

Are Large Multimodal Models Ready for Post-Thrombectomy mTICI Assessment? A Comparison with Physicians

Michael Knott

`michael.knott@klinik.uni-regensburg.de`

University Hospital Regensburg

Gerardo Napodano

University Hospital Regensburg

Ekaterina Noeva

University Hospital Regensburg

Liang Zhang

University Hospital Regensburg

Moritz Brandenstein

University Hospital Regensburg

Johannes Dinkel

University Hospital Regensburg

Lili-Laura Bruta

medbo District Hospital

Julia Pistorius

medbo District Hospital

David Olmes

University Hospital Regensburg

Isabel Wiesinger

medbo District Hospital

Christian Stroszczynski

University Hospital Regensburg

Christina Wendl

University Hospital Regensburg

Quirin D. Strotzer

University Hospital Regensburg

Keywords:

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11013237/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background and Purpose. Mechanical thrombectomy is standard of care for acute ischemic stroke due to large vessel occlusion (LVO), and the modified Treatment in Cerebral Ischemia (mTICI) scale is the universal technical endpoint for postprocedural reperfusion assessment. Large language models (LLMs) and large multimodal models (LMMs) have shown strong performance in medical information extraction, but their ability to infer procedural outcomes, particularly from multimodal input, remains untested. We therefore compared LMMs with clinicians in mTICI grading after mechanical thrombectomy.

Materials and Methods. In this retrospective single-center study, 90 patients with acute anterior-circulation LVO treated between 2011 and 2023 were included. Multimodal input consisted of representative pre- and postprocedural DSA frames and the mTICI-redacted procedural reports. Reference standard was the expert-defined mTICI grade. GPT-5 (OpenAI), MedGemma-1.5 (Google), and four human readers (two neurologists, two radiologists) graded every case under three input conditions: report only, images only, and images plus report. The primary endpoint was quadratic-weighted Cohen kappa (κ) against the reference. Secondary endpoints were exact six-level accuracy, mean absolute ordinal error, and sensitivity/specificity at the reperfusion-success threshold ($\text{mTICI} \geq 2b$).

Results. GPT-5 was virtually indistinguishable from pooled readers when a report was available (images plus report κ 0.87 [95% confidence interval: 0.78, 0.92]; report only 0.86 [0.78, 0.92]; readers 0.87 [0.81, 0.91] for both), but image-only agreement collapsed to κ 0.35 [0.08, 0.56] (readers 0.85 [0.78, 0.90]). MedGemma-1.5 performed poorly throughout (κ 0.34, 0.34, and 0.02). Sensitivity was uniformly high, but specificity separated raters: GPT-5 reached 81.2% with the report but only 18.8% on images only, while MedGemma-1.5 labeled nearly every procedure successful (specificity 0.0%-6.2%). Essentially all usable model performance came from the report (report increment for GPT-5: +0.52 [0.33, 0.75]; MedGemma-1.5: +0.33 [0.14, 0.54]).

Conclusions. Multimodal frontier models can achieve physician-level mTICI agreement when procedural reports are available, but performance drops substantially on images alone. Current models may support documentation, structured data extraction, and quality assurance, while expert human interpretation remains indispensable for direct image-based reperfusion assessment.

Introduction

Acute ischemic stroke remains one of the leading causes of mortality and long-term disability worldwide [1]. Over the past decade, mechanical thrombectomy has fundamentally transformed the management of patients with large vessel occlusion (LVO). Important randomized controlled trials established the superiority of endovascular thrombectomy over best medical therapy alone, thereby defining it as the standard of care for eligible patients [2–6]. Further evidence demonstrated benefit even in patients with extended time windows and large ischemic cores, substantially broadening eligibility for endovascular treatment [7–10].

Successful reperfusion is among the strongest predictors of favorable outcome after mechanical thrombectomy, making reliable angiographic assessment of treatment success central to both clinical practice and stroke research. The modified Treatment in Cerebral Ischemia (mTICI) scale is the standard angiographic measure for evaluating cerebral reperfusion following endovascular treatment and a primary technical endpoint in thrombectomy trials and registries [11, 12]. Accurate mTICI grading therefore carries substantial clinical, scientific, and economic importance.

Despite its widespread adoption, mTICI grading remains susceptible to interobserver variability. Reported agreement is consistently only moderate, with a recognized tendency toward operator overestimation and further reduced reliability for posterior circulation and distal territory occlusions [11, 14–19]. Since treatment decisions, quality assurance, and scientific analyses increasingly rely on standardized reperfusion metrics, methods that reduce observer-dependent variability are of growing interest. Automated methods could support more standardized grading, but their value depends on whether they merely reproduce information contained in procedural reports or can independently interpret angiographic findings.

In parallel with advances in stroke therapy, artificial intelligence has emerged as a transformative technology across medicine. Large language models (LLMs) have shown strong capabilities in medical question answering, clinical reasoning, and diagnostic support, with performance comparable to or exceeding that of physicians on selected benchmark tasks [20–21]. However, performance on text-based benchmarks does not necessarily translate into reliable interpretation of specialized angiographic imaging.

Large multimodal models (LMMs) that jointly process text and images have further raised the possibility that AI systems may contribute to image-based diagnostic assessment in settings traditionally reserved for clinical experts [22–23]. Early zero-shot evaluations of general-purpose vision-language models on diagnostic imaging tasks have begun to define both the promise and the current limits of this approach [24–26]. Within radiology, LMMs have shown promising capabilities in structured reporting and information extraction, including mTICI grades from free-text thrombectomy reports [27–29]. Whether such models genuinely infer post-thrombectomy reperfusion rather than extract an already stated grade, and whether they can do so from angiographic images without accompanying text, remain untested.

We therefore compared multimodal models with neurologists and radiologists in grading post-thrombectomy mTICI against an independent reference standard under three input conditions: report only, images only, and images plus report. We hypothesized that models would match readers when a report was available but not when grading from images alone, and that readers, unlike models, would perform similarly across conditions.

Materials and Methods

Study Design and Cohort

This retrospective, single-center study was approved by the institutional review board of the University Hospital Regensburg (vote no. 26-4706-104, clinical trial number: not applicable). The requirement for informed consent was waived owing to the retrospective study design and the use of anonymized imaging data. The study was conducted in accordance with the Declaration of Helsinki and its later amendments, and the REFINE checklist for reporting foundation and large language models in medical research was followed during study design and manuscript preparation [30]. No funding was received for this study

Consecutive cases of mechanical thrombectomy for acute LVO treated between 2011 and 2023 were identified by billing codes. Patients were eligible if the occlusion was located in the anterior circulation and if pre- and post-thrombectomy digital subtraction angiography (DSA) series and the procedural radiology report were available. Cases with posterior-circulation LVO, non-evaluable angiographic data or missing reports were excluded. See Fig. 1 for a study flow chart.

Reference Standard and Data Preparation

The mTICI grade was treated as an ordinal six-level scale (0, 1, 2a, 2b, 2c, and 3), and successful reperfusion was defined as mTICI 2b or higher as defined by the Cerebral Angiographic Revascularization Grading Collaborators [12, 13]. A reference mTICI grade was established for every included case independently of the study readers and of the evaluated models in a consensus using all available information by W.C. and W.I., with 19 and 14 years of experience in neuroradiology, respectively. This reference mTICI scale was used throughout the study. A.A. and B.B. also selected the single most representative frames of each of the four dynamic DSA series (posteroanterior and lateral projections, pre- and post-thrombectomy). Each image was labeled with its acquisition phase and projection. To ensure that image interpretation was tested rather than text recognition, all frames were manually reviewed for burned-in annotations that could reveal the occlusion site or procedural outcome. Where present, these were removed in the pixel data without obscuring vascular anatomy. The report, when supplied, had the final mTICI grade removed and was written in German. Figure 2 illustrates the input data for a representative case.

Model and Physician Assessment

Every rater graded the same cases under three input conditions: report only (procedural report with the final mTICI grade redacted and no images), images only (pre- and post-thrombectomy DSA images without the report), and images plus report (both sources provided).

Two multimodal models were evaluated. GPT-5 (gpt-5-2025-08-07; OpenAI), a general-purpose multimodal model, was queried through the OpenAI Responses application programming interface with images submitted at high detail. MedGemma-1.5 (medgemma-1.5-4b-it; Google), a medically tuned open-weights vision-language model, was run locally in 16-bit precision without quantization, using the Hugging Face transformers pipeline with greedy decoding. Both models received an identical task

specification for each condition: a standardized prompt that assigned an expert-neurointerventionalist persona, listed the operational mTICI definitions, and requested the final mTICI grade with free-text reasoning returned as a structured JSON object. No in-context examples and no task-specific fine-tuning were used (zero-shot setting). The mTICI grade was parsed from the JSON output, and responses that did not yield a valid grade were treated as missing. The full prompts for the three input conditions and the required output format are provided in Appendix S1 of the Supplementary Data. Each case × input condition was assessed once per model.

Four readers independently graded all cases under each input condition in a randomized order: an experienced neurologist (C.C. with CC years of experience in neurology), a less-experienced neurologist (D.D. with DD years of experience in neurology), an experienced radiologist (E.E. with EE years of experience in neuroradiology) and a less-experienced radiologist (F.F. with FF years of experience in neuroradiology). Readers were blinded to the reference grade. For analysis, readers were considered individually and pooled into three groups: neurologists, radiologists, and all readers. A washout period of at least 2 weeks was used to reduce recall bias.

Statistical Analysis

The primary endpoint was agreement with the reference standard. The primary metric was quadratic-weighted Cohen kappa (wk), which penalizes larger ordinal grade discrepancies more heavily and is therefore well suited to this task. Secondary metrics were exact six-level classification accuracy, mean absolute ordinal error (MAE), and, at the binary reperfusion-success threshold (mTICI ≥2b), sensitivity and specificity. Aggregated reader groups pooled their case-level grades against the reference. Cases with missing predictions were excluded from the analysis for the respective model, as reported in Table 1.

Table 1
Agreement with the reference standard by rater and input condition.

Rater	Input condition	n	<i>wk</i> [95% CI]	Exact Accuracy, %	MAE [95% CI]	Sensitivity, % [95% CI]	Specificity, % [95% CI]
GPT-5	Report only	90	0.864 [0.781, 0.919]	66.7 [56.4, 75.5]	0.41 [0.28, 0.54]	98.6 [92.7, 99.8]	81.2 [57.0, 93.4]
	Images only	90	0.345 [0.082, 0.555]	44.4 [34.6, 54.7]	1.01 [0.79, 1.23]	100.0 [95.1, 100.0]	18.8 [6.6, 43.0]
	Images + report	90	0.868 [0.781, 0.921]	67.8 [57.6, 76.5]	0.39 [0.27, 0.51]	100.0 [95.1, 100.0]	62.5 [38.6, 81.5]
MedGemma-1.5	Report only	89	0.336 [0.162, 0.527]	47.2 [37.2, 57.5]	0.87 [0.66, 1.08]	98.6 [92.6, 99.8]	6.2 [1.1, 28.3]
	Images only	79	0.017 [-0.087, 0.107]	24.1 [16.0, 34.5]	1.10 [0.87, 1.33]	100.0 [94.3, 100.0]	0.0 [0.0, 20.4]
	Images + report	86	0.337 [0.176, 0.512]	47.7 [37.4, 58.1]	0.94 [0.72, 1.19]	97.1 [90.2, 99.2]	6.2 [1.1, 28.3]
Neurologists	Report only	180	0.843 [0.766, 0.893]	63.9 [56.1, 71.7]	0.44 [0.34, 0.54]	98.0 [95.5, 100.0]	68.8 [47.2, 88.5]
	Images only	180	0.842 [0.753, 0.902]	63.3 [55.0, 71.7]	0.44 [0.33, 0.54]	95.9 [92.0, 99.3]	71.9 [50.0, 90.6]
	Images + report	180	0.841 [0.757, 0.894]	56.1 [47.8, 64.4]	0.50 [0.39, 0.60]	95.9 [92.4, 99.3]	71.9 [50.0, 90.6]
Radiologists	Report only	180	0.901 [0.852, 0.935]	71.7 [65.0, 78.9]	0.33 [0.24, 0.42]	97.3 [94.5, 99.4]	81.2 [65.4, 94.5]
	Images only	180	0.865 [0.795, 0.913]	62.8 [54.4, 70.6]	0.43 [0.33, 0.54]	95.3 [91.3, 98.7]	81.2 [62.5, 96.7]
	Images + report	180	0.902 [0.846, 0.958]	68.9 [61.7, 76.1]	0.34 [0.26, 0.42]	95.9 [93.0, 98.7]	81.2 [62.5, 97.1]

Note.—Values are point estimates with 95% CIs in square brackets. Abbreviations: *wk*, quadratic-weighted Cohen kappa; MAE, mean absolute ordinal error in grade steps. Sensitivity and specificity are reported at the reperfusion-success threshold (mTICI ≥ 2 b). *n* is the number of ratings (cases $\times$ readers) with both a reference and a non-missing grade. All metrics were computed on gradable cases only. Note that the MedGemma-1.5 model returned non-gradable outputs in a subset of cases.

Rater	Input condition	n	wk [95% CI]	Exact Accuracy, %	MAE [95% CI]	Sensitivity, % [95% CI]	Specificity, % [95% CI]
			0.937]		0.42]		
All readers	Report only	360	0.873 [0.812, 0.911]	67.8 [61.1, 74.4]	0.39 [0.30, 0.47]	97.6 [95.6, 99.3]	75.0 [57.8, 90.6]
	Images only	360	0.854 [0.783, 0.903]	63.1 [56.1, 70.0]	0.44 [0.35, 0.53]	95.6 [92.8, 98.1]	76.6 [57.7, 93.2]
	Images + report	360	0.872 [0.808, 0.911]	62.5 [55.8, 69.7]	0.42 [0.34, 0.49]	95.9 [93.2, 98.4]	76.6 [57.8, 93.2]
Note.—Values are point estimates with 95% CIs in square brackets. Abbreviations: wk, quadratic-weighted Cohen kappa; MAE, mean absolute ordinal error in grade steps. Sensitivity and specificity are reported at the reperfusion-success threshold (mTICI $\geq$2b). <i>n</i> is the number of ratings (cases $\times$ readers) with both a reference and a non-missing grade. All metrics were computed on gradable cases only. Note that the MedGemma-1.5 model returned non-gradable outputs in a subset of cases.							

All estimates are reported as point values with 95% confidence intervals (CIs), and a contrast was considered informative when its CI excluded zero. For individual raters (both models, each of the four readers), CIs for wk and MAE came from a case-resampled percentile bootstrap (1,000 resamples) with a common resample index across arrays so that paired contrasts retained their case-level pairing. Sensitivity, specificity, and exact accuracy CIs used the Wilson score method. For aggregated reader groups, where each case contributes one rating per member, all CIs came from a case-cluster bootstrap resampling whole cases. Contrasts were bootstrapped on shared cases using the same resample for both terms.

Two within-rater input-sensitivity contrasts quantified the value of each information source on shared cases: the report increment (wk with images plus report minus wk with images only, representing the gain from adding the report) and the imaging increment (wk with images plus report minus wk with report only, representing the gain from adding the images). The model-versus-reader gap was the paired wk difference (model minus reader or reader group) within each condition.

Inter-reader agreement was summarized descriptively as the pairwise quadratic-weighted kappa within each specialty pair and the mean pairwise kappa across all four readers, together with the exact four-way agreement rate. Analyses were performed in Python using the numpy v2.2.2, pandas v2.2.3, scikit-learn v1.6.1, matplotlib v3.10.0, and seaborn v0.13.2 libraries.

Results

General Results

Ninety of the initially screened thrombectomy cases met the eligibility criteria. Occlusions were left-sided in 48 of 90 cases (53.3%) and right-sided in 42 (46.7%), with occlusion sites distributed across M1 (n = 48, 53.3%), ICA (n = 28, 31.1%), M2 (n = 13, 14.4%), and M3 (n = 1, 1.1%). The reference mTICI distribution was skewed toward higher grades: mTICI 3 in 44.4% (40/90), 2c in 20.0% (18/90), 2b in 17.8% (16/90), 2a in 10.0% (9/90), 1 in 4.4% (4/90), and 0 in 3.3% (3/90). Successful reperfusion (mTICI $\geq$ 2b) was present in 82.2% (74/90) of cases. Binary performance estimates should therefore be interpreted in the context of substantial class imbalance. Grade distributions for the reference standard and for each rater-by-condition combination are shown in Fig. 3.

Agreement With the Reference Standard

When a procedural report was available, GPT-5 graded mTICI at a level comparable to the pooled readers. Quadratic-weighted κ was 0.87 with images plus report, 0.86 with report only, and 0.87 for all readers under the same conditions. When restricted to images only, GPT-5 agreement fell sharply to 0.35. MedGemma-1.5 performed poorly in all settings, with wk values of 0.34 with images plus report, 0.34 with report only, and 0.02 with images only. It also produced non-gradable outputs in 4, 1, and 11 cases, respectively.

Among the reader groups, radiologists showed the highest agreement overall, with wk of 0.90 with a report and 0.87 on images only, and neurologists performed similarly and remained stable across conditions, with wk of 0.84 in all three tasks. Detailed estimates are given in Table 1 and visualized in Fig. 4. Individual-reader estimates are reported in Supplementary Table 1.

Exact six-level accuracy and mean absolute ordinal error followed the same pattern. With a report, GPT-5 achieved exact accuracy of 66.7% to 67.8% and a MAE of about 0.4 grade steps, similar to the pooled readers, whose exact accuracy ranged from 62.5% to 67.8% with a MAE of about 0.4. By contrast, GPT-5 on images only fell to 44.4% exact accuracy with a MAE of 1.01 grade steps, and MedGemma-1.5 on images only fell to 24.1% with a MAE of 1.10. Even with a report available, MedGemma-1.5 remained suboptimal, with exact accuracy of 47.2% to 47.7% and a MAE of 0.87 to 0.94 grade steps.

At the binary reperfusion-success threshold, sensitivity was uniformly high, with most rater-by-condition cells at or above 95% and several at 100%, reflecting the high prevalence of successful reperfusion. Specificity, however, separated raters clearly. GPT-5 reached 81.2% specificity with the report only but only 18.8% on images only. MedGemma-1.5 identified almost no unsuccessful case, with specificity of 0.0% to 6.2%, effectively labeling nearly every procedure as successful. Reader groups retained acceptable specificity in every condition: radiologists reached 81.2% throughout, and all readers combined reached 75.0% to 76.6%.

Input Sensitivity

The two information sources, images and report text, contributed very differently to the models and to human readers (Table 2; Fig. 5). For the models, essentially all usable performance came from the report. Adding the report on top of images increased wk by + 0.52 [0.33, 0.75] for GPT-5 and + 0.33 [0.14, 0.54] for MedGemma-1.5, with both intervals excluding zero, whereas adding images on top of the report changed wk negligibly for GPT-5 (+ 0.00 [-0.05, 0.06]) and for MedGemma-1.5 (+ 0.01 [-0.10, 0.13]). For the readers, the pattern was reversed: neither increment materially changed agreement, indicating that they graded almost equally well from images alone and from the report alone. The only reader contrast whose interval excluded zero was a small report increment for radiologists (+ 0.04 [0.01, 0.07]).

Table 2
Input-sensitivity contrasts.

Rater	Report increment, Δwk [95% CI]	Imaging increment, Δwk [95% CI]
GPT-5	+ 0.522 [0.325, 0.753]	+ 0.004 [-0.053, 0.056]
MedGemma-1.5	+ 0.329 [0.137, 0.540]	+ 0.011 [-0.095, 0.126]
Neurologists	0.000 [-0.033, 0.032]	-0.002 [-0.053, 0.050]
Radiologists	+ 0.037 [0.010, 0.069]	+ 0.001 [-0.027, 0.029]
All readers	+ 0.018 [-0.004, 0.042]	-0.001 [-0.032, 0.028]
Note.—Change in quadratic-weighted kappa (wk) when one information source was added. The report increment compares images plus report with images only and the imaging increment compares images plus report with report only. Values are point estimates with 95% CIs in square brackets.		

Model-Versus-Reader Comparison

When a report was available, GPT-5 performed essentially identically to the pooled readers. The paired wk gap between GPT-5 and all readers was 0.00 [-0.05, 0.04] with images plus report and – 0.01 [-0.06, 0.04] with report only. On images only, GPT-5 performed substantially worse than every reader group, with a gap of -0.51 [-0.74, -0.31] versus all readers. Across all three conditions, MedGemma-1.5 fell short of the pooled readers, with gaps in wk of -0.53 [-0.69, -0.36] for images plus report, -0.54 [-0.71, -0.35] for report only, and – 0.84 [-0.95, -0.71] for images only.

Inter-Reader Agreement

Pairwise quadratic-weighted agreement among the four readers was substantial in every condition, with a mean wk of 0.84 to 0.86 across all readers, and was consistently higher within the radiology pair (0.85 to 0.89) than within the neurology pair (0.78 to 0.82) (Table 3; Supplementary Fig. 1). Exact four-way agreement on the full six-level scale was achieved in only about one third of cases (33.3% to 38.9%), indicating that even concordant readers often differed by adjacent grades.

Table 3
Inter-reader agreement by input condition.

Input condition	Neurology pair, <i>WK</i>	Radiology pair, <i>WK</i>	All readers, mean <i>WK</i>	Exact 4-way agreement, %
Images + report	0.823 [0.714, 0.893]	0.887 [0.807, 0.934]	0.857 [0.775, 0.906]	33.3 [24.5, 43.6]
Report only	0.778 [0.597, 0.878]	0.853 [0.756, 0.913]	0.837 [0.736, 0.896]	38.9 [29.5, 49.2]
Images only	0.808 [0.691, 0.884]	0.889 [0.815, 0.933]	0.845 [0.770, 0.897]	35.6 [26.4, 45.9]
Note.—Pairwise quadratic-weighted kappa within each specialty pair, the mean pairwise kappa across all four readers, and the exact four-way agreement rate across the six reader pairs on the six-level scale are reported. 95% CI in square brackets.				

Discussion

We compared neurologists and neuroradiologists with contemporary LMMs in grading post-thrombectomy reperfusion on the mTICI scale across three input conditions. The main finding was that GPT-5 achieved performance comparable to the readers when grading was based on the written procedural report. Adding angiographic images did not improve agreement, whereas image-only performance was substantially lower than that of the physician readers. MedGemma-1.5 performed below the readers in all three input conditions. Taken together, model performance was driven predominantly by textual information rather than meaningful multimodal integration.

Our finding that GPT-5 performed comparably to physicians on mTICI assessment from textual descriptions is consistent with the growing literature showing strong medical text-processing capabilities of contemporary LLMs and LMMs [20–21]. In radiology, they demonstrated free-text report structuring and interpretation at human-level accuracy [27–28]. The present results extend those observations to neurointerventional stroke care, suggesting that the information required for mTICI classification is often encoded in procedural reports with sufficient explicitness to support accurate interpretation by both human readers and AI systems. This capability may be useful for automated data extraction, registry population, retrospective research, and quality assurance.

Comparable performance in text-based assessment does not imply true understanding of angiographic findings by the models. Rather, it highlights the ability of modern language models to recognize and map linguistic patterns associated with reperfusion grading. Because thrombectomy reports often contain standardized descriptions of residual occlusion, distal branch filling, and reperfusion extent, the models may simply translate these textual cues into established mTICI categories. In a setting where operator overestimation and fatigue-related inaccuracy are recognized concerns, AI-based extraction of reperfusion grades from procedural reports would avoid fatigue and procedural pressure. However, it

may still inherit systematic biases embedded in the documentation, including operator-related overestimation of reperfusion [15, 17, 19].

MedGemma-1.5's limited performance should be read against its intended purpose. The 4-billion-parameter model is built for local, on-premise use on modest hardware, which is attractive for data protection, cost, and reproducibility in clinical research. Its weaknesses here follow from the same design: medical pretraining centered on imaging modalities other than cerebral DSA leaves angiographic reperfusion grading largely outside its adapted domain, and models of this size handle non-English clinical text less reliably, which likely compounded the difficulty of German procedural reports. Domain adaptation at this scale did not offset limited capacity for the present task.

By contrast, human readers clearly outperformed the evaluated AI models when only angiographic image data were included, underscoring the continued importance of visual expertise in neurointerventional imaging. Accurate mTICI grading requires integrating complex spatial and temporal information such as distal branch opacification, collateral filling, delayed perfusion, and the extent of reperfusion across the affected vascular territory, all of which experienced clinicians synthesize using years of domain-specific training. Although multimodal AI systems have made substantial progress in medical imaging, current general-purpose models remain primarily language optimized and appear far less capable of specialized angiographic assessment [22–28]. This is consistent with recent evaluations of general-purpose vision-language models in other diagnostic imaging tasks, which also found only modest standalone image-based accuracy compared with domain experts [24–25]. The advantage of human readers observed here supports the view that current AI systems should be considered complementary tools rather than replacements for expert interpretation in neurointerventional practice.

From a clinical perspective, these results suggest a pragmatic division of labor. Large language models may assist with interpretation of procedural documentation, automated extraction of outcome parameters, and standardization of reporting workflows. Final assessment of imaging, however, should continue to rely on expert review of the image data, particularly in cases where subtle distinctions may affect clinical evaluation, research outcomes, or quality metrics.

Several limitations warrant acknowledgment. First, the sample size and number of readers were modest, which limits subgroup analyses and, together with the skewed distribution, makes estimates of sensitivity and specificity less precise because they are based on only 16 negative cases. Second, this was a single-center study with standardized local reporting conventions and thrombectomy protocols, and the use of German-language reports may have contributed to the limited performance of MedGemma-1.5. Therefore, generalizability to other centers and report formats remains uncertain. Third, only a limited selection of LMMs was evaluated. GPT-5 and MedGemma-1.5 likely bracket the performance range, with GPT-5 near the upper end and MedGemma-1.5 toward the lower end. Given the rapid pace of model development, these results may not reflect currently available or near-future systems. Fourth, image-based assessment relied on static angiographic screenshots rather than full DSA series, which may underestimate performance and does not fully capture the potential of

multimodal AI in routine practice. The high performance of human readers, substantially different from the models, still shows that accurate mTICI assessment is feasible using static images.

Conclusion

A frontier general-purpose model demonstrated performance comparable to physicians when classifying mTICI scores from written thrombectomy reports. When only angiographic image data were provided, human readers retained a clear advantage. These findings highlight both the considerable potential and the current limitations of large multimodal models in neurointerventional stroke care: AI may support documentation, reporting, and data extraction, but expert human interpretation remains indispensable for accurate image-based assessment of reperfusion outcomes after mechanical thrombectomy.

Abbreviations

AI

Artificial intelligence

API

Application programming interface

CI

Confidence interval

DSA

Digital subtraction angiography

GPT

Generative pre-trained transformer

ICA

Internal carotid artery

JSON

JavaScript Object Notation

LLM

Large language model

LMM

Large multimodal model

LVO

Large vessel occlusion

MAE

Mean absolute ordinal error

M1

M1 segment of the middle cerebral artery

M2

M2 segment of the middle cerebral artery

M3

M3 segment of the middle cerebral artery
mTICI
Modified Thrombolysis in Cerebral Infarction
REFINE
Reporting foundation and large language models in medical research
RCT
Randomized controlled trial
WK
Quadratic-weighted Cohen's kappa

Declarations

Ethics approval and consent to participate

This retrospective, single-center study was approved by the institutional review board of the University Hospital Regensburg (vote no. 26-4706-104, clinical trial number: not applicable). The requirement for informed consent was waived owing to the retrospective study design and the use of anonymized imaging data. The study was conducted in accordance with the Declaration of Helsinki and its later amendments, and the REFINE checklist for reporting foundation and large language models in medical research was followed during study design and manuscript preparation.

Consent for publication

Not applicable.

Availability of data and materials

All data of this study is included in this manuscript.

Competing Interests

All authors do not have competing interests regarding this study.

Funding

No funding was received for this study

Authors' contributions

Michael Knott and Quirin D. Strotzer contributed to the conceptualization of the study, data analysis, and manuscript preparation. Gerardo Napodano, Ekaterina Noeva, Liang Zhang, Moritz Brandenstein, Johannes Dinkel, Lili-Laura Bruta, Julia Pistorius, David Olmes, Isabel Wiesinger, Christian Stroszczynski, and Christina Wendl contributed to data collection. Christian Stroszczynski and Christina Wendl provided supervision. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors would like to thank all colleagues, patients and relatives involved in the clinical care of the patients.

References

1. GBD 2021 Stroke Risk Factors Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* 2024;23:973–1003. 10.1016/S1474-4422(24)00369-7.
2. Berkhemer OA, Fransen PSS, Beumer D, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med.* 2015;372:11–20. 10.1056/NEJMoa1411587.
3. Goyal M, Demchuk AM, Menon BK, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med.* 2015;372:1019–30. 10.1056/NEJMoa1414905.
4. Campbell BCV, Mitchell PJ, Kleinig TJ, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med.* 2015;372:1009–18. 10.1056/NEJMoa1414792.
5. 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;372:2285–95. 10.1056/NEJMoa1415061.
6. Jovin TG, Chamorro A, Cobo E, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med.* 2015;372:2296–306. 10.1056/NEJMoa1503780.
7. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med.* 2018;378:11–21. 10.1056/NEJMoa1706442.
8. Albers GW, Marks MP, Kemp S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med.* 2018;378:708–18. 10.1056/NEJMoa1713973.
9. Sarraj A, Hassan AE, Abraham MG, et al. Trial of endovascular thrombectomy for large ischemic strokes. *N Engl J Med.* 2023;388:1259–71. 10.1056/NEJMoa2214403.
10. Huo X, Ma G, Tong X, et al. Trial of endovascular therapy for acute ischemic stroke with large infarct. *N Engl J Med.* 2023;388:1272–83. 10.1056/NEJMoa2213379.
11. Suh SH, Cloft HJ, Fugate JE, et al. Clarifying differences among thrombolysis in cerebral infarction scale variants. *Stroke.* 2013;44:1166–8. 10.1161/STROKEAHA.111.000399.
12. Zaidat OO, Yoo AJ, Khatri P, et al. Recommendations on Angiographic Revascularization Grading Standards for Acute Ischemic Stroke: A Consensus Statement. *Stroke.* 2013;44(9). 10.1161/STROKEAHA.113.001972.
13. Goyal M, Fargen KM, Turk AS, et al. 2C or not 2C: defining an improved revascularization grading scale and the need for standardization of angiography outcomes in stroke trials. *J Neurointerv Surg.* 2014. 10.1136/neurintsurg-2013-010665.
14. Gaha M, Roy C, Estrade L, et al. Inter- and intraobserver agreement in scoring angiographic results of intra-arterial stroke therapy. *AJNR Am J Neuroradiol.* 2014;35:1163–9. 10.3174/ajnr.A3828.

15. Heiferman DM, Pecoraro NC, Wozniak A, et al. Reliability of the modified TICl score among endovascular neurosurgeons. *AJNR Am J Neuroradiol*. 2020;41:1441–6. 10.3174/ajnr.A6696.
16. Volný O, Cimflova P, Szeder V. Inter-rater reliability for thrombolysis in cerebral infarction with TICl 2c category. *J Stroke Cerebrovasc Dis*. 2017;26:461–5. 10.1016/j.jstrokecerebrovasdis.2016.11.008.
17. Mokin M, Hirsch JA, Fiorella D, Jadhav AP. Lies, damned lies, and TICl. *J Neurointerv Surg*. 2021;13:1070–3. 10.1136/neurintsurg-2021-018072.
18. Findler M, Turjman A, Raymond J, et al. Interobserver agreement in scoring angiographic results of basilar artery occlusion stroke therapy. *AJNR Am J Neuroradiol*. 2021;42:1458–63. 10.3174/ajnr.A7182.
19. Pressman E, Waqas M, Sands V, et al. Factors associated with decreased accuracy of mTICl scoring among neurointerventionalists during thrombectomy. *Stroke*. 2021;52:2335–41. 10.1161/STROKEAHA.120.033372.
20. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models (Med-PaLM 2). *Nat Med*. 2025;31:943–58. 10.1038/s41591-024-03423-7.
21. Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. *Science*. 2026;392:524–7. 10.1126/science.adz4433.
22. Moor M, Banerjee O, Abad ZFH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259–65. 10.1038/s41586-023-05881-4.
23. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *NEJM AI*. 2024;1:Aloa2300138. 10.1056/Aloa2300138.
24. Altintas Mese C, Mese I, Akinci D, Antonoli T. Zero-shot performance of a general-purpose vision-language model for pediatric appendicitis diagnosis. *Pediatr Radiol* 2026 Jun 8. 10.1007/s00247-026-06687-y
25. Schramm S, Preis S, Metz MC, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. *Radiology*. 2025;314:e240689. 10.1148/radiol.240689.
26. Strotzer QD, Nieberle F, Kupke LS, et al. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. *Radiology*. 2024;313:e240955. 10.1148/radiol.240955.
27. Woźnicki P, Laqua C, Fiku I, et al. Automatic structuring of radiology reports with on-premise open-source large language models. *Eur Radiol*. 2024;34:8605–15. 10.1007/s00330-024-11074-y.
28. Mallio CA, Sertorio AC, Bernetti C, et al. Large language models for structured reporting in radiology: performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol Med*. 2023;128:1471–9. 10.1007/s11547-023-01651-4.
29. Lehnen NC, Dorn F, Wiest IC, et al. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. *Radiology*. 2024;311:e232741. 10.1148/radiol.232741.

Figures

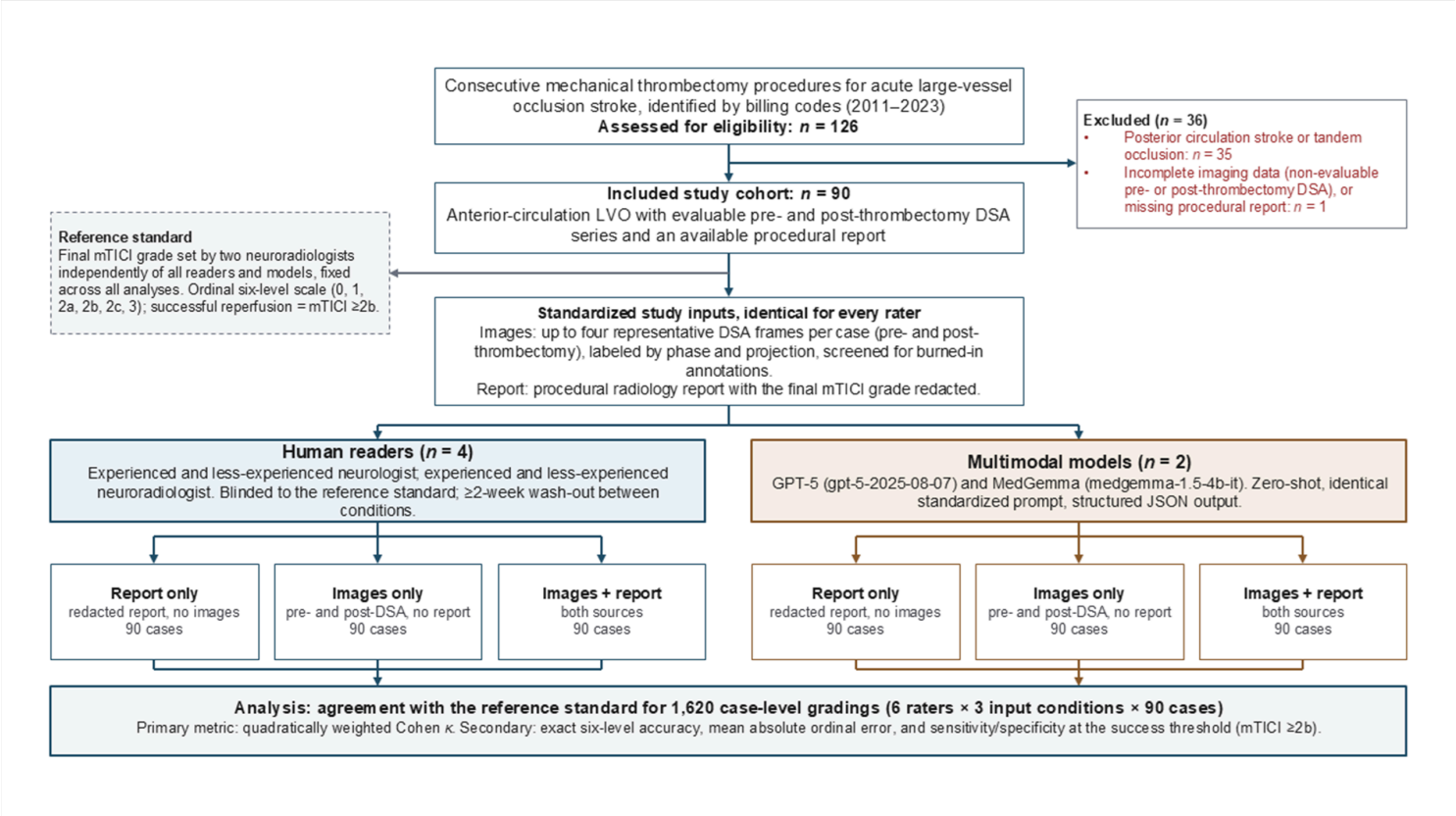


Figure 1

Study Flow Diagram. Inclusion/exclusion criteria and study setup are illustrated

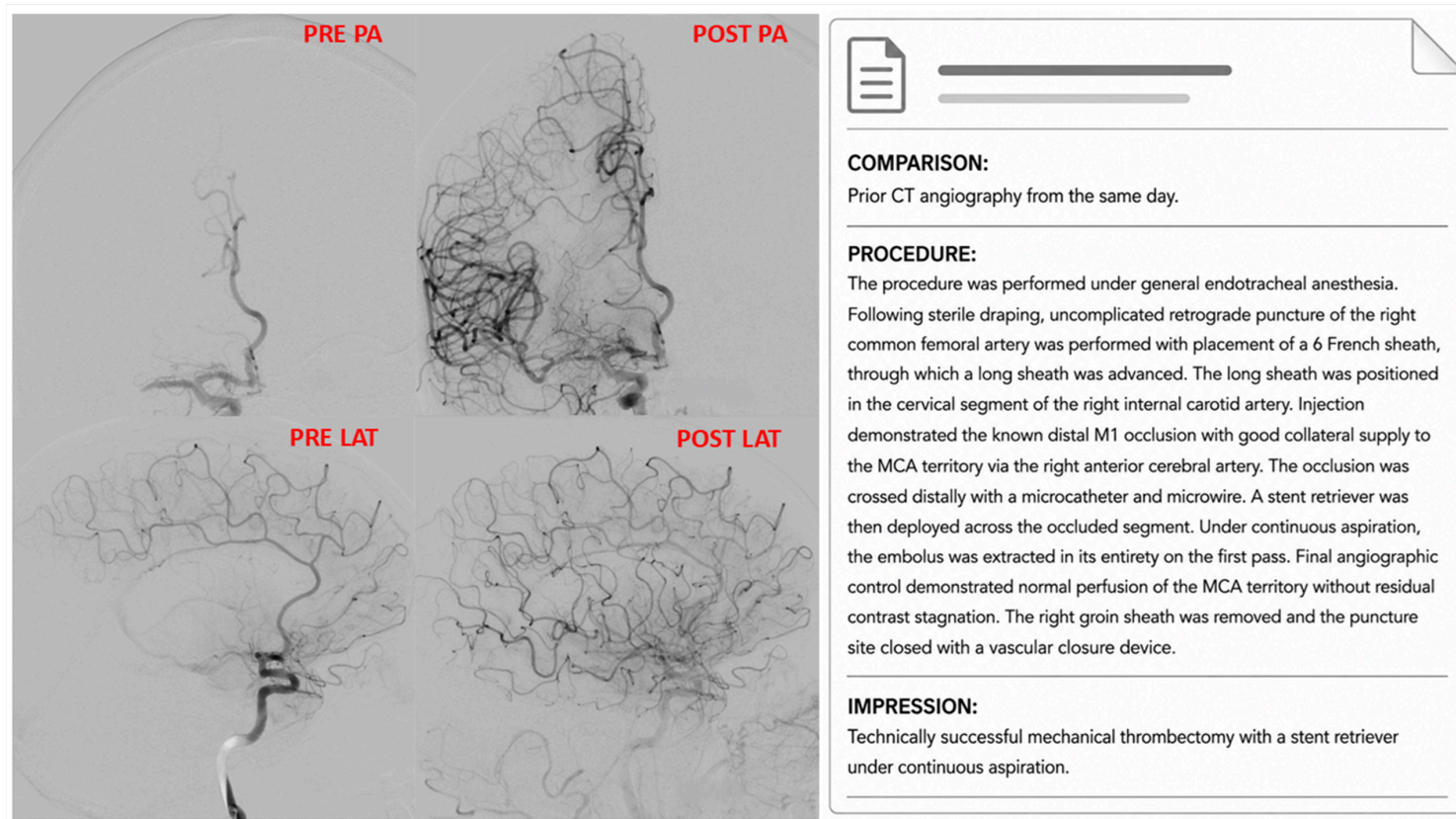


Figure 2

Example case input material. Left: digital subtraction angiography of a right distal M1 occlusion in posteroanterior (PA, top) and lateral (LAT, bottom) projections, before (left column) and after (right column) mechanical thrombectomy. Right: the corresponding procedure report as presented to raters and models, with the mTICI grade removed.

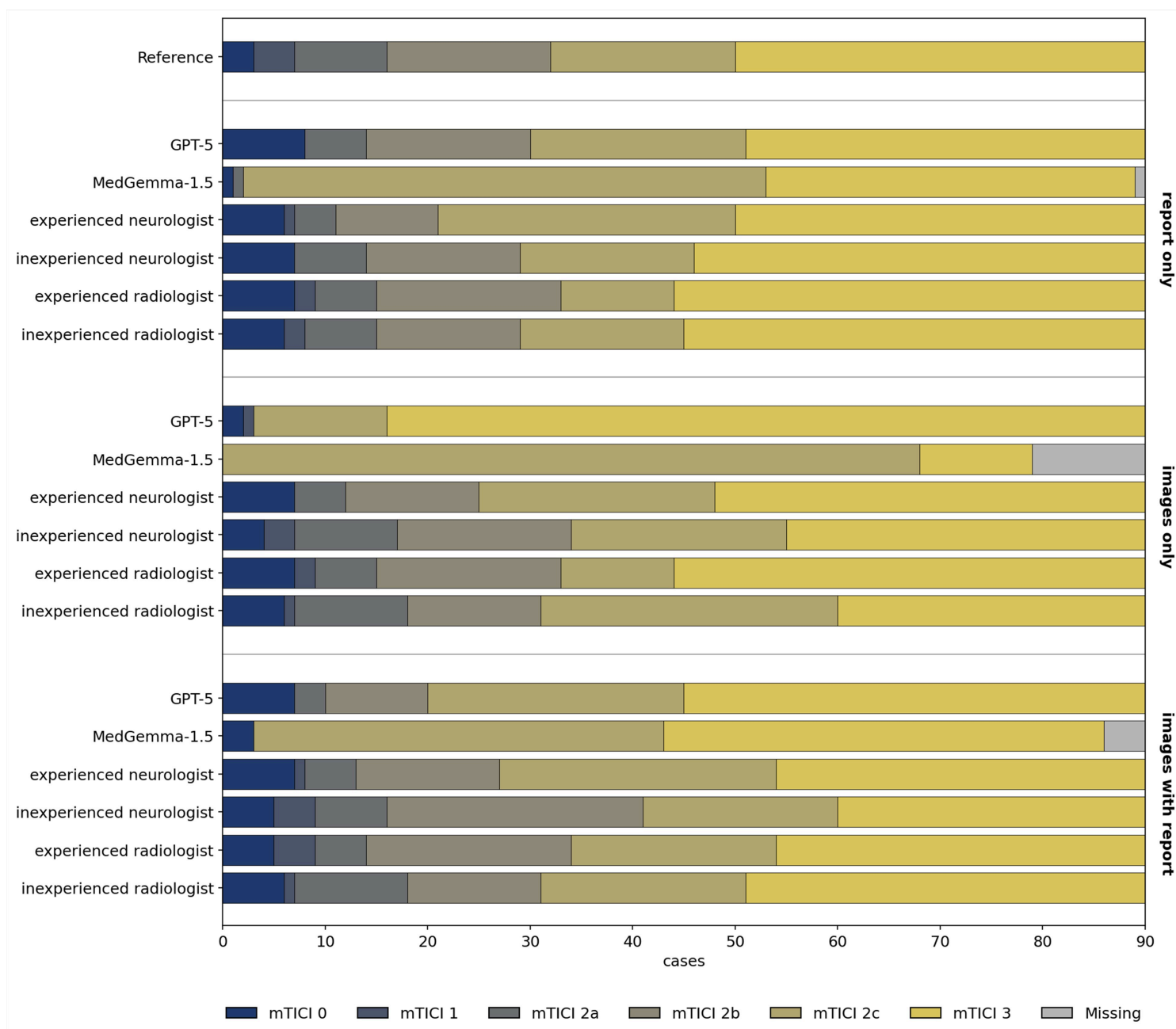


Figure 3

mTICI grade distributions. Distribution of assigned mTICI grades for the reference standard and for each rater under each input condition. The 'Missing' segment denotes non-gradable model outputs (only the MedGemma-1.5 model provided non-gradable outputs).

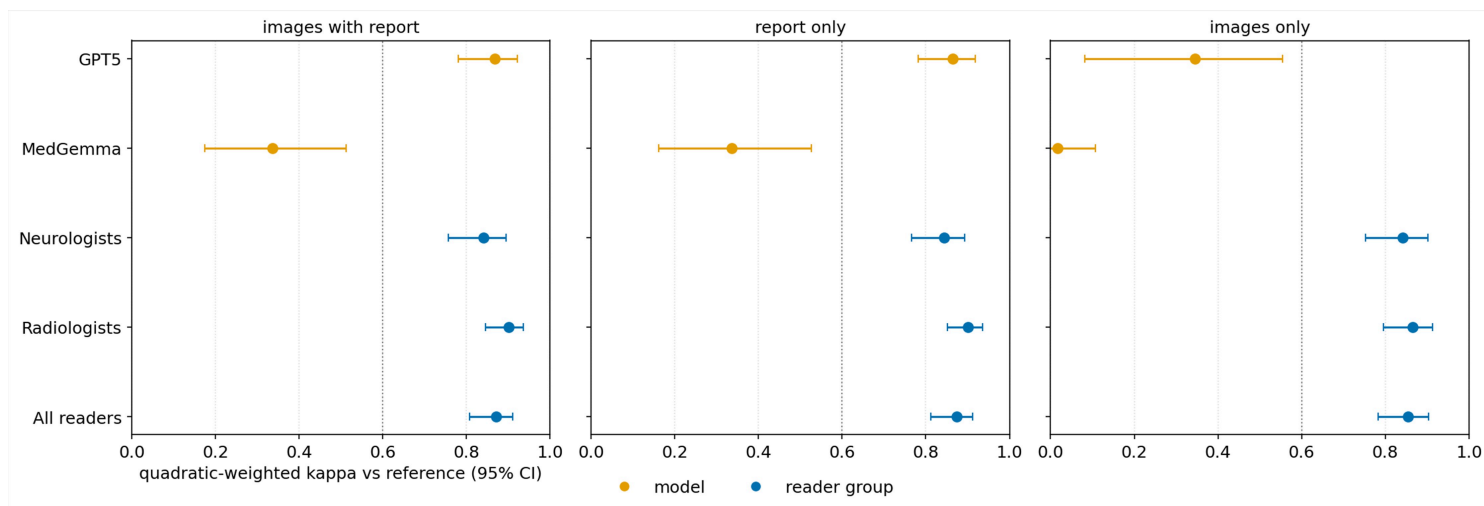


Figure 4

Agreement with the reference standard by rater and input condition. Quadratic-weighted Cohen kappa (point estimate and 95% CI) of each model and reader group versus the reference mTICI grade, by input condition. The dotted line marks $wk = 0.60$.

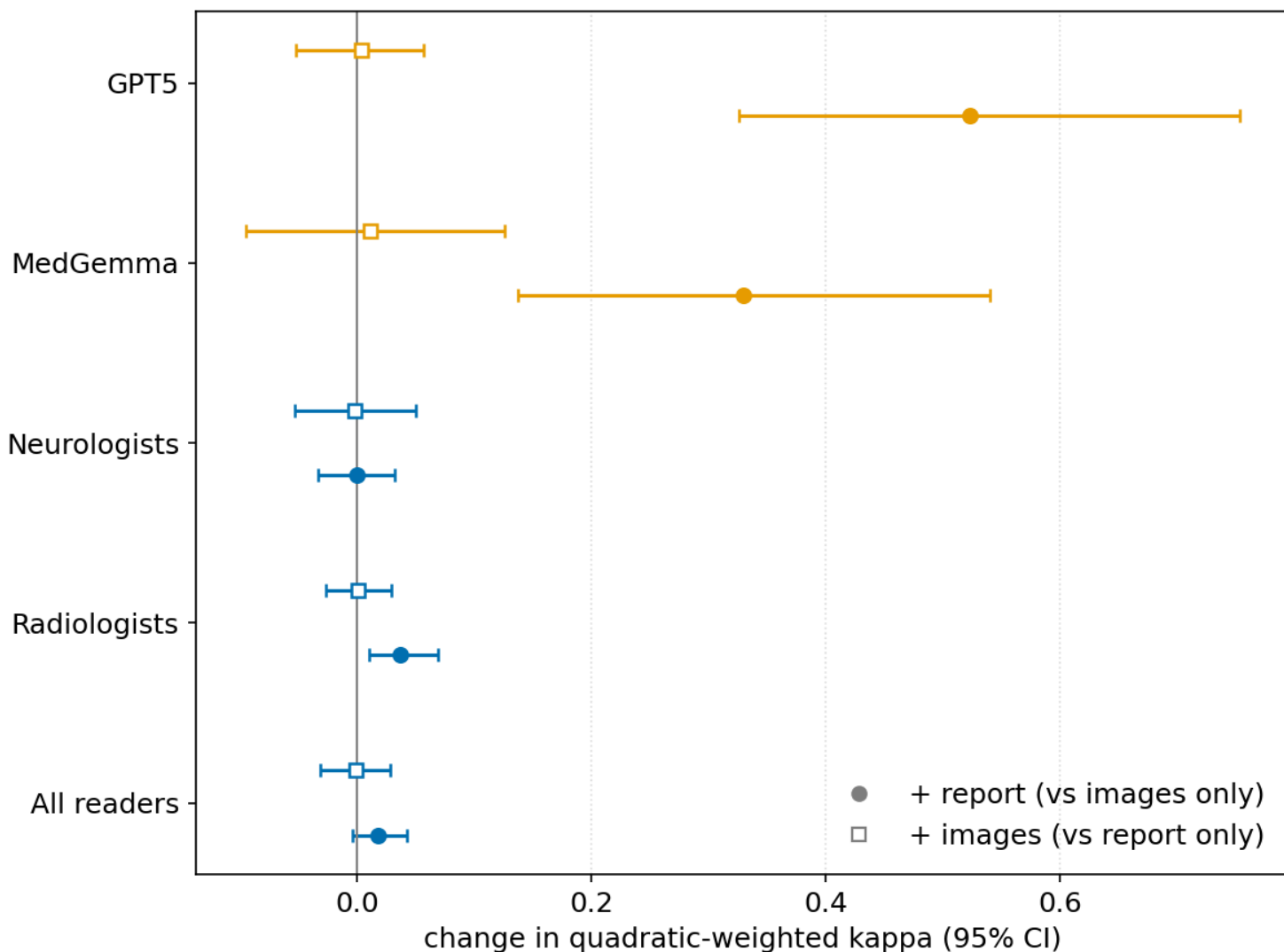


Figure 5

Input sensitivity. Change in quadratic-weighted kappa (95% CI) when one information source is added: filled circles represent report increment (images plus report versus images only) and open squares represent imaging increment (images plus report versus report only).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [mTICISupplement.docx](#)