


Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis

Nils C. Lehnen, MD • Franziska Dorn, MD • Isabella C. Wiest, MD, MSc • Hanna Zimmermann, MD • Alexander Radbruch, MD, JD • Jakob Nikolas Kather, MD, MSc • Daniel Paech, MD, PhD

From the Department of Neuroradiology, University Hospital Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Venusberg-Campus 1, 53127 Bonn, Germany (N.C.L., F.D., A.R., D.P.); Research Group Clinical Neuroimaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany (N.C.L., A.R.); Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany (I.C.W.); Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany (I.C.W., J.N.K.); Institute of Neuroradiology, University Hospital, LMU Munich, Munich, Germany (H.Z.); and Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Mass (D.P.). Received October 10, 2023; revision requested December 18; final revision received February 18, 2024; accepted March 12. Address correspondence to N.C.L. (email: nils.lehnen@ukbonn.de).

Conflicts of interest are listed at the end of this article.

Radiology 2024; 311(1):e232741 • <https://doi.org/10.1148/radiol.232741> • Content codes:  

Background: Procedural details of mechanical thrombectomy in patients with ischemic stroke are important predictors of clinical outcome and are collected for prospective studies or national stroke registries. To date, these data are collected manually by human readers, a labor-intensive task that is prone to errors.

Purpose: To evaluate the use of the large language models (LLMs) GPT-4 and GPT-3.5 to extract data from neuroradiology reports on mechanical thrombectomy in patients with ischemic stroke.

Materials and Methods: This retrospective study included consecutive reports from patients with ischemic stroke who underwent mechanical thrombectomy between November 2022 and September 2023 at institution 1 and between September 2016 and December 2019 at institution 2. A set of 20 reports was used to optimize the prompt, and the ability of the LLMs to extract procedural data from the reports was compared using the McNemar test. Data manually extracted by an interventional neuroradiologist served as the reference standard.

Results: A total of 100 internal reports from 100 patients (mean age, 74.7 years \pm 13.2 [SD]; 53 female) and 30 external reports from 30 patients (mean age, 72.7 years \pm 13.5; 18 male) were included. All reports were successfully processed by GPT-4 and GPT-3.5. Of 2800 data entries, 2631 (94.0% [95% CI: 93.0, 94.8]; range per category, 61%–100%) data points were correctly extracted by GPT-4 without the need for further postprocessing. With 1788 of 2800 correct data entries, GPT-3.5 produced fewer correct data entries than did GPT-4 (63.9% [95% CI: 62.0, 65.6]; range per category, 14%–99%; $P < .001$). For the external reports, GPT-4 extracted 760 of 840 (90.5% [95% CI: 88.3, 92.4]) correct data entries, while GPT-3.5 extracted 539 of 840 (64.2% [95% CI: 60.8, 67.4]); $P < .001$.

Conclusion: Compared with GPT-3.5, GPT-4 more frequently extracted correct procedural data from free-text reports on mechanical thrombectomy performed in patients with ischemic stroke.

© RSNA, 2024

Supplemental material is available for this article.

Mechanical thrombectomy in acute ischemic stroke has become the standard of care for patients with large vessel occlusion (1–6). Standard procedural details such as door-to-groin puncture time, door-to-reperfusion time, Alberta Stroke Program Early CT Score (ASPECTS), or number of thrombectomy maneuvers are collected in national or institutional stroke registries for research and quality assurance purposes. These data are usually extracted and transferred manually by human readers, a labor-intensive task with the risk of transmission errors.

OpenAI's generative pretrained transformer (GPT) versions 4 (GPT-4 [7]) and 3.5 (GPT-3.5 [8]) are large language models (LLMs) that have recently attracted great attention among the public. For radiologic purposes, it has been shown that GPT-4 could create standardized reports from free-text radiology reports without missing any key features (9) or simplify radiology reports without adding inaccurate or missing relevant information. In a recent study (10), 83.3% of simplified reports with GPT-4

contained no inaccurate information. Also, ChatGPT has been evaluated for its ability to answer patient questions about lung cancer (11) or breast cancer prevention or screening (12), with 70.8% and 88.0% correct answers, respectively. ChatGPT has also been shown to correctly answer 69% of radiology board-style examination questions (13) and solve 54% of Diagnosis Please quizzes when given patient medical history and imaging findings (14).

Several prior studies used natural large language processing models to extract data from radiology reports on neuroradiologic CT or MRI (15–22), but no studies, to our knowledge, have evaluated ChatGPT or other LLMs for data extraction from reports of neurointerventional procedures, such as mechanical thrombectomy. The aim of this work was to assess whether GPT-4 can extract data from free-text neuroradiology reports on mechanical thrombectomy in patients with acute ischemic stroke to create a database containing standard procedural data, clinical data, and data on materials or medication used. Another aim of

Abbreviations

ASPECTS = Alberta Stroke Program Early CT Score, LLM = large language model

Summary

By using data manually extracted by a radiologist as the ground truth, GPT-4 more frequently extracted correct data from free-text reports on mechanical thrombectomy in patients with ischemic stroke compared with GPT-3.5.

Key Results

- In a retrospective study including 100 reports on mechanical thrombectomy in patients with ischemic stroke, GPT-4 correctly extracted 2631 of 2800 (94.0% [range per category, 61%–100%]) data points while GPT-3.5 correctly extracted 1788 of 2800 (63.9% [range per category, 14%–99%]).
- Using 30 external reports on mechanical thrombectomy, GPT-4 extracted 90.5% correct data entries and GPT-3.5 extracted 64.2%.

this study was to evaluate if GPT-3.5, the predecessor of GPT-4, which was free of charge, can serve as an alternative to its more recent successor.

Materials and Methods

Ethics Approval

Institutional review board approval was obtained for this retrospective study and the need for written informed consent was waived.

Study Sample

Consecutive reports from patients with ischemic stroke who underwent mechanical thrombectomy between November 2022 and September 2023 were extracted from a picture archiving and communication system at a single institution. Inclusion criteria were patient age greater than 18 years and intracranial large or medium vessel occlusion confirmed at CT or MRI with intention to treat by means of mechanical thrombectomy. Exclusion criteria were the absence of a detailed report or the absence of intracranial occlusion at digital subtraction angiography. No statistical power analysis was performed prior to data acquisition to determine the study size, neither for the internal nor the external data set, but the inclusion of at least 100 free-text reports was the goal based on prior similar studies (10,23).

To investigate the generalizability of the LLM's ability to correctly extract data from reports, another 30 reports on mechanical thrombectomy performed between September 2016 and December 2019 at an outside German institution were obtained.

Data Extraction from Text Reports

A German prompt was created by one of the authors (N.C.L.) and tested on 20 reports to identify errors and optimize the instructions given to the LLMs. The experiments were conducted in German. An English version of the prompt is provided in Appendix S1. The LLMs were instructed to create comma-separated value (better known as CSV) tables for each report with procedural details. The internal and external reports and the prompt

were copied and pasted to the browser version of ChatGPT with default settings. No automated pipeline for processing the reports was used. The CSV code provided by the LLMs was copied to a text (txt) file and then converted to a CSV table. For each procedural detail, the LLMs were given detailed instructions that included the options for data entries such as “yes,” “no,” or “missing.” The categories that were extracted from the reports, as well as the instructions for the LLMs for each procedural detail, are summarized in Table 1.

The reports were also assessed by an interventional neuro-radiologist (N.C.L., 8 years of experience in radiology and 3 years of experience in interventional neuroradiology), and data were extracted into a CSV table manually to create the reference standard. During this process, the neuroradiologist was blinded to the results produced by the LLM. The data were only extracted from procedure reports, with no data acquired from other external material.

Evaluation of ChatGPT Data Entries

Only data entries of the LLMs that exactly matched the expert's readings were counted as correct. Any deviation from the given options in the prompt was counted as false, including synonyms, punctuation marks, or any additional symbols entered by the LLM. If a data point was not included in the report, it was declared as “missing” by the neuroradiologist. When the LLM also declared certain information as “missing,” the data entry was categorized as correct. If more than the 28 columns asked for in the prompt were created by the LLMs, surplus columns generated by the LLMs were neglected when importing them into the statistics software.

The incorrect data entries made by the LLMs were reviewed by the neuroradiologist to determine if they were content errors or format errors. A format error was defined as a data entry made by the LLM that did not meet the criteria for a correct data entry, but still was correct in terms of content (for example, “ICA” instead of “carotid,” as required by the prompt, or “Yes” with a capital letter instead of “yes,” as required by the prompt). A content error was defined as a data entry that was simply incorrect, like “M2” instead of “M1,” or “yes” instead of “no.”

For the category with the poorest result for GPT-4 for institution 1, the last thrombectomy maneuver with 39 incorrect data entries, the analysis was repeated for the reports with incorrect data entries with a more detailed prompt. The more detailed prompt included the information in the first-pass thrombectomy maneuver; the time of the first maneuver, which was also the same as that of the last maneuver; and clear instruction to fill in “missing” when the time of the last thrombectomy maneuver was not explicitly mentioned in the report.

Statistical Analysis

Statistical analyses were performed by one of the authors (N.C.L.) using R (version 4.0.3; <https://www.r-project.org/>) and RStudio (version 1.2.5033; <https://rstudio.org/download/desktop>). The CSV tables created by the LLMs were merged in R and compared with the data set created by the neurointerventionalist. Interrater agreement was assessed by using Cohen κ values, where κ less than 0.20 was indicative of poor agreement;

Table 1: Summary of Instructions and Options Given to the Large Language Models for Each Procedural Detail

Procedural Detail	Instructions and Options
Date of intervention	Format dd.mm.yyyy
Location of vessel occlusion	Choose from carotid, carotid terminus, M1, M2, M3, A1, A2, A3, P1, P2, P3, basilar, or unknown
Side of vessel occlusion	Left, right, or not applicable in the case of basilar artery occlusion
NIHSS score	NIHSS score as mentioned in the report or “missing” if not mentioned
ASPECTS	ASPECTS as mentioned in the report or “missing” if not mentioned
Intravenous thrombolysis	Yes, no, or “missing” if not mentioned
Procedure times	
Symptom onset	Format hh:mm, or “missing” if not mentioned
Arrival at thrombectomy center	Format hh:mm, or “missing” if not mentioned
Stroke imaging	Format hh:mm, or “missing” if not mentioned
Groin puncture	Format hh:mm, or “missing” if not mentioned
First intracranial run	Format hh:mm, or “missing” if not mentioned
First thrombectomy maneuver	Format hh:mm, or “missing” if not mentioned
Last thrombectomy maneuver	Format hh:mm, or “missing” if not mentioned
Last run	Format hh:mm, or “missing” if not mentioned
Technical details	
No. of thrombectomy maneuvers	No. of thrombectomy maneuvers if specified in the report. May be calculated, if possible, from the free-text report.
mTICI score	mTICI score as specified in the report. Options: 0, 1, 2a, 2b, 2c, 3. “0,” if a futile thrombectomy was performed; “3,” if complete recanalization is mentioned in the report; “missing,” if no mTICI score can be extracted from the report
Use of BGC	Yes, if a BGC is mentioned, or no. BGCs typically used in our institution were given as examples.
Use of distal aspiration	Yes, if distal aspiration was performed, or no. Typical aspiration catheters used in our institution were given as examples.
Use of stent retriever	Yes, if stent retrievers were used, or no. Typical stent retrievers used in our institution were given as examples.
Extracranial stent	Yes, if an extracranial stent was placed, or no
Intracranial stent	Yes, if an intracranial stent was placed, or no
Periprocedural medication	
ASA	Yes, if ASA was administered, or no
Clopidogrel	Yes, if clopidogrel was administered, or no
Ticagrelor	Yes, if ticagrelor was administered, or no
Tirofiban	Yes, if tirofiban was administered, or no
Heparin	Yes, if heparin was administered, or no
FPCT performed	Yes, if an FPCT was performed, or no
ICH	Yes, if ICH is mentioned in the report, or no, if no hemorrhage is described.

Note.— A1 = anterior cerebral artery A1 segment, A2 = anterior cerebral artery A2 segment, A3 = anterior cerebral artery A3 segment, ASA = acetylic salicylic acid, ASPECTS = Alberta Stroke Program Early CT Score, BGC = balloon guide catheter, FPCT = flat-panel CT, ICH = intracranial hemorrhage, M1 = middle cerebral artery M1 segment, M2 = middle cerebral artery M2 segment, M3 = middle cerebral artery M3 segment, mTICI = modified Thrombolysis in Cerebral Infarction, NIHSS = National Institutes of Health Stroke Scale, P1 = posterior cerebral artery P1 segment, P2 = posterior cerebral artery P2 segment, P3 = posterior cerebral artery P3 segment.

0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, good agreement, and 0.81–1.00, very good agreement. For the comparison of GPT-4 versus GPT-3.5, the McNemar test was used. The level of statistical significance was set at $P = .05$. No correction for multiple testing was performed.

Results

Study Sample

A total of 107 reports from 107 patients were initially acquired. No patients were excluded due to absence of a written

report, but seven patients were excluded due to absence of intracranial occlusion. Thus, a total of 100 patients (mean age, 74.7 years \pm 13.2 [SD]; 47 male, 53 female) and corresponding reports were included. A flowchart of the inclusion and exclusion process, as well as further data analysis, is provided in Figure 1.

The reports included in this study were written by six different neurointerventionalists. All reports were successfully processed by both GPT-4 and GPT-3.5. GPT-4 generated tables with one column per extracted variable ($n = 28$), whereas GPT-3.5 generated tables with varying numbers of columns (range, 26–30 columns).

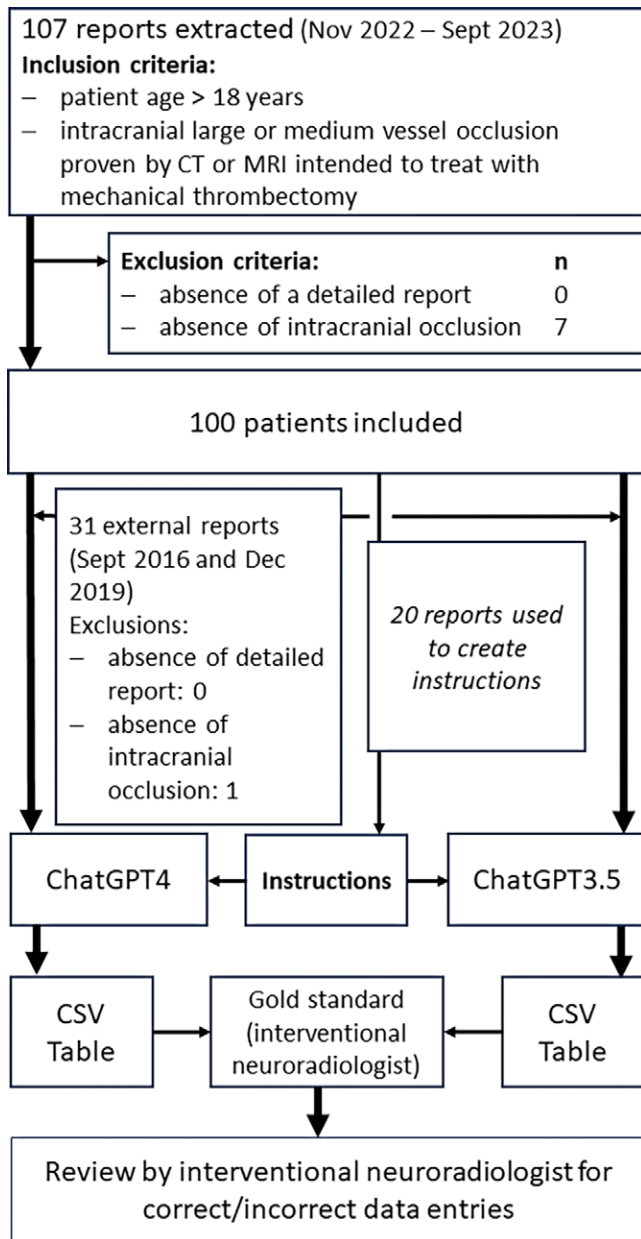


Figure 1: Flowchart of patient inclusion and exclusion criteria, data analysis, and generation of results. The same 30 external reports were used for GPT-4 and GPT-3.5. CSV = comma-separated value.

In addition, 31 external reports were available, of which one was excluded due to absence of intracranial vessel occlusion, and none was excluded due to absence of a written report. The remaining 30 external reports from 30 patients (mean age, 72.7 years ± 13.5; 18 male, 12 female) were written by five different neurointerventionalists. All external reports were successfully processed by both GPT-4 and GPT-3.5. Patient characteristics and basic clinical and procedural details are summarized in Table 2.

Data Extraction by GPT-4 and GPT-3.5

Of the 2800 data entries made by the interventional neuroradiologist, 2631 (94.0% [95% CI: 93.0, 94.8]) were correctly entered by GPT-4, which was higher than the 1788 (63.9%

Table 2: Patient Characteristics and Key Clinical and Procedural Information

Characteristic	Institution 1	Institution 2
Sex		
F	47 (47.0)	12 (40.0)
M	53 (53.0)	18 (60.0)
Mean age (y)*	74.7 ± 13.2	72.7 ± 13.5
Age range (y)	23–98	39–92
Occlusion site		
ICA	19 (19.0)	3 (10.0)
M1	51 (51.0)	19 (63.3)
M2	22 (22.0)	0 (0.0)
A1	0 (0.0)	0 (0.0)
A2	0 (0.0)	1 (3.3)
BA	7 (7.0)	7 (23.3)
P1	0 (0.0)	0 (0.0)
P2	1 (1.0)	0 (0.0)
Median NIHSS score†	8 (0–24)	Not applicable
Median ASPECTS†	9 (3–10)	Not applicable
Intravenous thrombolysis	23 (23.0)	13 (43.3)
Mean door-to-groin puncture time (min)*	43 ± 24	Not applicable
Mean groin-to-reperfusion time (min)*	71 ± 172	28 ± 16
mTICI		
0	2 (2.0)	1 (3.3)
1	0 (0.0)	0 (0.0)
2a	3 (3.0)	1 (3.3)
2b	26 (26.0)	15 (50.0)
2c	20 (20.0)	2 (6.7)
3	48 (48.0)	9 (30.0)
Mean no. of maneuvers*	2.9 ± 2.5	2.4 ± 1.9
Extracranial stent placement	9 (9.0)	2 (6.7)
Intracranial stent placement	2 (2.0)	3 (10.0)
Postinterventional ICH	17 (17.0)	2 (6.7)

Note.— Unless otherwise indicated, data are numbers of patients and data in parentheses are percentages. Not applicable indicates that the data included in the reports were not sufficient to provide the respective information. A1 = anterior cerebral artery A1 segment, A2 = anterior cerebral artery A2 segment, ASPECTS = Alberta Stroke Program Early CT Score, BA = basilar artery, ICA = internal carotid artery, ICH = intracranial hemorrhage, M1 = middle cerebral artery M1 segment, M2 = middle cerebral artery M2 segment, mTICI = modified Thrombolysis in Cerebral Infarction, NIHSS = National Institutes of Health Stroke Scale, P1 = posterior cerebral artery P1 segment, P2 = posterior cerebral artery P2 segment.

* Data are mean ± SDs.

† Data in parentheses are the range.

[95% CI: 62.0, 65.6]; *P* < .001) that were correctly entered by GPT-3.5. When considering individual categories, correct entries by GPT-4 ranged from 61 of 100 (61.0% [95% CI: 50.7, 70.6]) for time of last thrombectomy maneuver to 100 of 100 (100.0% [95% CI: 96.4, 100.0]) for date of intervention, National Institutes of Health Stroke Scale score, ASPECTS, stent retriever, extracranial stent, and intracranial

Table 3: Correct Data Entries Made by ChatGPT for Each Procedural Detail

Procedural Detail	Correct Data Entries by GPT-4	Correct Data Entries by GPT-3.5	P Value
Total no. of data entries	94.0 (93.0, 94.8)	63.9 (62.0, 65.6)	<.001
Date of intervention	100.0 (96.4, 100.0)	99.0 (94.6, 100.0)	>.99
Location of vessel occlusion	87.0 (78.8, 92.9)	80.0 (70.8, 87.3)	.09
Side of vessel occlusion	98.0 (93.0, 99.8)	86.0 (77.6, 92.1)	.001
NIHSS score	100.0 (96.4, 100)	83.0 (74.2, 89.8)	<.001
ASPECTS	100.0 (96.4, 100)	88.0 (80.0, 93.6)	.001
Intravenous thrombolysis	88.0 (80.0, 93.6)	39.0 (29.4, 49.3)	<.001
Procedure time			
Symptom onset	96.0 (90.1, 98.9)	64.0 (53.8, 73.4)	<.001
Arrival at thrombectomy center	90.0 (82.4, 95.1)	72.0 (62.1, 80.5)	<.001
Stroke imaging	98.0 (93.0, 99.8)	81.0 (71.9, 88.2)	<.001
Groin puncture	96.0 (90.1, 98.9)	88.0 (80.0, 93.6)	.03
First intracranial run	94.0 (87.4, 97.8)	84.0 (75.3, 90.6)	.02
First thrombectomy maneuver	95.0 (88.7, 98.4)	79.0 (69.7, 86.5)	.001
Last thrombectomy maneuver	61.0 (50.7, 70.6)	24.0 (16.0, 33.6)	<.001
Last run	96.0 (90.1, 98.9)	24.0 (16.0, 33.6)	<.001
Technical details			
No. of thrombectomy maneuvers	91.0 (83.6, 95.8)	22.0 (14.3, 31.4)	<.001
mTICI score	93.0 (86.1, 97.1)	14.0 (7.9, 22.4)	<.001
Use of BGC	89.0 (81.2, 94.4)	51.0 (40.1, 61.1)	<.001
Use of distal aspiration	99.0 (94.6, 100.0)	81.0 (71.9, 88.2)	<.001
Use of stent retriever	100.0 (96.4, 100.0)	65.0 (54.8, 74.3)	<.001
Extracranial stent	100.0 (96.4, 100.0)	91.0 (83.6, 95.8)	.008
Intracranial stent	100.0 (96.4, 100.0)	89.0 (81.2, 94.4)	.003
Periprocedural medication			
ASA	97.0 (91.5, 99.4)	64.0 (53.8, 73.4)	<.001
Clopidogrel	97.0 (91.5, 99.4)	70.0 (60.0, 78.8)	<.001
Ticagrelor	90.0 (82.4, 95.1)	62.0 (51.7, 71.5)	<.001
Tirofiban	98.0 (93.0, 99.8)	37.0 (28.6, 47.2)	<.001
Heparin	99.0 (94.6, 100.0)	61.0 (50.7, 70.6)	<.001
FPCT performed	86.0 (77.6, 92.1)	49.0 (38.9, 59.2)	<.001
ICH	93.0 (86.1, 97.1)	41.0 (31.3, 51.3)	<.001

Note.—Unless otherwise indicated, data are percentages and data in parentheses are 95% CIs. Comparisons between GPT-4 and GPT-3.5 were made using the McNemar test. ASA = acetylsalicylic acid, ASPECTS = Alberta Stroke Program Early CT Score, BGC = balloon guide catheter, FPCT = flat-panel CT, ICH = intracranial hemorrhage, mTICI = modified Thrombolysis in Cerebral Infarction, NIHSS = National Institutes of Health Stroke Scale.

stent. Correct entries per category for GPT-3.5 ranged from 14 of 100 (14.0% [95% CI: 7.9, 22.4]) for modified Thrombolysis in Cerebral Infarction score to 99 of 100 (99.0% [95% CI: 94.6, 100.0]) for date of intervention. A detailed overview of the number of correct entries made by the LLMs for each category is given in Table 3 and Figure 2. Very good agreement was observed between GPT-4 and the neuroradiologist ($\kappa = 0.93$), and moderate agreement was observed between GPT-3.5 and the neuroradiologist ($\kappa = 0.59$).

Two example reports from institutions 1 and 2 along with the data entries of the neuroradiologist and both LLMs can be found in Tables S1 and S2, respectively.

Incorrect Data Entries by GPT-4 and Chat GPT-3.5

A total of 169 data entries by GPT-4 were deemed incorrect, of which 19 (11.2%) were due to format errors but were correct in terms of content and 150 (88.8%) were incorrect in terms

of content. When a more detailed prompt was used to try to improve the number of correct data entries by GPT-4 for the category of last thrombectomy maneuver, the number of incorrect entries decreased from 39 to 19. For GPT-3.5, 1012 data entries were deemed incorrect, of which 86 (8.5%) were due to format errors and 926 (91.5%) were incorrect in terms of content. A complete list of errors made by GPT-4 is available in Table S3, including the differentiation between format errors and errors in terms of content.

Data Extraction by ChatGPT in an External Data Set

For the external reports, 760 of 840 (90.5% [95% CI: 88.3, 92.4]) data entries made by GPT-4 were correct, which was higher than the 539 of 840 (64.2% [95% CI: 60.8, 67.4]; $P < .001$) that were correctly entered by GPT-3.5. A detailed overview of the number of correct entries made by the LLMs on the external reports can be found in Table 4.

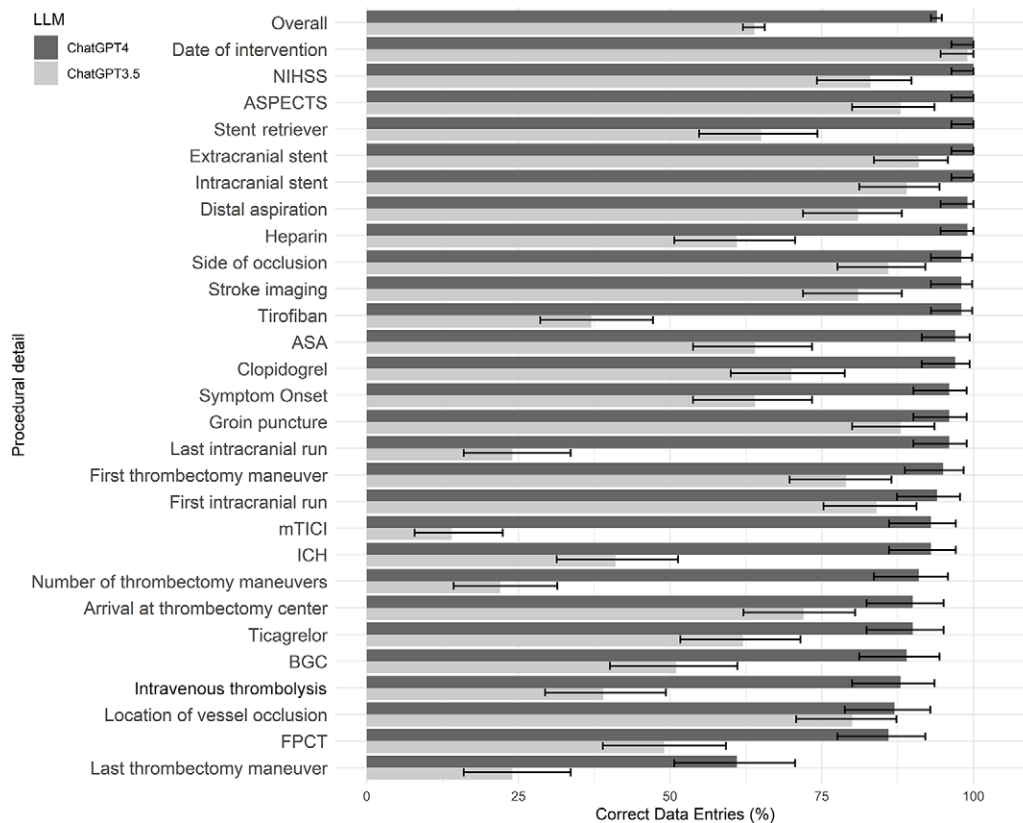


Figure 2: Bar plot shows the percentage of correct data entries extracted from radiology reports by GPT-4 and GPT-3.5, overall and for each procedural detail. Error bars represent 95% CIs. ASA = acetylsalicylic acid, ASPECTS = Alberta Stroke Program Early CT Score, BGC = balloon guide catheter, FPCT = flat-panel CT, ICH = intracranial hemorrhage, LLM = large language model, mTICI = modified Thrombolysis in Cerebral Infarction, NIHSS = National Institutes of Health Stroke Scale.

Discussion

Procedural details of mechanical thrombectomy in patients with ischemic stroke are predictors of clinical outcome and are collected for prospective studies or national stroke registries. To date, these data are collected by human readers, which is a labor-intensive task that is prone to errors. In this retrospective study, we assessed the ability of GPT-4 and GPT-3.5 to extract data from free-text reports on mechanical thrombectomy. GPT-4 correctly extracted 94.0% (2631 of 2800) of data entries, which was higher than those extracted by GPT-3.5 (63.9% [1788 of 2800]; $P < .001$), compared with the data extracted by a neurointerventionalist as the reference. We validated our results by performing the same analysis for 30 external reports, for which GPT-4 achieved 90.5% (760 of 840) correct data entries and GPT-3.5 achieved 64.2% (539 of 840; $P < .001$).

Recently, numerous studies have shown that GPT-4 could be used for data mining from free-text reports in oncology, with correct extraction of lesion parameters in 98.6% (24); generating differential diagnoses from written reports, with a 68.8% concordance to an expert panel of radiologists (23); or determining the correct radiologic study from the radiology request form, with a success rate of 84% (25). Ong et al (15) developed several models to identify characteristics of ischemic stroke from radiologic reports and achieved accuracy of 89% and 91% for detecting stroke and middle cerebral artery location, respectively.

Using their natural language processing approach, Yu et al (17) reported an accuracy of 97.3% for detecting large vessel occlusion from radiologic reports. Li et al (16) reported a 97% accuracy for stroke detection for their algorithm. Gunter et al (20) reported an accuracy of greater than 90.0% for identifying a variety of stroke characteristics. Siddiqui et al (21) reported 98% correct data extractions for core and penumbra volumes for their natural language processing approach. Although the studies did not investigate the ability of LLMs to extract data from reports on mechanical thrombectomy, our results are in line with their results and confirm the usability of LLMs for data extraction from free-text reports.

We observed that GPT-4 had higher rates of correct data entries than did GPT-3.5, as has been described by previous studies such as that by Rao et al (26), who found that GPT-4 was superior to GPT-3.5 for using the American College of Radiology appropriateness criteria for breast pain and breast cancer screening, or Li et al (27), who showed that GPT-4 was superior to GPT-3.5 for correctly solving Diagnosis Please quizzes. However, the question of whether GPT-3.5 could also deliver sufficient results for the specific task of data extraction from thrombectomy reports may be of interest to potential users and was thus included in our study.

We observed poor results for certain data points, which highlights the fact that human supervision is still needed. For the last thrombectomy maneuver, we repeated the analysis for the 39 reports with incorrect data entries made by GPT-4 with a

Table 4: Correct Data Entries Made by ChatGPT for Each Procedural Detail (External Reports)

Procedural Detail	Correct Data Entries by GPT-4	Correct Data Entries by GPT-3.5	<i>P</i> Value
Total no. of data entries	90.5 (88.3, 92.4)	64.2 (60.8, 67.4)	<.001
Date of intervention	100.0 (88.4, 100.0)	100.0 (88.4, 100.0)	NA
Location of vessel occlusion	90.0 (73.5, 97.9)	80.0 (61.4, 92.3)	.25
Side of vessel occlusion	93.3 (77.9, 99.2)	70.0 (50.6, 85.3)	.02
NIHSS score	100.0 (88.4, 100.0)	86.7 (69.3, 96.2)	.13
ASPECTS	100.0 (88.4, 100.0)	86.7 (69.3, 96.2)	.13
Intravenous thrombolysis	93.3 (77.9, 99.2)	70.0 (50.6, 85.3)	.02
Procedure times			
Symptom onset	83.3 (65.3, 94.4)	70.0 (50.6, 85.3)	.39
Arrival at thrombectomy center	100.0 (88.4, 100.0)	70.0 (50.6, 85.3)	.008
Stroke imaging	86.7 (69.3, 96.2)	66.7 (47.2, 82.7)	.11
Groin puncture	56.7 (37.4, 74.5)	46.7 (28.3, 65.7)	.55
First intracranial run	96.7 (82.8, 99.9)	60.0 (40.6, 77.3)	.006
First thrombectomy maneuver	90.0 (73.5, 97.9)	43.3 (25.5, 62.6)	.002
Last thrombectomy maneuver	53.3 (34.3, 71.2)	26.7 (12.3, 45.9)	.03
Last run	93.3 (77.9, 99.2)	46.7 (28.3, 65.7)	.002
Technical details			
No. of thrombectomy maneuvers	46.7 (28.3, 65.7)	33.3 (17.3, 52.8)	.29
mTICI score	93.3 (77.9, 99.2)	53.3 (34.3, 71.2)	.001
Use of BGC	90.0 (73.5, 97.9)	76.7 (57.7, 90.1)	.34
Use of distal aspiration	100.0 (88.4, 100.0)	86.7 (69.3, 96.2)	.13
Use of stent retriever	93.3 (77.9, 99.2)	86.7 (69.3, 96.2)	.68
Extracranial stent	100.0 (88.4, 100.0)	90.0 (73.5, 97.9)	.25
Intracranial stent	100.0 (88.4, 100.0)	90.0 (73.5, 97.9)	.25
Periprocedural medication			
ASA	96.7 (82.8, 99.9)	73.3 (54.1, 87.7)	.046
Clopidogrel	93.3 (77.9, 99.2)	33.3 (17.3, 52.8)	<.001
Ticagrelor	100.0 (88.4, 100.0)	66.7 (47.2, 82.7)	.004
Tirofiban	100.0 (88.4, 100.0)	33.3 (17.3, 52.8)	<.001
Heparin	100.0 (88.4, 100.0)	73.3 (54.1, 87.7)	.01
FPCT performed	83.3 (65.3, 94.4)	23.3 (9.9, 42.3)	<.001
ICH	100.0 (88.4, 100.0)	53.3 (34.3, 71.7)	<.001

Note.— Unless otherwise indicated, data are percentages and data in parentheses are 95% CIs. Comparisons between GPT-4 and GPT-3.5 were made by using the McNemar test. ASA = acetylsalicylic acid, ASPECTS = Alberta Stroke Program Early CT Score, BGC = balloon guide catheter, FPCT = flat-panel CT, ICH = intracranial hemorrhage, mTICI = modified Thrombolysis in Cerebral Infarction, NA = not applicable, NIHSS = National Institutes of Health Stroke Scale.

more detailed prompt and could achieve another 20 correct data entries. Thus, we hypothesize that prompt optimization can partially reduce incorrect data entries. A discussion on how our prompt was generated can be found in Appendix S2.

We acknowledge several limitations of our study. First, its retrospective nature may limit the generalization of results. Second, data extraction by GPT-4 and GPT-3.5 was only tested on a small number of reports from an external institution, thus additional studies are necessary to validate the generalizability of our results. Third, the LLM was given examples of materials used at our institution. This may reduce the generalizability of our study to centers that use a different set of materials. Fourth, the reproducibility of GPT-4's output was not assessed, which may be a topic for future research. In a study by Samaan et al (28), patient questions on bariatric surgery were posed twice to GPT-4, and 90.7% of answers were reproducible. Fifth, the reports and the prompt were created in German language; thus, the results of our

study may need confirmation in other languages. Sixth, ChatGPT will undergo further updates, limiting the reproducibility of our results. Seventh, only one reader was used as the reference standard. However, the reader was an experienced interventional neuroradiologist with advanced understanding of the procedure, which we regarded as sufficient for this exploratory study. Last, GPT-4 is not a Food and Drug Administration– or Conformité Européenne–labeled product; thus, care must be taken when using it in the context of medical care (29).

In conclusion, GPT-4 more frequently extracted correct procedural data from reports on mechanical thrombectomy performed in patients with ischemic stroke than its predecessor GPT-3.5. This suggests that GPT-4, or other large language models, could provide an alternative to retrieving these data manually. Although GPT-4 may facilitate this process and possibly improve data extraction from radiology reports, errors currently still occur and surveillance by human readers is needed.

Deputy Editor: Yoshimi Anzai

Scientific Editor: Ariane Panzer

Acknowledgment: The authors used large language models GPT-4 and GPT-3.5 (OpenAI; <https://chat.openai.com>) to generate the reports.

Author contributions: Guarantors of integrity of entire study, **N.C.L., D.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **N.C.L., F.D., A.R., D.P.**; clinical studies, **H.Z., D.P.**; experimental studies, **N.C.L., J.N.K.**; statistical analysis, **N.C.L., J.N.K., D.P.**; and manuscript editing, all authors

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **N.C.L.** No relevant relationships. **F.D.** Research grants from Cerenovus, Johnson and Johnson; consulting fees from Cerenovus, Johnson and Johnson, Microvention, Balt, P Rctoring for Balt, Cerus Endovascular; speakers honoraria from Asahi, Acandis, Stryker, Cerenovus, Medtronic, QÀpel; payment for testimony from Cerenovus, Johnson and Johnson, Balt; advisory board member for Cerenovus; associate editor for KLNE and JCM; former associate editor for JNIS; executive board member of ESMINT and DE-GIR. **I.C.W.** No relevant relationships. **H.Z.** No relevant relationships. **A.R.** No relevant relationships. **J.N.K.** Royalties or licenses from GSK; consulting services for Owkin, France, DoMore Diagnostics, Norway, Panakeia, UK, Scailyte, Switzerland, Cancilico, Germany, Mindpeak, Germany, and Histofy, UK; honoraria for lectures from AstraZeneca, Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer, Fresenius; holds shares in StratifAI GmbH, Germany. **D.P.** Grants from German Research Foundation (DFG), Else-Kroener-Fresenius Foundation (EKFS), payment or honoraria for lectures from Siemens Healthineers.

References

- Berkhemer OA, Fransen PSS, Beumer D, et al. A randomized trial of intra-arterial treatment for acute ischemic stroke. *N Engl J Med* 2015;372(1):11–20. [Published correction appears in *N Engl J Med* 2015;372(4):394.]
- Campbell BCV, Mitchell PJ, Kleinig TJ, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med* 2015;372(11):1009–1018.
- Goyal M, Demchuk AM, Menon BK, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med* 2015;372(11):1019–1030.
- Goyal M, Menon BK, van Zwam WH, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet* 2016;387(10029):1723–1731.
- Jovin TG, Chamorro A, Cobo E, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med* 2015;372(24):2296–2306.
- Saver JL, Goyal M, Bonafe A, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med* 2015;372(24):2285–2295.
- ChatGPT-4. OpenAI. OpenAI. <https://chat.openai.com>. Accessed September 10, 2023.
- ChatGPT-3.5. OpenAI. <https://chat.openai.com>. Accessed September 10, 2023.
- Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology* 2023;307(4):e230725.
- Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. *Radiology* 2023;309(2):e232561.
- Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* 2023;307(5):e230922.
- Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology* 2023;307(4):e230424.
- Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology* 2023;307(5):e230582.
- Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. *Radiology* 2023;308(1):e231040.
- Ong CJ, Orfanoudaki A, Zhang R, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS One* 2020;15(6):e0234908.
- Li MD, Lang M, Deng F, et al. Analysis of Stroke Detection during the COVID-19 Pandemic Using Natural Language Processing of Radiology Reports. *AJNR Am J Neuroradiol* 2021;42(3):429–434.
- Yu AYZ, Liu ZA, Pou-Prom C, et al. Automating Stroke Data Extraction From Free-Text Radiology Reports Using Natural Language Processing: Instrument Validation Study. *JMIR Med Inform* 2021;9(5):e24381.
- Torres-Lopez VM, Rovenolt GE, Olcese AJ, et al. Development and Validation of a Model to Identify Critical Brain Injuries Using Natural Language Processing of Text Computed Tomography Reports. *JAMA Netw Open* 2022;5(8):e2227109.
- Miller MI, Orfanoudaki A, Cronin M, et al. Natural Language Processing of Radiology Reports to Detect Complications of Ischemic Stroke. *Neurocrit Care* 2022;37(Suppl 2):291–302.
- Gunter D, Puac-Polanco P, Miguel O, et al. Rule-based natural language processing for automation of stroke data extraction: a validation study. *Neuroradiology* 2022;64(12):2357–2362.
- Siddiqui Z, Bhatia K, Corbin A, Dhar R. Rules-based natural language processing to extract features of large vessel occlusion and cerebral edema from radiology reports in stroke patients. *Neurosci Inform* 2023;3(2):100129.
- Sung SF, Chen CH, Pan RC, Hu YH, Jeng JS. Natural Language Processing Enhances Prediction of Functional Outcome After Acute Ischemic Stroke. *J Am Heart Assoc* 2021;10(24):e023486.
- Kottlors J, Bratke G, Rauen P, et al. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. *Radiology* 2023;308(1):e231167.
- Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology* 2023;308(3):e231362.
- Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for Automated Determination of Radiological Study and Protocol based on Radiology Request Forms: A Feasibility Study. *Radiology* 2023;307(5):e230877.
- Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol* 2023;20(10):990–997.
- Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 Accuracy and Drift in *Radiology* Diagnosis Please Cases. *Radiology* 2024;310(1):e232411.
- Samaan JS, Yeo YH, Rajeev N, et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes Surg* 2023;33(6):1790–1796.
- López-Úbeda P, Martín-Noguerol T, Luna A. Radiology in the era of large language models: the near and the dark side of the moon. *Eur Radiol* 2023;33(12):9455–9457.