EJNMMI Research

ORIGINAL RESEARCH

Open Access

Can chatGPT-4o reliably standardize PSMA PET/CT and PET/MRI reports using PROMISE V2 criteria? - An exploratory study

Anna Hinterberger[1,2], Maurin H. Mangold[1,3,4], Caroline Weigel[1,2], Henri Hartmann[1,2], Dominik Nörenberg[5], Matthias F. Froelich[5], Ricarda Ebner[6], Caelán Max Haney-Aubert[4], Karl-Friedrich Kowalewski[1,3,4], Stefan O. Schönberg[5] and Freba Grawe[1,2,5,7*]

Abstract

Background Structured reporting standardizes and facilitates reporting, improves accurate communication, and ultimately clinical decision-making. Although standardized frameworks such as PROMISE criteria are available for prostate-specific membrane antigen positron emission tomography (PSMA PET) for prostate cancer patients, free-text reporting remains predominant in both clinical routine and trials. Large language models (LLMs) may enable low-effort, time-efficient extraction of structured classifications from narrative reports. This study evaluated the performance of ChatGPT-4o for extracting PROMISE V2-based classifications from unstructured PSMA-PET/CT and PET/MRI reports.

Results For PSMA-PET/CT, overall miTNM accuracy was 79.8%, whereas PSMA-PET/MRI achieved a significantly higher accuracy of 91.0% (OR = 2.80, 95% CI: 1.32–6.51, $p = 0.003$). Component-wise, PET/MRI outperformed PET/CT in T-stage classification (83.8% vs. 57.7%; OR = 3.83, 95% CI: 1.34–12.69, $p = 0.006$) and demonstrated numerically higher N-stage classification accuracy (100% vs. 85.9%, $p = 0.014$), while M-stage classification was comparable between modalities (89.1% vs. 95.7%; OR = 0.84, 95% CI: 0.20–4.19, $p = 0.748$). PRIMARY score accuracy was also comparable for PET/CT and PET/MRI (70.4% vs. 88.1%; OR = 0.43, 95% CI: 0.05–2.14, $p = 0.315$). ChatGPT-4o's rationale for classifications was rated highly plausible across modalities, with a minimum Likert score of ≥ 4.8 for miTNM and 4.1 for PRIMARY.

Conclusion ChatGPT-4o enables reliable extraction of PROMISE V2–based N- and M-stage classifications from free-text PSMA-PET reports, with limited accuracy for T-stage. This work provides a first step toward leveraging LLMs to support structured and efficient reporting in PSMA PET imaging and points out present limitations.

Keywords PROMISE, ChatGPT4o, miTNM, Prostate cancer

*Correspondence:
Freba Grawe
freba.grawe@dkfz-heidelberg.de
[1]DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Heidelberg, Germany
[2]Junior Clinical Cooperation Unit Translational Molecular Imaging in Oncologic Therapy Monitoring (E310), German Cancer Research Center, Heidelberg, Germany
[3]Junior Clinical Cooperation Unit Intelligent Systems and Robotics in Urology (ISRU), German Cancer Research Center, Heidelberg, Germany
[4]Department of Urology and Urologic Surgery, University Medical Centre Mannheim, University of Heidelberg, Mannheim, Germany
[5]Department of Radiology and Nuclear Medicine, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany
[6]Department of Radiology, LMU University Hospital, LMU Munich, Munich, Germany
[7]German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Large language models (LLMs) are advanced artificial intelligence systems trained on large-scale text datasets to process and generate human language. Recent LLM-based models, such as the Generative Pre-Trained Transformer (GPT) developed by OpenAI, have demonstrated expert-level performance in a wide range of language-based tasks across different domains [1, 2]. These capabilities have sparked increasing interest in radiology, particularly in the context of structured reporting, which is known to guide image interpretation, enhance reproducibility, and facilitate communication with referring clinicians [3–5]. However, the integration of LLMs into imaging workflows remains limited.

For the disease extent of prostate cancer, the PROMISE (Prostate Cancer Molecular Imaging Standardized Evaluation) framework, which was updated in 2023 (V2), was developed to standardize the interpretation of prostate-specific membrane antigen (PSMA) targeted positron emission tomography (PET) [6]. PROMISE provides a unified system for comparing imaging, biopsy, and biomarker findings. It forms the basis for the molecular imaging Tumour-Node-Metastasis (miTNM) classification, which is a molecular imaging-based analogue to the conventional TNM system for tumour staging and includes the PRIMARY score, which characterizes intraprostatic PSMA uptake patterns to guide local staging and biopsy targeting [7]. Since its introduction, PROMISE has been validated in multiple studies and is associated with improved diagnostic accuracy and consistency in staging and treatment planning for prostate cancer [8–10]. Moreover, the PROMISE criteria demonstrated their potential to enable precise risk stratification for overall survival prediction in both early and advanced stages of prostate cancer [11]. Despite these advantages, structured reporting frameworks such as PROMISE have not been implemented in routine clinical practice nor systematically in clinical trials. Free-text reporting remains predominant, and the transition to structured formats often requires training, time, and system-level changes [12, 13]. To overcome these challenges, LLMs may offer a scalable, low-effort solution by enabling real-time extraction of standardized classifications from narrative reports or by assisting in automated structured reporting [5, 14]. However, before clinical integration, such tools must be thoroughly validated for accuracy, reliability, and reproducibility in a medical context [15, 16].

In this study, we evaluated the feasibility of ChatGPT-4o, accessed via the ChatGPT interface, to extract PROMISE V2-based miTNM classifications and PRIMARY scores from unstructured PSMA-ligand PET/CT and PET/MRI reports. Given that MRI offers superior soft-tissue contrast for local staging of prostate cancer compared to CT, we further hypothesized that ChatGPT-4o would show improved performance when interpreting PET/MRI reports.

Methods

Study design and patient cohort

In this single-center retrospective study, 137 free-text PSMA-PET reports (written in German) from 137 consecutive patients with histologically confirmed prostate cancer were screened. Ten reports were used for prompt development and were excluded from further analysis. An additional 19 patients were excluded due to missing clinical information, resulting in a final study cohort of 108 patients. The cohort included 71 PSMA-PET/CT and 37 PSMA-PET/MRI scans performed between January 2021 and December 2024. Indications comprised initial staging, biochemical recurrence, and therapy monitoring. Ethics approval was obtained from the local institutional review board (EK II 2023-886-AF 11).

Imaging protocols

PET/CT and PET/MRI scans were acquired using standard institutional protocols.

PSMA PET/CT imaging

PET imaging was performed according to established institutional protocols according to current guidelines using [^{18}F]PSMA. Low-dose CT for attenuation correction was acquired with 10 mA, 120 kV, 512 × 512 matrix. After intravenous injection of contrast agent ($n = 33$) diagnostic CT scans of the neck, thorax, abdomen, and pelvis (100–190 mAs; 120 kV) were acquired. Patients received diagnostic CT scans without contrast enhancement ($n = 38$) in case of known allergic reactions to iodinated contrast agent, renal impairment/failure, or hyperthyroidism or because contrast enhanced CT was already performed up to two weeks prior PET/CT. All acquired PET/CT scans were analysed using a dedicated software package (syngo.via, Siemens Healthcare).

PSMA PET/MRI imaging

18 patients received [^{18}F]PSMA and 19 patients received [^{68}Ga]PSMA according to established institutional protocols. In parallel, MRI included T2-HASTE, DWI, and contrast-enhanced T1-VIBE sequences, which were acquired in all patients. PET data were acquired in 3D mode. The raw data was reconstructed iteratively and evaluated using the ROI technique (regions of interest) and determination of the SUV (Standardized Uptake Value).

PROMISE extraction via ChatGPT-4o prompting

All reports were manually extracted from the radiology information system (RIS) by medical staff including radiotracer type and dose, the body of the report text

and the conclusion of the report without patient related information (e.g., name, birth date). All PSMA-PET/CT and PET/MRI reports were additionally reviewed by one investigator to ensure completeness for quality proposes. Findings in the report were minimally adjusted to anonymize the report. A standardized prompt was iteratively developed based on ten representative raw text reports (which were not further included in the dataset) and optimized for accurate extraction of PROMISE V2-based classifications. The prompt was evaluated in both German and English to assess whether language-specific semantic interpretations or subtle, staging-relevant nuances in the German language influenced the output. Moreover, we performed repeated test runs on identical reports to investigate the stability of the web-application. No differences in the output were observed between the two languages and the multiple test-runs on identical reports, leading to the decision to continue with the web-application rather than switching to the Application Programming Interface (API). The web-application is an accessible, state-of-the-art commercial model and could be readily implemented in clinical practice without specialized hardware or technical infrastructure, offering a suitable option for our exploratory study. Every report was submitted to ChatGPT-4o as a raw text in addition to the prompt. Moreover, all reports were submitted to ChatGPT-4o in one day to ensure that changes on the version due to ChatGPT updates do not influence the output. To reflect routine clinical workflow, we deliberately chose a simplified staging approach. Therefore, both the expert reference reading and the ChatGPT-based extraction followed an AJCC-oriented hierarchical interpretation.

Prompt for miTNM classification

Given the following PSMA-PET report, classify findings using PROMISE V2 published by Robert Seifert et al. in 2023. Provide the PROMISE classification (only include the following categories for local tumour: miT0, miT2 (u and m) and miT3 (a and b), miT4, miTr; for intrapelvic nodes: miN0, miN1, miN2; for distant metastases: miM0, miM1 (a, b and c) a short rationale based on the report text why you chose this PROMISE classification. Only state the result and the rational. Only state the highest scores for each category T, N and M. Please report results in English.

Prompt for PRIMARY classification

"Given the following PSMA-PET report, classify findings using the table attached and a short rationale why you chose this PRIMARY score. Only state the result and the rational. Please report results in English."

PRIMARY score.

1: No dominant intraprostatic pattern on PSMA. Low grade activity, miT0.

2: Diffuse transition zone activity or symmetrical central zone activity that does not extend to the prostate margin on CT background, miT0.

3: Focal transition zone activity visually twice above background, miT2, miT3, or T4.

4: Focal peripheral zone activity (no minimum intensity), miT2, miT3, or T4.

5: Intense uptake (visual very high intensity or SUVmax > 12), miT2, miT3, or miT4.

For PRIMARY scoring, only patients without prior prostatectomy were included to ensure intraprostatic lesion assessment. Each report was then entered into the ChatGPT-4o interface (Open AI, San Francisco, CA, USA) together with the final version of the prompt via the official web application. ChatGPT-4o generated a structured output including the miTNM classification, PRIMARY score (if applicable), along with a free-text rationale for each decision. No patient-identifying information or external datasets were uploaded. All outputs were exported into Excel (Microsoft, version 16.96.1) for further analysis.

Reference standard and validation

For comparison of results between ChatGPT-4o, all written reports were classified according to PROMISE V2 for PRIMARY and miTNM by one radiologist with expertise in hybrid imaging and one urologist specialized (= human reading) in prostate cancer, with additional validation by a board-certified nuclear medicine physician. Figure 1 depicts the extraction and validation process. For further analysis, instances of underestimation and overestimation were evaluated. Underestimation was defined as ChatGPT-4o assigning a lower T, N, or M category than the human reader, whereas overestimation was defined as ChatGPT-4o assigning a higher T, N, or M category than the human reader.

Evaluation of model output

In addition to classification accuracy, the ChatGPT-4o generated rationales for each miTNM component and PRIMARY scoring, which was independently evaluated by two readers using a 5-point Likert scale (1 = poor, 5 = excellent) to assess plausibility and factual correctness.

Statistical analysis

For accuracy of ChatGPT-4os miTNM and PRIMARY classification the percentual correctness, based on the human reading, and 95% confidence interval (CI) was calculated. Accuracy was compared between imaging modalities (PET/CT vs. PET/MRI) and clinical indication (initial staging vs. recurrence) using the Fisher's

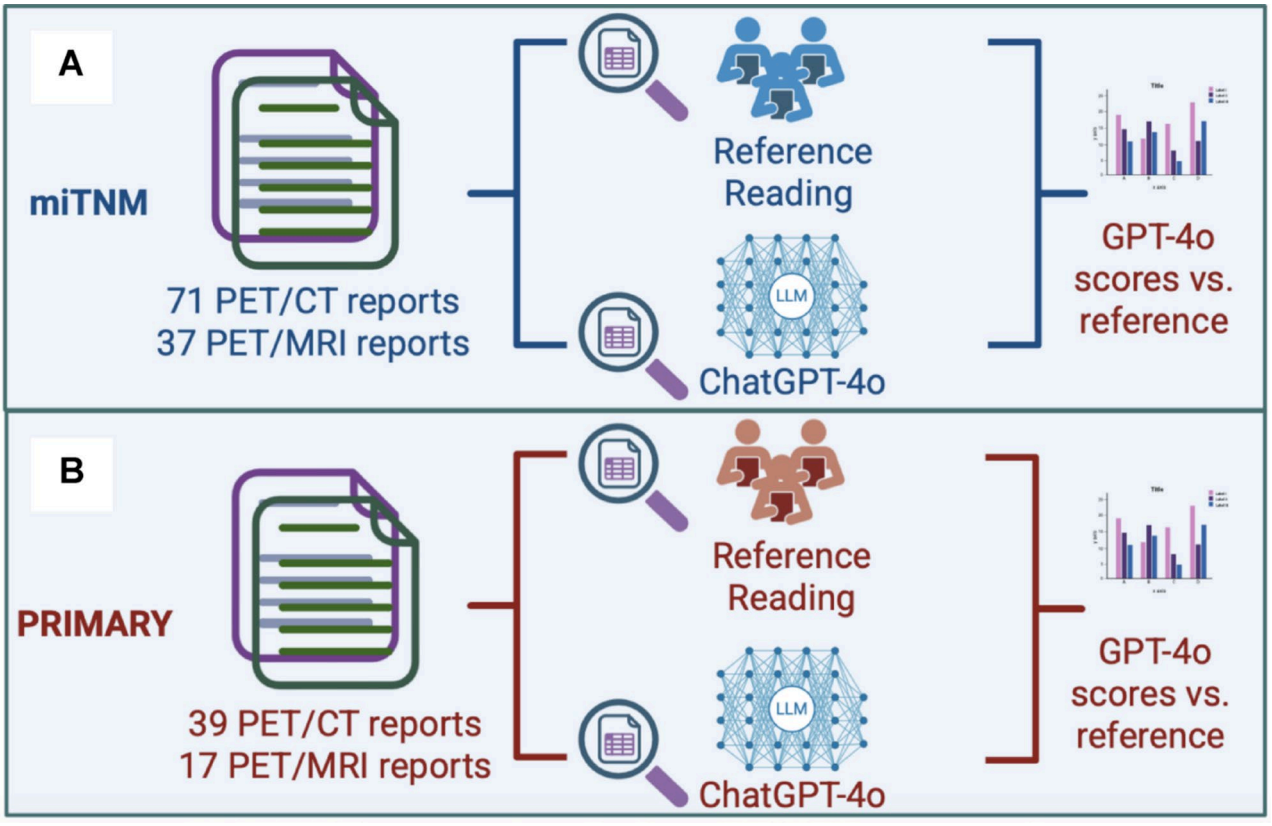


Fig. 1 Workflow-diagram. 108 PSMA-PET (71 PET/CTs and 37 PET/MRIs) written reports were analysed. For miTNM classification, all 108 reports were included (**A**). For PRIMARY classification, 56 reports of patients without prior prostatectomy were included (**B**). PRIMARY and miTNM was assessed based on the output of ChatGPT-4o and human reading, using the information provided in the free-text-reports

exact test. Likert ratings of ChatGPT-4o's rationales were obtained independently by two readers and averaged, results were calculated based on a two-tailed t-test. Data are presented as mean and standard deviation (SD). Cohens kappa was calculated for ChatGPT-4o and the validation set as well as between the two readers (M.M. and A.H.). Agreement was considered poor for $\kappa < 0.4$, fair for κ between 0.4 and 0.59, good for κ between 0.6 and 0.74, and excellent for $\kappa > 0.74$. Results were considered significant if $p < 0.05$. Statistical analyses were performed using RStudio (PositPBC, Boston, Massachusetts, US, Version 2024.12.1).

Results

Baseline patient characteristics

In total, 108 reports were included (71 PET/CT, 37 PET/MRI).

PROMISE classification extracted from written PET/CT reports

In 71 PET/CT reports, overall accuracy of miTNM classification by ChatGPT-4o from written reports compared to human readers was 79.8% (95%CI: 0.74–0.85). By component, percentage of correctly classified miTNM was 57.7% for miT (95%CI: 0.45–0.68), 85.9% for miN (95%CI: 0.77–0.92), and 95.7% for miM (95%CI: 0.90-1.00).

ChatGPT-4o *underestimated* 21 miT, 8 miN, and 3 miM categories and *overestimated* 9 miT, 2 miN. Most T-stage underestimations resulted from information that was not explicitly stated in the reports, nevertheless, these misclassifications were mostly minor shifts as T2m was often rated as T2u by ChatGPT-4o (see Table 2). We found that misclassification, particularly of miT2, was predominantly driven by information-limited discordances rather than true model-related errors. For further details, see Supplementary Table S1. N- and M-stage misclassifications similarly arose from subtle or insufficiently detailed descriptions that ChatGPT-4o did not capture (see Fig. 2). For miT, in 17 cases the most common source of error ($n = 23$) was that multifocal (m) or unifocal (u) tracer uptake was not explicitly described but only implied though phrases such as "tracer uptakes in projection of the prostate" or "tracer uptake in projection on the prostate." In five cases, ChatGPT-4o produced hallucinated information, and in two cases therapy-related details were not recognized because the report only stated "R1", without explicitly mentioning prostatectomy. For miN, eight cases were underestimated

Table 1 Summarizes patient characteristics

Patient characteristics	
mean age in years ± SD	70 ±7.7
mean PSA in ng/ml ± SD	11.1 ± 30.7
Gleason score (n, %)	
Gleason 6	6 (5.5)
Gleason 7	55 (50.9)
Gleason 8	17 (15.7)
Gleason 9	26 (24.2)
Gleason 10	4 (3.7)
Clinical/pathological TNM	
R0	10 (16.0)
T1	25 (22.6)
cT1	20
pT1	5
T2	38 (34.4)
cT2	21
pT2	17
T3	31 (23.4)
cT3	12
pT3	19
T4	4 (3.6)
cT4	2
pT4	2
N0	72 (69.3)
cN0	42
pN0	30
N1	32 (30.7)
cN1	20
pN1	12
M0	75 (69.5)
cM0	42
pM0	33
M1a	15 (13.8)
cM1a	9
pM1a	6
M1b	16 (14.9)
cM1b	9
pM1b	7
M1c	2 (1.8)
cM1c	2
pM1c	0
Clinical indications for PET imaging	
initial staging	56 (51.8)
biochemical recurrence	52 (48.2)
after prostatectomy	39(75.0)
after radiation therapy	7 (13.5)
after combination of both or other	6 (11.5)

Characteristics of patients undergoing PSMA-PET/CT and PSMA-PET/MRI between 2021 and 2024. SD: standard deviation; PSA: prostate specific antigen; T: tumour; N: nodes; M: metastases; R0: no residual tumour after therapy. Gleason scores were derived only from histopathology of TRUS-guided biopsy specimens

because ChatGPT-4o did not capture subtle wording that implied multiple lymph node metastases (e.g. "all the PET-positive lymph nodes in the small pelvis") or vague formulations such as "possible lymph node metastases". As a result, one N category was underestimated due to

Table 2 ChatGPT-4o vs. human validation reading for molecular imaging (mi) tumour (T) classification applied on PSMA PET/CT reports

PSMA PET/CT			**Incorrect predictions**								
Validation	**Total count**	**Correct prediction**	**Overestimations**					**Underestimations**			
			T2u	**T2m**	**T3a**	**T3b**	**T4**	**T0**	**T2u**	**T2m**	**T3a**	**T3b**
T0	5	**5**	0	0	0	0	0	0	0	0	0	0
Tr	3	**1**	**2**	0	0	0	0	0	0	0	0	0
T2u	17	**15**	0	**2**	0	0	0	0	0	0	0	0
T2m	40	**14**	0	0	**2**	**3**	0	0	**21**	0	0	0
T3a	0	na	na	na	na	na	na	na	na	na	na	na
T3b	3	**3**	0	0	0	0	0	0	0	0	0	0
T4	3	**3**	0	0	0	0	0	0	0	0	0	0

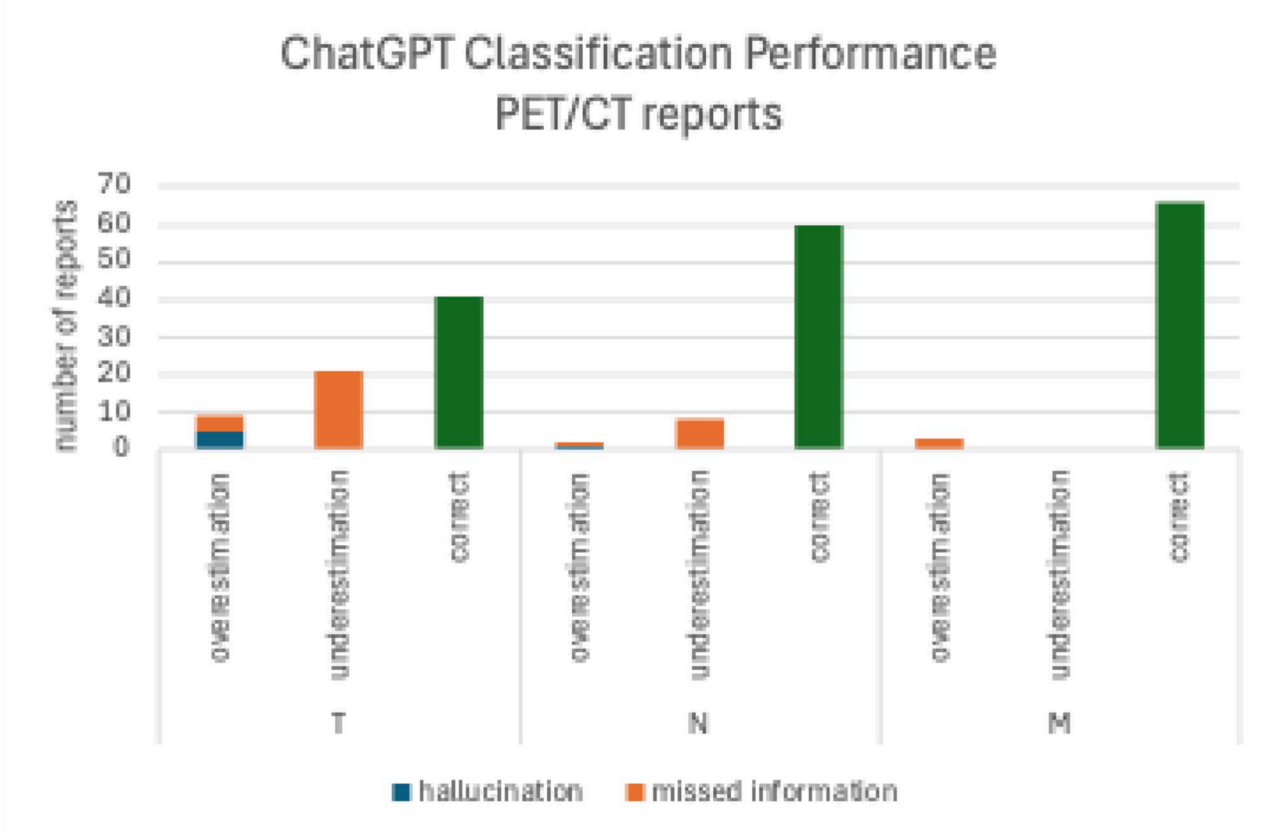


Fig. 2 Performance of ChatGPT-4o for molecular imaging (mi) tumour (T), node (N) and metastases (M) classification applied on PSMA PET/CT reports

insufficiently recognized information and another due to hallucination. For miM, three cases were underestimated because ChatGPT-4o overlooked ambiguous descriptions such as "possible ganglion or lymph node adjacent to the abdominal aorta" or "multiple paraaortic lymph nodes". In both examples, the expert reader rated the findings as miM1a as the possible lymph node adjacent to the aorta showed a SUVmax of 6.0 and should be seen as suspicious. Also the multiple paraaortic lymph nodes were described as enlarged with high SUVmax, leading to the decision of rating them as miM1a by the expert reader. For miT, the interreader agreement between the two readers from PET/CT reports was $\kappa = 0.42$ (CI: 0.25–0.58). The agreement was $\kappa = 0.80$ (CI: 0.64–0.93) for miN and $\kappa = 0.79$ (CI: 0.65–0.92) for miM. Agreement between ChatGPT-4o and the validation reading was $\kappa = 0.38$ (CI: 0.23–0.53) for miT, $\kappa = 0.69$ for miN (CI: 0.55–0.84), and $\kappa = 0.90$ (CI: 0.79–1.0) for miM.

In the 39 PET/CT reports eligible for PRIMARY scoring (local tumor present), ChatGPT-4o achieved correct ratings in 70.4% (95%CI: 0.50–0.85) of patients with an interreader agreement of $\kappa = 0.77$ (CI: 0.16–0.95) when comparing the validation reading with ChatGPT-4o.

PROMISE classification extracted from PET/MRI reports

In 37 PET/MRI reports, overall miTNM classification accuracy was at 91.0% (95%CI: 0.84–0.94). Component-wise accuracy was 83.8% for miT (95%CI: 0.70–0.95), 100% for miN, and 89.1% for miM (95%CI: 0.77–0.97).

ChatGPT-4o *underestimated* one miT category, and *overestimated* 5 miT and 4 miM categories. All T and M stages misclassifications resulted from information that was not explicitly stated or insufficiently captured in the reports. For miT, a frequent cause of overestimation was that ChatGPT-4o inferred capsular infiltration and classified cases as T3, even though capsular breach was not described ($n = 4$), detailed information of this shift can be seen in Table 3. For further details on misclassification of PET/MRI reports, see Supplementary Table S1. Additionally, some cases were underestimated because subtle distinctions between unifocal (u) or multifocal (m) tracer uptake were not captured ($n = 2$). For miT, the interreader agreement between the two readers from PET/MRI reports was $\kappa = 0.59$ (CI: 0.41–0.75). The agreement was $\kappa = 0.58$ (CI: 0.30–0.81) for miN and $\kappa = 0.65$ (CI: 0.35–0.85) for miM. Agreement between ChatGPT-4o and the validation reading was $\kappa = 0.78$ (CI: 0.62–0.92) for miT, $\kappa = 1.00$ for miN, and $\kappa = 0.78$ (CI: 0.55–0.95) for miM.

Table 3 ChatGPT-4o vs. human validation reading for molecular imaging (mi) tumour (T) classification applied on PSMA PET/MRI reports

PSMA PET/MRI			Incorrect predictions										
Validation	Total count	Correct prediction	Overestimations						Underestimations				
			T2u	T2m	T3a	T3b	T3c	T4	T0	T2u	T2m	T3a	T3b	T3c
T0	15	**15**	0	0	0	0	0	0	0	0	0	0	0	0
Tr	9	**8**	0	0	0	0	0	**1**	0	0	0	0	0	0
T2u	5	**2**	**1**	0	**2**	0	0	0	0	0	0	0	0	0
T2m	4	**2**	0	0	**1**	0	0	0	0	**1**	0	0	0	0
T3a	0	na	na	na	na	na	na	na	na	na	na	na	na	na
T3b	1	**1**	0	0	0	0	0	0	0	0	0	0	0	0
T4	3	**3**	0	0	0	0	0	**0**	0	0	0	0	0	0

For miN, underestimation occurred when ChatGPT-4o did not fully interpret ambiguous formulations such as "pathologic lymph nodes in the small pelvis extending up to the paracaval region." (Fig. 3).

Among the 17 PET/MRI images eligible for PRIMARY scoring, ChatGPT-4o achieved an accuracy of 88.1% (95%CI: 0.69-1.00) with an interreader agreement of $\kappa = 0.96$ (CI: 0.78-1.00). when comparing the validation reading with ChatGPT-4o.

Comparison of miTNM and PROMISE results for PET/CT and PET/MRI

ChatGPT-4o achieved significantly higher accuracy for PET/MRI compared to PET/CT in T-stage classification (83.8% vs. 57.7%; OR = 3.83, 95% CI: 1.34–12.69, $p = 0.006$) and N-stage classification (100% vs. 85.9%, $p = 0.014$). In contrast, performance was comparable for M-stage classification (89.1% vs. 95.7%; OR = 0.84, 95% CI: 0.20–4.19, $p = 0.748$) and PRIMARY score assessment (70.4% vs. 88.1%; OR = 0.43, 95% CI: 0.05–2.14, $p = 0.315$). Overall, ChatGPT-4o demonstrated high accuracy for miTNM classification, with significantly more accurate performance on PET/MRI reports compared to PET/CT reports (91.0% vs. 79.8%; OR: 2.80 95% CI: 1.32–6.51, $p = 0.003$).

Accuracy analysis by clinical indications (initial staging vs. recurrence assessment) for PET

For initial staging ($n = 56$), overall miTNM classification accuracy was 85.9% (95% CI: 0.74–0.86). Component-wise accuracy was 60.7% for miT, 87.5% for miN, and 96.4% for miM. For recurrence ($n = 52$), overall miTNM classification accuracy was 81.5% (95% CI: 0.79–0.90). Component-wise accuracy was 73.0% for miT, 94.2% for miN, and 90.4% for miM. Comparisons between initial staging and recurrence showed no significant differences for miT classification (OR = 1.76, 95% CI: 0.74–4.27, $p = 0.228$), miN classification (OR = 2.03, 95% CI: 0.50–9.83, $p = 0.362$), or miM classification (OR = 0.45, 95% CI: 0.07–2.23, $p = 0.312$). Overall, ChatGPT-4o performance was comparable between both clinical indications (OR = 1.31, 95% CI: 0.73–2.52, $p = 0.315$), as illustrated in Fig. 4. Hallucinations were observed in only six cases.

Plausibility assessment of ChatGPT-4o explanations to results

The factual plausibility of ChatGPT-4o's explanations were rated consistently high among modalities for miTNM and PRIMARY scoring. Mean Likert scores were:

- miT = 4.8 ± 0.37, miN = 4.9 ± 0.30, miM = 4.8 ± 0.30 for PET/CT.

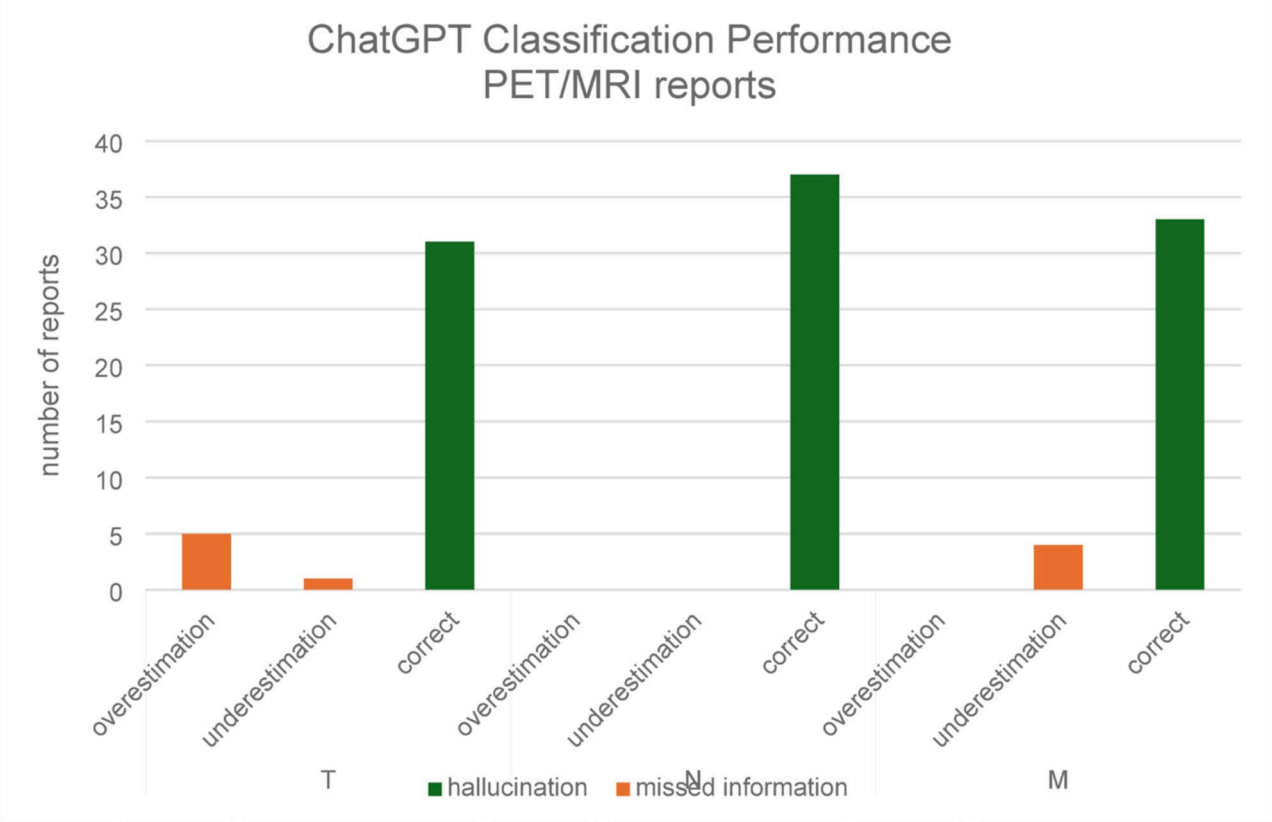


Fig. 3 Performance of ChatGPT-4o for molecular imaging (mi) tumour (T), node (N) and metastases (M) classification applied on PSMA PET/MRI reports

- miT = 4.9 ± 0.38, miN = 4.9 ± 0.38, miM = 4.9 ± 0.51 for PET/MRI.
- 4.1 ± 1.4 for PRIMARY in PET/CT.
- 4.9 ± 0.32 for PRIMARY in PET/MRI.

Discussion

To our knowledge, this is the first study to evaluate the utility of LLMs for the automated extraction of PROMISE V2-based classifications from unstructured PSMA PET/CT and PET/MRI reports in patients with prostate cancer. Our results demonstrate that ChatGPT-4o enables extraction of miTNM and PRIMARY classifications from free-text reports in clinical routine, with high accuracy for nodal (miN) and distant metastasis (miM) staging across both modalities. However, accuracy for miT classification was significantly higher for PSMA-PET/MRI reports.

For PSMA-PET/CT, the accuracy was moderate for miT (57.7%), but high for miN (85.9%) and miM (95.7%). For PSMA-PET/MRI, the accuracy was sufficient for miT (83.8%), miN (100%) and miM (89.1%). These results are consistent with Yuan et al., who reported high accuracy rates of up to 94% for TNM staging using LLMs in prostate cancer [17]. However, classification of the primary tumour (miT) was less accurate in our PET/CT cohort. One explanation is the frequent use of vague or missing information on extracapsular extension/no extracapsular extension (T3) and potential infiltration (T4), as well as missing or vague description on unifocal or multifocal presence of cancer lesions in free-text reports, making precise staging difficult. For instance, ChatGPT-4o often underrated T stage due to missing detailed information on uni-/multifocality of the primary tumor, which has been similarly observed by Nakamura et al. who reported high accuracy for N (78.9%) and M (86.7%) staging but markedly lower accuracy for T staging (52.2%) in lung cancer [18]. Suzuki et al., identified a related issue in pancreatic cancer reports, where underestimation of T4 tumours occurred when vessel "contact" was described without explicitly stating vascular invasion [19].

These findings suggest that the performance of LLMs is closely linked to the precision and clarity of the underlying report, highlighting a significant bottleneck for the implementation of automated structured reporting and underscoring the importance of standardized phrasing within free-text reports.

Therefore, we emphasize the importance of either using highly structured language within free-text reports or explicitly incorporating linguistic nuances, tailored to each center's typical terminology into the prompt to minimize misclassification. Even the incorporation of domain specific knowledge into the prompt already demonstrated

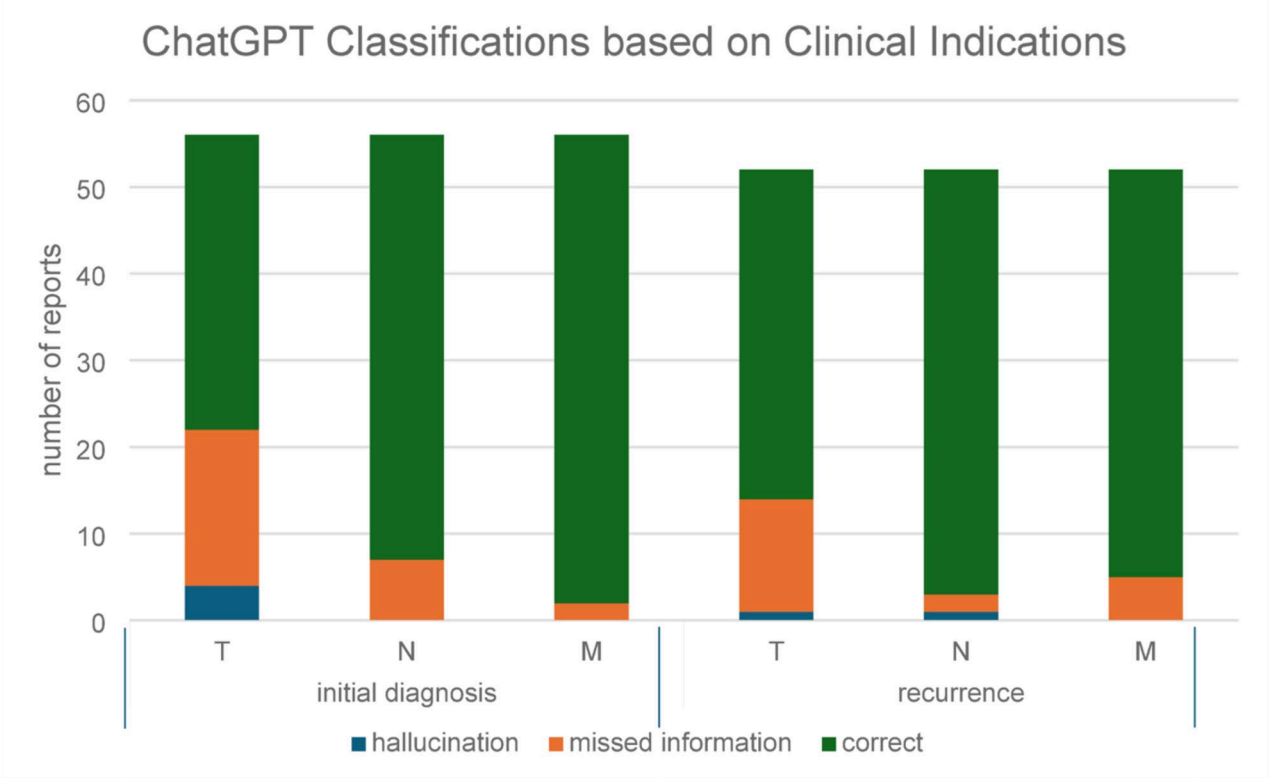


Fig. 4 Performance of ChatGPT-4o for molecular imaging (mi) tumour (T), node (N) and metastases (M) classification among clinical indications recurrence vs. initial staging

high performance through expert-informed prompt engineering, achieving a T-staging accuracy of 91.3% [20].

In our analysis ChatGPT-4o achieved notably higher miT and miN accuracy from PET/MRI reports compared to PET/CT reports (miT: 83.8% vs. 57.7%, $p = 0.006$; miN: 100% vs. 85.9%, $p = 0.014$), interestingly, ChatGPT-4o showed higher interreader agreement compared to the validation set than interreader agreement between the two human readers for PET/MRI reports ($\kappa = 0.78$ vs. $\kappa = 0.59$) and lower agreement for PET/CT reports (vs. 0.38 vs. $\kappa = 0.42$) for T stage and also for N stage (PET/MRI: $\kappa = 1.00$ vs. $\kappa = 0.58$; PET/CT: $\kappa = 0.69$ vs. $\kappa = 0.80$). miM accuracy remained comparable for PET/CT and PET/MRI (89.1% vs. 95.7%, $p = 0.748$) and ChatGPT-4o showed higher agreements compared to the validation sets than the agreements between the two human readers in both modalities (PET/MRI: $\kappa = 0.78$ vs. $\kappa = 0.65$; PET/CT: $\kappa = 0.90$ vs. $\kappa = 0.79$). This variability in performance has also been reported by Alalawi et al., who demonstrated that GPT-4o's performance varies considerably across different radiology subspecialties and imaging modalities [21]. However, the observed differences in miTNM classification performance between PET/MRI and PET/CT in our study should be interpreted with caution, given the limited number of PET/MRI reports included.

From a clinical perspective, accurate miT classification is critical, as misclassification can lead to inappropriate treatment strategies. In our study, most errors, which were mainly minor and caused by missing detailed information of tumour extent for accurate use of PROMISE in free-text-reports, were underestimations. These observations underline the importance of not relying solely on automated extraction tools yet, making human oversight remain essential to ensure clinical accuracy and patient safety.

Given the limited diagnostic value of PSMA-PET/CT for primary prostate tumours [22–24], the PRIMARY score was developed to enhance local staging accuracy and guide PSMA-PET-guided biopsy, particularly in newly diagnosed or biopsy-naïve patients [25]. Applying the PRIMARY score to both modalities, our results demonstrated comparable accuracy for PET/MRI reports (88.1%) compared to PET/CT (70.4%). The slightly superior performance of PET/MRI (even though there was no significant difference) in the PRIMARY score may also be attributed to its higher soft-tissue contrast, which allows for more precise description of the anatomical localization and consequently reporting of the primary tumour extent e.g. exact discrimination of peripheral vs. transitional zones).

Across all parameters, the rationale for each classification using a five-point Likert scale was rated very high between 4.1 and 4.9. This supports its use as an explainable AI system in clinical practice. However, rather than replacing physicians, ChatGPT-4o can function as an augmented intelligence assistant, helping structure complex reports, support tumour board preparation, and reduce time-to-decision in high-throughput settings – especially as PSMA-PET/CT is gaining widespread clinical use following the EMA/FDA approval of [^{177}Lu]Lu-PSMA-617 (Pluvicto) and growing demand for PSMA-targeted diagnostics and therapy. Moreover, ChatGPT-4o could support not only structured data extraction for clinical use, but also for clinical trials to ensure adherence, particularly if standardized reporting schemes incorporating PSMA response metrics such as PSMA-PET Progression (PPP) and the Response Evaluation Criteria in PSMA-PET/CT (RECIP) are used. For example, Andrei et al. have shown that the RECIP criteria are highly relevant for assessing overall survival (OS), as patients classified as having RECIP-progressive disease exhibit significantly lower OS compared to those with stable disease or a therapy response [26, 27]. Prior studies, such as those by Fink et al. and Bhayana et al., have already highlighted this role in radiology [5, 14].

Several limitations must be acknowledged. First, this was a retrospective single-center study using German-language reports, which may limit generalizability. Although ChatGPT-4o is multilingual, the combination of German-language source reports and English-language prompts could have introduced subtle comprehension challenges [28]. Even though we pre-evaluated reports using the prompt in both German and English and no differences in the output were observed between the two languages, larger cohorts are needed to identify multilingual differences. Second, both [^{68}Ga]- and [^{18}F]-labelled PSMA tracers were included, which may differ in biodistribution and influence interpretation, especially for the PRIMARY score, as the score 5 was originally developed and validated for ^{68}[Ga]-PSMA-11 PET/CT imaging. Third, sample size was significantly smaller for PET/MRI, mainly because numbers of PET/MRIs in our centre are generally low compared to PET/CTs, which may influence statistical power. Moreover, we used the web interface and not the API, which might lead to slightly variable results as the parameters such as temperature or seed are not fixed in the web application. Ultimately, LLMs are improving fast with newer versions being introduced regularly. Therefore, ChatGPT-4o might already be outdated at time of publication.

Conclusion

Overall, this study supports the feasibility of using ChatGPT-4o as a clinical companion for structured extraction of PROMISE V2-based classifications from PSMA PET reports. By extracting miN and miM as well as the PRIMARY scores from unstructured clinical free-text reports, ChatGPT-4o offers a scalable solution to improve standardization and efficiency in oncologic imaging workflows. Caution is required for T-stage classification, as our results show limited performance for extraction of miT, mainly due to misinterpretation of ambiguous language in free-text reports (for instance, the use of the plural term "uptakes" in the projection of the prostate led ChatGPT-4o to assign T2u, while readers classified the finding as T2m). This underscores that insufficiently structured or imprecise reporting can significantly impair reliable information extraction by LLMs; they cannot compensate for vague or ambiguous input. Accurate performance depends on clear and explicit textual information. However, given the increasing clinical relevance of PSMA PET/CT – especially in the context of PSMA-targeted therapies – such approaches may offer added value in daily practice, particularly under conditions of high workload or limited specialist availability. Further prospective and multi-center validation studies will be essential to confirm generalizability and ensure safe implementation, particularly in high-stakes settings such as oncological stagings. Nonetheless, this work provides a first step toward leveraging language-based AI to support structured, explainable, and efficient reporting in molecular imaging.

Abbreviations

AI	Artificial Intelligence
CI	Confidence Interval
CT	Computed Tomography
DWI	Diffusion Weighted Imaging
EMA	European Medicines Agency
FDA	Food and Drug Administration
GPT	Generative Pre–Trained Transformer
HASTE	Half–Fourier Acquisition Single–shot Turbo spin Echo
LLM	Large Language Model
mi	Molecular Imaging
miM	Molecular Imaging Metastasis Classification
miN	Molecular Imaging Node Classification
miT	Molecular Imaging Tumour Classification
miTNM	Molecular Imaging Tumour–Node–Metastasis Classification
MRI	Magnetic Resonance Imaging
OS	Overall Survival
PET	Positron Emission Tomography
PPP	PSMA–PET Progression
PRIMARY	Prostate Imaging Reporting system for Intra–prostatic PSMA uptake
PROMISE	Prostate Cancer Molecular Imaging Standardized Evaluation
PSA	Prostate Specific Antigen
PSMA	Prostate–Specific Membrane Antigen
RECIP	Response Evaluation Criteria in PSMA–PET/CT
ROI	Region of Interest
SD	Standard Deviation
SUV / SUVmax	Standardized Uptake Value / maximum Standardized Uptake Value

T1	VIBE–T1–weighted Volumetric Interpolated Breath–hold Examination
TNM	Tumour–Node–Metastasis
u / m	Unifocal / Multifocal

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13550-026-01475-z.

Supplementary Material 1

Acknowledgements
Not applicable.

Author Contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Anna Hinterberger, Maurin H. Mangold, Caroline Weigel, Henri Hartmann and Freba Grawe. The first draft of the manuscript was written by Anna Hinterberger and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability
The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval
This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Heidelberg (EK II 2023-886-AF 11).

Consent to participate
Informed consent was waived due to the retrospective character of this study.

Consent for publication
Not applicable.

Competing Interests
The authors have no relevant financial or non-financial interests to disclose.

Received: 30 January 2026 / Accepted: 22 June 2026
Published online: 26 June 2026

References

1. OpenAI. GPT-4 technical report. arXiv 2303.08774 [preprint] https://arxiv.org/abs/2303.08774. Published March 15, 2023. Updated March 27, 2023. Accessed September 22, 2023.
2. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, Nattenmueller J, Neubauer J, Bamberg F, Reisert M, Russe MF. A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology. 2023;308(1):e230970. https://doi.org/10.1148/radiol.230970. PMID: 37489981.
3. Lee KL, Kessler DA, Caglic I, Kuo YH, Shaida N, Barrett T. Br J Radiol. 2025;98(1167):368–74. https://doi.org/10.1093/bjr/tqae236 . PMID: 39535870; PMCID: PMC11840166. Assessing the performance of ChatGPT and Bard/Gemini against radiologists for Prostate Imaging-Reporting and Data System classification based on prostate multiparametric MRI text reports.
4. Bahl M, Balthazar P, Davis MA, Makary MS, Tirumani SH, Whitlow CT. ChatGPT and Large Language Models in Radiology: Perspectives From the Field. AJR Am J Roentgenol. 2024;223(4):e2432022. https://doi.org/10.2214/AJR.24.32022 . Epub 2024 Sep 18. PMID: 39291940.
5. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology. 2024;310(1):e232756. https://doi.org/10.1148/radiol.232756. PMID: 38226883.
6. Eiber M, Herrmann K, Calais J, Hadaschik B, Giesel FL, Hartenbach M, Hope T, Reiter R, Maurer T, Weber WA, Fendler WP. Prostate Cancer Molecular Imaging Standardized Evaluation (PROMISE): Proposed miTNM Classification for the Interpretation of PSMA-Ligand PET/CT. J Nucl Med. 2018;59(3):469–478. https://doi.org/10.2967/jnumed.117.198119. Epub 2017 Nov 9. Erratum in: J Nucl Med. 2018;59(6):992. PMID: 29123012.
7. Seifert R, Emmett L, Rowe SP, Herrmann K, Hadaschik B, Calais J, Giesel FL, Reiter R, Maurer T, Heck M, Gafita A, Morris MJ, Fanti S, Weber WA, Hope TA, Hofman MS, Fendler WP, Eiber M. Second Version of the Prostate Cancer Molecular Imaging Standardized Evaluation Framework Including Response Evaluation for Clinical Trials (PROMISE V2). Eur Urol. 2023;83(5):405–12. Epub 2023 Mar 17. PMID: 36935345.
8. Toriihara A, Nobashi T, Baratto L, Duan H, Moradi F, Park, Iagaru A. Comparison of 3 interpretation criteria for 68Ga-PSMA11 PET based on inter-and intra-reader agreement. J Nucl Med. 2020;61(4):533–9.
9. Ma L, Hao Y, Zhai L, Zhang W, Cao X, Jia K. Which PSMA PET/CT interpretation criteria most effectively diagnose prostate cancer? a retrospective cohort study. BMC Med Imaging. 2025;25(1):23. https://doi.org/10.1186/s12880-025-01557-9 . PMID: 39833713; PMCID: PMC11749428.
10. Oprea-Lager DE, van Elst T, Aluwini S, Dewulf E, der Poel HV, Stoevelaar H, Bangma CH, Beeker A, Boudewijns S, Budiharto T, de Jong IJ, de Vries KC, Donswijk ML, Fütterer JJ, Hamberg P, Heijmen L, Hoekstra RJ, Kerkhofs TMA, Lavalaye J, Bruin DL, Noordzij W, Oving IM, Robbrecht DGJ, Schaake EE, van de Luijtgaarden ACM, van den Bergh RCN, van den Berkmortel F, van der Hulle T, van der Mijn JCK, van Dodewaard-de Jong JM, van Kruchten M, van Leeuwen PJ, van Limbergen E, van Moorselaar RJA, van Oort IM, van Roermund JGH, van Soest RJ, Veninga T, Vis AN, Voortman J, Willemse PM, Yakar D, Mehra N. How Does Routine Prostate-specific Membrane Antigen Positron Emission Tomography/Computed Tomography Modify the Current Management of Prostate Cancer? A Multidisciplinary View. Eur Urol Open Sci. 2025;75:69–79. https://doi.org/10.1016/j.euros.2025.03.007 . PMID: 40256661; PMCID: PMC12008539.
11. Karpinski MJ, Hüsing J, Claassen K, Möller L, Kajüter H, Oesterling F, Grünwald V, Umutlu L, Kleesiek J, Telli T, Merkel-Jens A, Hüsing A, Kesch C, Herrmann K, Eiber M, Hoberück S, Meyer PT, Kind F, Rahbar K, Schäfers M, Stang A, Hadaschik BA, Fendler WP. Combining PSMA-PET and PROMISE to re-define disease stage and risk in patients with prostate cancer: a multicentre retrospective study. Lancet Oncol. 2024;25(9):1188–1201. https://doi.org/10.1016/S1470-2045(24)00326-7. Epub 2024 Jul 29. Erratum in: Lancet Oncol. 2024;25(12):e626. doi: 10.1016/S1470-2045(24)00635-1. PMID: 39089299.
12. Faggioni L, Coppola F, Ferrari R, et al. Usage of structured reporting in radiological practice: results from an Italian online survey. Eur Radiol. 2017;27:1934–43.
13. Goldberg-Stein S, Chernyak V. 3 Change management—How to implement structured reporting. In: Brook OR, Sommer WH, editors. Radiology structured reporting handbook. New York: Georg Thieme Verlag KG; 2021. pp. 9–13.
14. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, Heußel CP, Kauczor HU, Weber TF. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology. 2023;308(3):e231362. https://doi.org/10.1148/radiol.231362. PMID: 37724963.
15. Sorin V, Barash Y, Konen E, et al. Large language models for oncological applications. J Cancer Res Clin Oncol. 2023;149:9505–8.
16. Tessler I, Wolfovitz A, Livneh N, et al. Advancing medical practice with artificial intelligence: ChatGPT in healthcare. Isr Med Assoc J. 2024;26:80–5.
17. Yuan LH, Huang SW, Chou D, Tsai CY. The In-depth Comparative Analysis of Four Large Language AI Models for Risk Assessment and Information Retrieval from Multi-Modality Prostate Cancer Work-up Reports. World J Mens Health. 2024 Dec 2. https://doi.org/10.5534/wjmh.240173. Epub ahead of print. PMID: 39743220.
18. Nakamura Y, Kikuchi T, Yamagishi Y, Hanaoka S, Nakao T, Miki S, et al. ChatGPT for automating lung cancer staging: feasibility study on open radiology report dataset. medRxiv. 2023;2023. https://doi.org/10.1101/2023.12.11.23299107.
19. Suzuki K, Yamada H, Yamazaki H, Honda G, Sakai S. Preliminary assessment of TNM classification performance for pancreatic cancer in Japanese radiology reports using GPT-4. Jpn J Radiol. 2025;43(1):51–5. https://doi.org/10

.1007/s11604-024-01643-y . Epub 2024 Aug 20. PMID: 39162781; PMCID: PMC11717849.
20. Ismayilov R, Aktas A, Gencoglu EA, Oguz A, Altundag O, Akcali Z. Staging Prostate Cancer with AI: A Comparative Study of Large Language Models and Expert Interpretation on PSMA PET-CT Reports. Mol Imaging Biol. 2026;28(1):93–105. https://doi.org/10.1007/s11307-025-02072-7 . Epub 2025 Dec 5. PMID: 41350971.
21. Alalawi S, Alchoghari R, Hakami A, Alhazmi A, Almutari E, Asseiri R, Alzahrani A, Aladhyani M, Alqarni S, Al-Sharydah AM. Evaluating the role of ChatGPT in structured radiology reporting: A systematic review. Med (Baltim). 2026;105(6):e47541. PMID: 41650069; PMCID: PMC12885729.
22. Berger I, Annabattula C, Lewis J, et al. 68Ga-PSMA PET/CT vs. mpMRI for locoregional prostate cancer staging: correlation with final histopathology. Prostate Cancer Prostatic Dis. 2018;21:204–11.
23. Ferraro DA, Becker AS, Kranzbühler B, et al. Diagnostic performance of 68Ga-PSMA-11 PET/MRI-guided biopsy in patients with suspected prostate cancer: a prospective single-center study. Eur J Nucl Med Mol Imaging. 2021;48:3315–24.
24. Wang L, Yu F, Yang L, et al. 68Ga-PSMA-11 PET/CT combining ADC value of MRI in the diagnosis of naive prostate cancer: perspective of radiologist. Med (Baltim). 2020;99:e20755.
25. Emmett L, Papa N, Buteau J, Ho B, Liu V, Roberts M, Thompson J, Moon D, Sheehan-Dare G, Alghazo O, Agrawal S, Murphy D, Stricker P, Hope TA, Hofman MS. The PRIMARY Score: Using Intraprostatic 68Ga-PSMA PET/CT Patterns to Optimize Prostate Cancer Diagnosis. J Nucl Med. 2022;63(11):1644–50. https://doi.org/10.2967/jnumed.121.263448 . Epub 2022 Mar 17. PMID: 35301240; PMCID: PMC9635676.
26. Fanti S, Hadaschik B, Herrmann K. Proposal for Systemic-Therapy Response-Assessment Criteria at the Time of PSMA PET/CT Imaging: The PSMA PET Progression Criteria. J Nucl Med. 2020;61(5):678–82. https://doi.org/10.2967/jnumed.119.233817 . Epub 2019 Dec 5. PMID: 31806774; PMCID: PMC7198387.
27. Gafita A, Rauscher I, Weber M, Hadaschik B, Wang H, Armstrong WR, Tauber R, Grogan TR, Czernin J, Rettig MB, Herrmann K, Calais J, Weber WA, Benz MR, Fendler WP, Eiber M. Novel Framework for Treatment Response Evaluation Using PSMA PET/CT in Patients with Metastatic Castration-Resistant Prostate Cancer (RECIP 1.0): An International Multicenter Study. J Nucl Med. 2022;63(11):1651–1658. https://doi.org/10.2967/jnumed.121.263072. Epub 2022 Apr 14. Erratum in: J Nucl Med. 2023;64(9):1503. PMID: 35422442; PMCID: PMC9635677.
28. Jiao W, Wang W, Huang J, Wang X, Tu Z. Is ChatGPT a good translator? yes with GPT-4 as the engine. arXiv [Preprint]. arXiv:2301.08745v4 (2023). Available online at: https://arxiv.org/abs/2301.08745 (Accessed March 23, 2023.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.