelines. The approach may be better suited to expediting initial data processing, enabling human professionals to rapidly obtain and interpret the results for meaningful research or clinical use. The rapid obsolescence of older models observed in this study also underscores the need for continuous, prospective benchmarking as new model generations emerge. Our study is unique in its use of prospectively collected patient records, obtained with explicit patient consent, ensuring a robust and ethically sound approach to this type of research. We discuss ethical and legal aspects of LLM use for such purposes and believe that the consent-based, fully anonymized framework demonstrated here provides a replicable model for future studies seeking to balance analytical capability with data protection obligations.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Funding

This work was supported by the Donatio Universitatis Carolinae fund of Charles University in Prague (research project "Právní podpora nových technologií a inovací v medicíně: Legal Support for New Technologies and Innovations in Medicine"), the Cooperatio Program of the 1st Medical Faculty of Charles University, and a project of the Ministry of Defense of the Czech Republic (grant MO 1012). The funders had no role in study design, data collection, analysis, interpretation of results, or writing of the manuscript.

Data Availability

Due to the sensitive nature of the data used in this study, the dataset supporting the conclusions of this article is not openly available.

Authors' Contributions

MČ and MM contributed equally as main authors.
Conceptualization: MČ, MM
Data curation: MČ, MM, KH, SH, KS, KE-HA
Formal analysis: MČ, MM
Funding acquisition: GRS, DN, PŠ
Investigation: MČ, MM, SL, RS, EG, HR-C, KH, SH
Methodology: MČ, MM, HH
Project administration: MČ, MM
Software: MČ
Supervision: GRS, DN
Writing—original draft: MČ, MM, PŠ, LŠ, IT
Writing—review and editing: MČ, MM, HH, SL, RS, EG, HR-C, PŠ, LŠ, IT, KH, SH, KS, KE-HA, GRS, DN

Conflicts of Interest

None declared.

Multimedia Appendix 1

Standardized prompt used for all model submissions, in the original Czech language. Patient documentation is represented by a placeholder.
[TXT File (Text, File), 5 KB-Multimedia Appendix 1]

Multimedia Appendix 2

English translation of the standardized prompt used for all model submissions. Patient documentation is represented by a placeholder.
[TXT File (Text, File), 5 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Machine readability of JSON outputs by model, including absolute counts and percentages for each output category and pairwise McNemar test results against the reference model with Bonferroni correction.
[XLSX File (Microsoft Excel File), 10 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Extraction accuracy by model and question type, including mean accuracy with 95% confidence intervals and pairwise paired-samples 2-tailed *t*-test results against the reference model with Bonferroni correction and Cohen d effect sizes.
[XLSX File (Microsoft Excel File), 9 KB-Multimedia Appendix 4]

References

1. Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv. Preprint posted online on Sep 7, 2020. [doi: 10.48550/arXiv.2009.03300]
2. Zhao Q, Huang Y, Lv T, et al. MMLU-CF: a contamination-free multi-task language understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2025:13371-13391. [doi: 10.18653/v1/2025.acl-long.656]
3. Wang Y, Ma X, Zhang G, et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark. arXiv. Preprint posted online on Jun 3, 2024. [doi: 10.48550/arXiv.2406.01574]

4. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
5. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
6. Du X, Zhou Z, Wang Y, et al. Testing and evaluation of generative large language models in electronic health record applications: a systematic review. medRxiv. Jun 22, 2025:2024.08.11.24311828. [doi: 10.1101/2024.08.11.24311828] [Medline: 39228726]
7. Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: a scoping review. J Int Med Res. Jul 2025;53(7):3000605251347556. [doi: 10.1177/03000605251347556] [Medline: 40615349]
8. Agrawal M, Chen IY, Gulamali F, Joshi S. The evaluation illusion of large language models in medicine. NPJ Digit Med. Oct 7, 2025;8(1):600. [doi: 10.1038/s41746-025-01963-x] [Medline: 41057566]
9. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
10. Chen Q, Hu Y, Peng X, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nat Commun. 2025;16(1):3280. [doi: 10.1038/s41467-025-56989-2] [Medline: 40188094]
11. Lu Z, Peng Y, Cohen T, Ghassemi M, Weng C, Tian S. Large language models in biomedicine and health: current research landscape and future directions. J Am Med Inform Assoc. Sep 1, 2024;31(9):1801-1811. [doi: 10.1093/jamia/ocae202] [Medline: 39169867]
12. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. Bioengineering (Basel). Jun 10, 2025;12(6):631. [doi: 10.3390/bioengineering12060631] [Medline: 40564447]
13. Wu J, Gu B, Zhou R, et al. BRIDGE: benchmarking large language models for understanding real-world clinical practice texts. Nat Biomed Eng. Jun 17, 2026. [doi: 10.1038/s41551-026-01719-2]
14. Jesus BG, Figueiredo TC, Anjos BH, et al. Benchmarking large language models for clinical data extraction from Portuguese medical notes in a university hospital. TechRxiv. Preprint posted online on Jun 23, 2025. [doi: 10.36227/techrxiv.175070182.28016701/v1]
15. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform. Jan 19, 2025;32(1):e101139. [doi: 10.1136/bmjhci-2024-101139] [Medline: 39832824]
16. Naemi A, Sahafi A. Benchmarking large language models for MIMIC-IV clinical note summarization. J Healthc Inform Res. 2025;10(1):95-115. [doi: 10.1007/s41666-025-00221-9] [Medline: 41658398]
17. OpenAI, Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
18. OpenAI, Hurst A, Lerer A, Goucher AP, et al. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
19. GPT-3.5 Turbo. OpenAI Developers. URL: https://platform.openai.com/docs/models/gpt-3.5-turbo [Accessed 2026-08-18]
20. GPT-4.1 model. OpenAI Developers. URL: https://platform.openai.com/docs/models/gpt-4.1 [Accessed 2026-08-18]
21. OpenAI GPT-4.5 system card. OpenAI Developers. 2025. URL: https://openai.com/index/gpt-4-5-system-card/ [Accessed 2026-08-18]
22. Introducing GPT-5. OpenAI. 2025. URL: https://openai.com/index/introducing-gpt-5/ [Accessed 2026-08-18]
23. An important next step on our AI journey. Google. Feb 6, 2023. URL: https://blog.google/technology/ai/bard-google-ai-search-updates/ [Accessed 2026-08-18]
24. Gemini 2.5: our most intelligent AI model. Google. Mar 25, 2025. URL: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ [Accessed 2026-08-18]
25. Claude 2. Anthropic. Jul 11, 2023. URL: https://www.anthropic.com/news/claude-2 [Accessed 2026-08-18]
26. Introducing Claude 4. Anthropic. May 22, 2025. URL: https://www.anthropic.com/news/claude-4 [Accessed 2026-08-18]
27. Claude Opus 4.1. Anthropic. Aug 5, 2025. URL: https://www.anthropic.com/news/claude-opus-4-1 [Accessed 2026-08-18]
28. Grok 3 models launch on API. SpaceXAI Docs. 2025. URL: https://docs.x.ai/developers/release-notes#grok-3-models-launch-on-api [Accessed 2026-08-19]
29. Grok 4.6. SpaceXAI Docs. URL: https://docs.x.ai/docs/models/grok-4 [Accessed 2026-08-19]

30. DeepSeek-AI, Liu A, Feng B, Wang B, Wang B, Liu B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. arXiv. Preprint posted online on May 7, 2024. [doi: 10.48550/arXiv.2405.04434]
31. Guo E, Gupta M, Sinha S, et al. neuroGPT-X: toward a clinic-ready large language model. J Neurosurg. 2023;140(4):1041-1053. [doi: 10.3171/2023.7.JNS23573] [Medline: 38564804]
32. Cohere. URL: https://cohere.com [Accessed 2026-08-19]
33. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on Jul 18, 2023. [doi: 10.48550/arXiv.2307.09288]
34. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
35. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
36. Moor M, Banerjee O, Abad ZS, et al. Foundation models for generalist medical artificial intelligence. Nature. Apr 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
37. Builtjes L, Bosma J, Prokop M, van Ginneken B, Hering A. Leveraging open-source large language models for clinical information extraction in resource-constrained settings. JAMIA Open. Oct 2025;8(5):ooaf109. [doi: 10.1093/jamiaopen/ooaf109] [Medline: 41041625]
38. Ahsan H, McInerney DJ, Kim J, et al. Retrieving evidence from EHRs with LLMs: possibilities and challenges. Proc Mach Learn Res. Jun 2024;248:489-505. [Medline: 39224857]
39. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2022:1998-2012. [doi: 10.18653/v1/2022.emnlp-main.130]
40. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. Int J Med Inform. Mar 2024;183:105321. [doi: 10.1016/j.ijmedinf.2023.105321] [Medline: 38157785]
41. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof. 2013;10:12. [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]
42. Yang J, Liu C, Deng W, et al. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. Patterns. Jan 12, 2024;5(1):100887. [doi: 10.1016/j.patter.2023.100887] [Medline: 38264716]
43. Yin AL, Guo WL, Sholle ET, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. Int J Med Inform. Jan 2022;157:104622. [doi: 10.1016/j.ijmedinf.2021.104622] [Medline: 34741892]
44. Joseph N, Lindblad I, Zaker S, et al. Automated data extraction of electronic medical records: validity of data mining to construct research databases for eligibility in gastroenterological clinical trials. Ups J Med Sci. 2022;127. [doi: 10.48101/ujms.v127.8260] [Medline: 35173908]
45. Goel A, Gueta A, Gilon O, et al. LLMs accelerate annotation for medical information extraction. arXiv. Preprint posted online on Dec 4, 2023. [doi: 10.48550/arXiv.2312.02296]
46. Qiu P, Wu C, Zhang X, et al. Towards building multilingual language model for medicine. Nat Commun. Sep 27, 2024;15(1):8384. [doi: 10.1038/s41467-024-52417-z] [Medline: 39333468]
47. Langarizadeh M, Orooji A, Sheikhtaheri A. Effectiveness of anonymization methods in preserving patients' privacy: a systematic literature review. Stud Health Technol Inform. 2018;248:80-87. [Medline: 29726422]
48. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). European Union. 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [Accessed 2025-12-09]
49. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. Jun 2023;5(6):e333-e335. [doi: 10.1016/S2589-7500(23)00083-3] [Medline: 37120418]

Abbreviations

EHR: electronic health record
LLM: large language model
NLP: natural language processing
TRIPOD-LLM: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Large Language Model

Edited by Luke MacNeill; peer-reviewed by Fabian Kahl, Wenzheng Sun; submitted 04.Feb.2026; initial comments to author 28.Apr.2026; final revised version received 09.Jul.2026; accepted 10.Jul.2026; published 24.Nov.2026

Please cite as:
Černý M, Májovský M, Lama S, Guo E, Hallak H, Singh R, Riva-Cambrin H, Šustek P, Široká L, Horčičáková K, Hužava S, Sajfrídová K, El-Haj Ahmadová K, Sutherland GR, Netuka D
Large Language Model Performance in Structured Data Extraction From Neurosurgical Electronic Health Records: Comparative Benchmark Study
JMIR Form Res 2026;10:e92871
URL: https://formative.jmir.org/2026/1/e92871
doi: 10.2196/92871

© Martin Černý, Martin Májovský, Sanju Lama, Edward Guo, Hana Hallak, Rahul Singh, Homer Riva-Cambrin, Petr Šustek, Lucie Široká, Katarina Horčičáková, Samuel Hužava, Kateřina Sajfrídová, Klára El-Haj Ahmadová, Garnette Roy Sutherland, David Netuka. Originally published in JMIR Formative Research (https://formative.jmir.org), 24.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.