

Benchmarking GPT-5 in Radiation Oncology: Measurable Gains, but Persistent Need for Expert Oversight

Ugur Dinc^{1,2,3*}, Jibak Sarkar^{1,2,3}, Philipp Schubert^{1,2,3}, Sabine Semrau^{1,2,3}, Thomas Weissmann^{1,2,3}, Andre Karius^{1,2,3}, Johann Brand^{1,2,3}, Bernd-Niklas Axer^{1,2,3}, Ahmed Gomaa^{1,2,3}, Pluvio Stephan^{1,2,3}, Ishita Sheth^{1,2,3}, Sogand Beirami^{1,2,3}, Annette Schwarz^{1,2,3}, Udo Gaipf^{1,2,3}, Benjamin Frey^{1,2,3}, Christoph Bert^{1,2,3}, Stefanie Corradini^{4,3}, Rainer Fietkau^{1,2,3} and Florian Putz^{1,2,3}

¹Department of Radiation Oncology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

²Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), Erlangen, Germany

³Bavarian Cancer Research Center (BZKF), Munich, Germany

⁴Department of Radiation Oncology, University Hospital, Ludwig Maximilian University of Munich, Munich, Germany

Correspondence*:

Dr. Dr. Ugur Dinc
ugur.dinc@fau.de

ABSTRACT

Introduction: Large language models (LLM) have shown great potential in clinical decision support and medical education. GPT-5 is a novel LLM system that has been specifically marketed towards oncology use. This study comprehensively benchmarks GPT-5 for the field of radiation oncology.

Methods: Performance was assessed using two complementary benchmarks: (i) the American College of Radiology Radiation Oncology In-Training Examination (TXIT, 2021), comprising 300 multiple-choice items, and (ii) a curated set of 60 authentic radiation oncologic vignettes representing diverse disease sites and treatment indications. For the vignette evaluation, GPT-5 was instructed to generate structured therapeutic plans and concise two-line summaries. Four board-certified radiation oncologists independently rated outputs for correctness, comprehensiveness, and hallucinations. Inter-rater reliability was quantified using Fleiss' κ . GPT-5 results were compared to published GPT-3.5 and GPT-4 baselines.

Results: On the TXIT benchmark, GPT-5 achieved a mean accuracy of 92.8%, outperforming GPT-4 (78.8%) and GPT-3.5 (62.1%). Domain-specific gains were most pronounced in dose specification and diagnosis. In the vignette evaluation, GPT-5's treatment recommendations were rated highly for correctness (mean 3.24/4, 95% CI: 3.11–3.38) and comprehensiveness (3.59/4, 95% CI: 3.49–3.69). Hallucinations were rare (mean 10.0%), and no case reached majority consensus for their presence. Inter-rater agreement was low (Fleiss' κ 0.083 for correctness),

reflecting inherent variability in clinical judgment. Errors clustered in complex scenarios requiring precise trial knowledge or detailed clinical adaptation.

Discussion: GPT-5 clearly outperformed prior model variants on the radiation oncology multiple-choice benchmark. Although GPT-5 exhibited favorable performance in generating real-world radiation oncology treatment recommendations, correctness ratings indicate room for further improvement. While hallucinations were infrequent, the presence of substantive errors underscores that GPT-5-generated recommendations require rigorous expert oversight before clinical implementation.

Keywords: GPT-5, Artificial Intelligence, Radiation Oncology, Large Language Models, Treatment Recommendation, Oncologic Decision Support, Hallucination, Real-World Evaluation

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly in recent years, driven by scaling of parameters (1), reinforcement-learning-based alignment (2, 3), and the development of modular architectures such as Mixture-of-Experts (MoE) (4, 5). These innovations have enabled broad use of LLMs across scientific and clinical domains (6, 7, 8). In biomedicine, domain-specific pretraining and clinical fine-tuning have enhanced representation of medical terminology and workflows (9, 10, 11), while general-purpose models have achieved exam-level performance in several evaluations (12, 13, 14). Nonetheless, accuracy remains heterogeneous across specialties and problem types (15, 16, 17, 18). Current consensus emphasizes transparent communication of model limitations and the need for sustained human oversight (17, 18).

Within radiation oncology, deep learning methods are established for tasks such as segmentation, image enhancement, dose estimation, and outcome prediction (19, 20, 21, 22, 23, 24, 25). LLMs extend this toolkit with text-centric applications such as guideline summarization, structured rationale generation, question answering, and automated documentation (26, 27, 28). Evaluation of such models now include physics-focused question sets (29, 30) as well as surgical and board-style assessments (31, 32). Radiation oncology-specific studies underscore both the promise of LLMs and the persistence of domain-specific limitations (e.g., dose prescription, differentiation between percutaneous and interstitial techniques) as well as challenges in keeping pace with evolving trial evidence (16, 17).

A key area of ongoing research is the use of large language models (LLMs) as clinical decision support (CDS) systems, with prominent initiatives including Med-PaLM, Med-PaLM 2, and Google AMIE (33, 13, 34). LLM agents like AMIE illustrate how LLM-based assistants may retrieve, synthesize, and contextualize medical evidence for patient-specific recommendations under expert supervision (33). Early evaluations report promising accuracy in case-based reasoning and treatment planning, while underscoring the necessity of explicit uncertainty handling and clinician oversight (16, 18, 35).

GPT-5, the latest generation of OpenAI's foundation models, represents a fundamental shift compared to GPT-3.5 and GPT-4 by explicitly incorporating reasoning-focused reinforcement learning reward models (36). In combination with a larger MoE backbone and improved calibration of probabilistic outputs, GPT-5 achieves stronger logical consistency, longer-context reasoning, and higher factual accuracy. Importantly, GPT-5 is a major OpenAI model explicitly positioned as a reasoning model, designed to generate structured, interpretable rationales in addition to predictions. These advances have translated into improved performance across biomedical benchmarks, USMLE-style exams, radiology case reasoning,

and oncology-specific tasks, while also reducing hallucination rates (37, 38, 17, 18). Despite these improvements, supervised use remains essential, particularly in high-stakes oncology settings.

Building on this progress, the present work provides the first comprehensive evaluation of GPT-5 in radiation oncology. We investigate two complementary settings: (i) a benchmark against the American College of Radiology Radiation Oncology In-Training Examination (ACR TXIT) subset using an automated Responses API pipeline directly comparable to GPT-3.5/4 results, and (ii) a novel real-world scenario dataset comprising 60 complex clinical cases without a single established standard of care (39).

By jointly analyzing standardized benchmarking and novel scenario-based evaluation, we assess GPT-5's accuracy, comprehensiveness and hallucination frequency. This dual design allows rigorous quantification of performance while also examining GPT-5's practical usability and failure modes in clinically ambiguous situations. Given the explicit positioning of GPT-5 as a reasoning model for scientific and medical tasks, our study provides a timely and domain-specific benchmark of its potential and limitations in radiation oncology.

2 MATERIALS AND METHODS

This study comprised two complementary evaluations of GPT-5 in radiation oncology: (i) performance on a standardized multiple-choice knowledge benchmark, and (ii) structured decision-support recommendations on real-world oncologic case vignettes. All analyses were performed on de-identified data, using isolated sessions without cross-case information transfer. No protected health information was processed.

2.1 GPT-5 model and prompting framework

The large language model (LLM) under test corresponds to the GPT-5 family as characterized in the publicly released system card (37). GPT-5 is a transformer-based model trained on large text corpora with mixture-of-experts routing and reinforcement learning from human and artificial intelligence (AI) feedback. For reproducibility, standardized instructions were used for all experiments, and prompts/outputs were logged. Each API call was executed in a fresh session to avoid context leakage between cases. Automation was implemented in Python.

2.2 ACR Radiation Oncology In-Training Examination benchmark

Knowledge-based performance was assessed using the 2021 American College of Radiology (ACR) Radiation Oncology In-Training Examination (TXIT) (40, 39), which comprises 300 multiple-choice questions spanning statistics, physics, biology, and clinical radiation oncology across disease sites. Fourteen questions included medical images, of which seven required visual interpretation (Q17, Q86, Q112, Q116, Q125, Q143, Q164). Since only GPT-5 is capable of processing image-based items, these questions were included exclusively in its evaluation. For GPT-3.5 and GPT-4, the image-based items were removed prior to testing, and the total number of eligible questions was adjusted accordingly, resulting in 293 scorable items, consistent with prior work (39).

Questions were presented as stem plus options without auxiliary text. Prompts instructed the model to select exactly one option and return the format `Final answer: X` with $X \in \{A, B, C, D\}$. No external tools (especially web search) or interactive feedback were permitted. Scoring followed established methodology: 1.0 for a correct choice and 0.0 otherwise.

For content analysis, items were mapped to ACR knowledge domains and to a clinical care-path framework (diagnosis, treatment decision, treatment planning, prognosis, toxicity, brachytherapy, and

dosimetry) (40). Items explicitly referencing major clinical trials or guidelines (e.g., Stockholm III, CRITICS, PORTEC-3, ORIOLE, AJCC 8th edition) were flagged for subgroup reporting (41, 42, 43). Results were compared directly to published GPT-3.5 and GPT-4 benchmarks.

To assess clinical usability, we curated a set of 60 anonymized oncologic case vignettes representing a broad spectrum of disease sites and treatment indications, including definitive, adjuvant, salvage, palliative, and reirradiation scenarios. Source cases were sampled from patients treated in 2025 and subsequently stripped of all identifiers. Patient name, birth date, and ID were removed automatically, while age and sex were retained. Clinical information such as diagnosis, stage, grading, comorbidities, and oncologic history was condensed by two physicians into vignettes, ensuring privacy while preserving clinical representativeness.

The final case set was designed to be balanced across tumor sites and treatment contexts (Table 1). Specifically, it included 10 brain tumor cases (glioblastoma, lower-grade glioma, meningioma, vestibular schwannoma, paraganglioma), 10 breast cancer cases (stratified by nodal status, recurrence, and DCIS), 10 lung cancer cases (NSCLC stage III, SBRT, reirradiation, SCLC), 10 rectal/anal cancer cases (neoadjuvant rectal, definitive anal, local recurrence), 10 prostate cancer cases (risk-adapted definitive therapy, biochemical recurrence, local recurrence post-prostatectomy), and 10 metastatic cases (brain, bone, and SBRT). This stratification allowed for evaluation of GPT-5 across both common and complex clinical scenarios, covering a representative cross-section of real-world radiation oncology practice.

2.3 Real-world oncologic decision-support benchmark

Each vignette was paired with a standardized instruction asking GPT-5 to propose a single most appropriate therapeutic plan and briefly justify the recommendation. Required elements included: disease stage, treatment intent, prior therapy, modality/technique, dose/fractionation, target volumes and OAR constraints, expected toxicities, and follow-up considerations. In addition, GPT-5 was instructed to generate a concise two-line summary of the proposed management, which was ultimately used for evaluation.

Outputs were evaluated by four senior radiation oncologists from a tertiary university hospital. Correctness and comprehensiveness were rated on 4-point Likert scales (4 = fully correct / comprehensive, 1 = not clinically justifiable). Hallucinations were flagged per case by each reviewer (binary). For analysis, we computed the *hallucination score*, defined as the mean fraction of reviewers flagging a hallucination across cases (range 0–1). Consensus thresholds were summarized at levels of any ($\geq 1/4$ raters), majority ($\geq 2/4$), strong ($\geq 3/4$), or unanimous (4/4). Inter-rater agreement was estimated using Fleiss' κ for correctness, comprehensiveness, and hallucination scores.

Exploratory subgroup analyses stratified cases by disease site, prior radiotherapy, and treatment intent. Endpoints included correctness, comprehensiveness and hallucination rate.

Below we provide two illustrative vignettes with GPT-5 therapy suggestions; the complete list of benchmark cases is available in the Supplementary Material. Note that dates have been removed from the published vignettes.

Example 1 – Case 2 (83-year-old female)

Diagnosis: Rectal carcinoma (5–11 cm from anal verge, 1 cm above linea dentata), cT4b cN2b cM0, CRM+, EMVI+, G2, mismatch repair–proficient (pMMR).

Anamnesis and work-up: ##.##.#####: Rectoscopy with biopsy → adenocarcinoma, G2, pMMR (MLH1, MSH2, MSH6, PMS2 intact). ##.##.#####: CT chest/abdomen → cT3 cN+ cM0, no distant metastases.

###.###.#####: Pelvic MRI → cT4b N2b, CRM+, EMVI+, suspected vaginal infiltration. ###.###.#####: Multidisciplinary tumor board recommended total neoadjuvant therapy. ###.###.#####: Transvaginal ultrasound → no sonographic vaginal invasion, fat plane preserved.

GPT-5 suggestion (TNT, long course): Radiochemotherapy with 50.4 Gy/28 fractions (IMRT) plus capecitabine 825 mg/m² BID (Mon–Fri), consolidation chemotherapy with mFOLFOX6 q14d × 4 cycles, followed by total mesorectal excision (TME) 8–12 weeks post-RCT.

Example 2 – Case 52 (male, 52 years).

Diagnosis: Prostate carcinoma cT1c cN0 cM0, Gleason 4+3=7b, iPSA 10.8 ng/ml, estimated lymph node risk (YALE) ~7%.

Anamnesis and work-up: ###.###.#####: Prostate biopsies → poorly differentiated adenocarcinoma, max. Gleason 7b (4+3). Since ###.#####: LHRH 3-monthly depot. ###.###.#####: Tumor board (urologic cancers) → recommendation for local curative therapy, patient in good general condition. ###.###.#####: Prostate MRI (pelvis) → no suspicious lymph nodes, indeterminate bone lesion (likely atypical hemangioma), MRI performed for radiotherapy planning. ###.###.#####: Uncomplicated implantation of 3 fiducial gold markers.

GPT-5 suggestion (curative RT + short-term ADT): Definitive radiotherapy to prostate ± proximal seminal vesicles, 60 Gy in 20 fractions (IG-IMRT/VMAT with daily IGRT); no elective pelvic lymph node irradiation; short-term ADT for 4–6 months (if already >6 months, terminate now).

2.4 Statistical analysis

For the TXIT benchmark, we report overall accuracy and descriptive breakdowns by domain, care-path category, and trial/guideline-anchored items. For the clinical decision-support benchmark, primary endpoints were expert-rated correctness, comprehensiveness, hallucinations, and concordance with delivered care. Exploratory subgroup analyses were prespecified; no multiplicity adjustment was applied. All automation, randomization seeds, prompts, and raw outputs are provided as Supplementary Material. Use of de-identified, retrospective vignettes complied with institutional policies for research on non-human-subjects data.

3 RESULTS

3.1 Overall TXIT performance

Across the 293 scorable items of the ACR TXIT (2021), previously reported baselines reproduced web–interface performance of 63.1% for GPT-3.5 and 74.1% for GPT-4. Using the application programming interface (API) with a fixed prompt over five repeated runs, GPT-3.5 achieved 62.1% ± 1.1% and GPT-4 78.8% ± 0.9%, consistent with earlier reports (39).

For GPT-5, we conducted five independent runs that included both text-only and image-based questions. Overall accuracy ranged from 92.3% to 93.0%, with a mean of 92.8%. Performance on the subset of image-based items was lower, with only 2 of 7 questions answered correctly. Given that the item pool, scoring criteria, and adjudication procedures were identical to those used for prior models, the observed improvement in accuracy reflects genuine advances in model capability rather than differences in test format or evaluation methodology.

3.2 Domain-wise analysis

Stratified by ACR knowledge domains, GPT-5 preserved historical strengths, reaching at least 95% accuracy in Statistics, CNS/Eye, Biology, and Physics. Performance remained lower for Gynecology (75.0%), and moderately reduced for Gastrointestinal and Genitourinary topics (both around 90%). Compared with GPT-4, the largest absolute gains were observed in Dose (from 59.4% to 87.5%) and Diagnosis (from 76.5% to 91.2%).

3.3 Clinical care-path analysis

When mapped to clinical care-path categories, GPT-5 demonstrated consistently high accuracy:

- 100% in Treatment Planning (11/11), Local Control (2/2), Diagnosis Methodology (3/3), Anatomy (8/8), and Pharmacology (1/1).
- 95.9% (47/49) in Treatment Decision and 95.2% (20/21) in Prognosis Assessment.
- 92.3% in Toxicity, 92.9% in Trial/Study/Guideline, and 91.2% in Diagnosis.
- 88.9% in Brachytherapy and 87.5% in Dose.

On items explicitly anchored to named trials or staging systems, GPT-5 achieved 92.9% (13/14), outperforming GPT-4 (85.7%) and GPT-3.5 (50.0%).

3.4 Evaluation on clinical case vignettes

In the evaluation of 60 real-world radiation oncology vignettes, GPT-5's treatment recommendations were rated as follows:

- Correctness: mean 3.24/4 (95% CI: 3.11–3.38).
- Comprehensiveness: mean 3.59/4 (95% CI: 3.49–3.69).

Hallucinations were infrequent. In total, 24 of 240 individual ratings (60 cases \times 4 raters) were classified as hallucinations, corresponding to an overall hallucination rate of 10%. Thus, the vast majority of ratings (90%) did not identify hallucinations, indicating that such occurrences were infrequent across cases. No case was flagged by the majority of experts (at least two out of four raters). Distribution was 36/60 (60%) with zero hallucination flags and 24/60 (40%) with exactly one.

Inter-rater reliability was low, with Fleiss' $\kappa = 0.083$ for correctness, $\kappa = -0.016$ for comprehensiveness, and $\kappa = -0.111$ for hallucinations, indicating variability in individual reviewer judgments.

When stratified by the six major tumor groups, distinct patterns emerged (Figures 5–7). Hallucinations were rare overall, with prostate and brain tumor cases showing almost none, whereas breast, rectal/anal, lung, and metastasis cases exhibited higher variability (Figure 5). Comprehensiveness was generally high across all groups (median $\geq 3.5/4$), with breast, prostate, and brain tumors rated most consistently complete, and rectal/anal and lung cancers showing broader variability (Figure 6). Correctness displayed the clearest differentiation: prostate and brain tumors achieved the highest median scores ($\geq 3.5/4$), breast and metastases performed intermediately, while rectal/anal and lung cancers scored lowest and most variably (Figure 7).

Beyond these main groupings, more granular subgroup analyses revealed clear data-defined differences. Highest correctness was observed in *prostate, intermediate risk* (correctness 3.83; comprehensiveness 3.83; hallucinations 0%) and *prostate, biochemical recurrence after RPE* (3.67; 3.83; 0%), as well as *small-cell*

lung cancer (SCLC) (3.67; 4.00; 0%). Brain primaries were generally strong: *meningioma* (3.67; 3.67; 0%), *vestibular schwannoma* (3.67; 3.33; 0%), *glioma grade 2/3* (3.50; 3.67; 0%), and *glioblastoma* (3.33; 3.83; 0%); a weaker brain subgroup was *pituitary adenoma* (2.67; 3.50; 12.5%).

Breast adjuvant scenarios were solid for correctness and highly complete but showed higher hallucination rates in several subgroups: *adjuvant low risk* (3.33; 4.00; 25%), *adjuvant node-negative* (3.50; 4.00; 25%), *adjuvant node-positive* (3.17; 3.50; 12.5%), *DCIS* (3.00; 3.33; 12.5%), and *loco-regional recurrence* (3.00; 3.33; 0%).

Within lung cancer, *NSCLC stage III (definitive)* performed moderately (3.33; 3.44; 8.33%), whereas settings requiring finer adaptation were lower: *NSCLC re-irradiation* (2.54; 3.50; 12.5%) and *NSCLC SBRT* (2.78; 3.22; 25%).

For metastatic disease, *palliative bone metastases* performed well (3.67; 3.89; 0%) and *metastatic SBRT* was acceptable (3.11; 3.33; 0%), while *brain metastases* were lower and more error-prone (2.67; 3.33; 18.75%).

Rectum–anal cases were the weakest overall: *anal cancer* (3.00; 3.00; 0%), *neoadjuvant rectal cancer* (2.75; 3.25; 25%), and *local recurrence* (2.33; 3.56; 25%).

These patterns localize remaining challenges to problem settings that demand precise trial knowledge, dose/fractionation choices, or complex multimodality sequencing.

Several representative cases illustrate these limitations:

- Case 7 (low-risk prostate cancer): The system recommended definitive therapy, which some raters considered overtreatment given active surveillance would also have been guideline-concordant, though others accepted it because the patient requested therapy.
- Case 8 (neoadjuvant rectal cancer): Biomarker analysis (e.g. MSI) was omitted, a gap increasingly relevant for therapy planning.
- Case 11 (brain metastases): The combination of ipilimumab and nivolumab was proposed, but with a non–guideline-concordant dosing scheme, lowering correctness.
- Case 17 (DCIS): A radiotherapy boost dose was recommended, which is not guideline-supported and was judged overtreatment.
- Case 29 (NSCLC with SBRT): Systemic therapy options were suggested despite definitive SBRT being the standard in this context.
- Case 36 (lung SBRT): Chemotherapy cycles were not specified, reducing precision despite an otherwise acceptable plan.
- Case 42 (brain metastases): Therapy recommendations were given without histologic confirmation, a prerequisite step that reduced guideline adherence.
- Case 58 (SBRT): Outdated regimens were mixed with correct ones, producing polarized ratings.

Where historical records permitted, GPT-5's recommendations showed high concordance with delivered care across treatment intent, modality/technique, and dose/target ranges, although multiple reasonable options often existed.

4 DISCUSSION

This work extends the literature on large language models (LLMs) in radiation oncology along two complementary axes: standardized examination performance and real-world decision support. On the ACR TXIT subset, the present model attained 92.8% under the same item pool and adjudication rules previously used for GPT-3.5/4, which yielded 63.1% and 74.1% respectively (39). The magnitude of this gain is consistent with the architectural and training changes that distinguish GPT-5 from its predecessors: scaling of transformer capacity, more stable long-context attention, preference optimization with richer feedback, and—critically—its explicit positioning as a reasoning model. Unlike GPT-3.5 and GPT-4, GPT-5 is designed not only to recall information but to generate structured, stepwise rationales, yielding more consistent and interpretable outputs (37, 38). Our use of a constrained response format (“Final answer: X”) reduced adjudication noise without contributing domain knowledge. Examination accuracy nevertheless remains an imperfect surrogate for clinical competence: persistent weaknesses were observed in brachytherapy, fine-grained dosimetry, and evolving trial-specific details, mirroring deficits documented for earlier models (39).

Comparisons with adjacent evaluations clarify where improvements reflect reasoning advances rather than test artifacts. In radiation oncology physics, Holmes *et al.* showed that item structure and distractors meaningfully influence performance (29), while Wang *et al.* demonstrated that simply shuffling answer options alters accuracy (30). Our reproduction of prior TXIT baselines with identical stems, options, and scoring therefore supports the interpretation that GPT-5’s higher scores reflect genuine advances in reasoning and knowledge synthesis rather than format effects. At the same time, gynecologic oncology, brachytherapy, and trial-anchored items remained more challenging, consistent with topic-level variability in prior reports (39, 16, 15).

A more robust and clinically relevant benchmark was established through our 60-case evaluation based on authentic oncologic vignettes, in which GPT-5 was tasked with generating structured management plans. The model produced coherent and comprehensive drafts, extending earlier findings with GPT-4 that had been reported in a smaller Red Journal–style Gray Zone cases (39). Hallucinations were infrequent and did not pose a substantive concern; rather, errors were concentrated in areas requiring detailed trial-specific knowledge, nuanced clinical adaptation or complex multimodality treatments (e.g., SBRT, DCIS, brain metastases, ano-rectal cancer, lung cancer with comorbidities). These results highlight the potential of reasoning-oriented models: GPT-5 is able to synthesize case-relevant rationales, thereby moving closer to providing the deliberative support that is directly applicable in tumor board settings.

Our evaluation complements emerging work such as the *Articulate Medical Intelligence Explorer (AMIE)* system, which was tested in synthetic breast oncology vignettes (33). While AMIE incorporated retrieval and self-critique pipelines and demonstrated performance above trainees and fellows, our study extends this line of research by benchmarking GPT-5 on real, anonymized, multi-disease radiation oncology cases rated by board-certified specialists.

Our findings align with broader medical LLM studies, which consistently show topic-level heterogeneity, stronger outputs under expert scaffolding, and improved interpretability when reasoning steps are made explicit (16, 15, 32, 31). Reviews and meta-analyses converge on supervised applications—education, tumor-board summarization, pre-board preparation—rather than autonomous decision-making (18, 17, 44). Within this trajectory, GPT-5’s positioning as a reasoning model represents a qualitative step: it enables explicit rationale generation and structured synthesis across complex oncology cases, something prior models handled only inconsistently.

Methodologically, our study also connects to emerging multimodal planning assistants that couple LLM reasoning to imaging or dose engines (27, 28). Such systems may compensate for GPT-5's current blind spots in image review and dosimetry but will require rigorous validation and regulatory oversight before clinical use (15, 45). From a governance perspective, near-term deployment should remain supervised, with retrieval-augmented pipelines, auditable links to guidelines and trials (e.g., AJCC, PORTEC-3, ORIOLE), explicit uncertainty labeling, and human sign-off (46, 43, 41, 42, 45).

Several limitations frame the interpretation of our findings. First, standardized examinations sample knowledge in a manner that differs from real-world practice; thus, high TXIT performance does not guarantee reliability in rare, ambiguous, or evolving scenarios (13). Second, the retrospective case cohort—while enriched for evaluability through follow-up and therapy verification—may be affected by survivorship and documentation biases. Third, despite the use of prespecified rubrics, inter-rater variability persisted, typical for complex oncologic vignettes (17). Fourth, our evaluation did not incorporate tool use or external retrieval, which could plausibly further improve accuracy through real-time access to guidelines and trial data. Fifth, the TXIT and 60-case sets cover only a subset of radiation oncology knowledge, and therefore cannot capture the full diversity of clinical decision-making. Finally, model versioning, decoding parameters, and chat memory constraints influence outputs; we mitigated these factors through fresh conversations and repeated runs, but residual variability remains.

Future research should prioritize prospective studies, ideally randomizing tumor-board workflows to model-assisted versus control arms. Additional directions include systematic evaluation of reasoning under tool use and retrieval conditions, direct comparisons of general-purpose and domain-adapted models, explicit assessment of target and dose coherence as well as toxicity forecasting, and development of auditable retrieval pipelines tightly coupled to primary evidence (18). Ultimately, the goal is not to replace clinician judgment, but to generate reproducible, evidence-linked drafts and option sets that accelerate multidisciplinary deliberation while preserving safety, accountability, and trust.

5 LIMITATIONS

Limitations include the absence of tool use and external retrieval, which could further improve performance by enabling real-time access to guidelines, trial updates, and dose-constraint references. The TXIT and 60-case sets cover only a subset of radiation oncology knowledge and therefore cannot capture the full spectrum of clinical decision-making. Outcomes are also influenced by model versioning, decoding parameters, and chat memory constraints; we mitigated these effects through fresh conversations and repeated runs (five each) for GPT-3.5, GPT-4, and GPT-5, but residual variability remains.

6 CONCLUSION

On TXIT benchmarks, GPT-4 outperformed GPT-3.5, and **GPT-5 further increased accuracy to 92.8%**, with strengths in statistics, CNS/eye, physics, diagnostic methods, and toxicity, and persistent gaps in gynecology, brachytherapy, dosimetry, and trial-specific details.

More importantly, our novel 60-case benchmark of real-world oncologic scenarios showed that GPT-5 can generate coherent, comprehensive management drafts. Hallucinations were rare and not a substantive concern; limitations instead reflected occasional inaccuracies in guideline details, multi-modal treatments, and nuanced adaptation of medical knowledge to complex clinical case descriptions.

Overall, GPT-5 emerges as a reasoning model with value in supervised applications such as education, pre-board preparation, and draft generation for tumor boards. Its near-term role is as an augmentative assistant, with human verification and evidence retrieval remaining essential safeguards.

CONFLICT OF INTEREST STATEMENT

The authors declare no commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

All prompts, grading rules, and aggregated results (per-run accuracies) are shared as Supplementary Material. The ACR TXIT 2021 exam is publicly accessible from ACR (usage per their terms).

ETHICS STATEMENT

The use of de-identified, case vignettes complied with institutional research policies and local legislation (BayKrG Art. 27) as well as with the Helsinki declaration and its subsequent amendments. All patients had given their full consent to the use of patient data for secondary scientific purposes. Patient case vignettes had been stripped of all identifiers. Moreover, clinical information such as diagnosis, stage, grading, comorbidities, and oncologic history was condensed by two physicians into vignettes, enabling full privacy while preserving clinical representativeness.

REFERENCES

1. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
2. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30** (2017).
3. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35** (2022) 27730–27744.
4. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
5. Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **23** (2022) 1–39.
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in Neural Information Processing Systems* (2017), vol. 30, 1–13.
7. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems* (2020), vol. 33, 1877–1901.
8. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35** (2022) 24824–24837.
9. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36** (2020) 1234–1240.

10. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3** (2021) 1–23.
11. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323* (2019).
12. OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023) 1–100.
13. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* **620** (2023) 172–180.
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health* **2** (2023) e0000198.
15. Ebrahimi B, Howard A, Carlson DJ, Al-Hallaq H. Chatgpt: can a natural language processing tool be trusted for radiation oncology use? *International Journal of Radiation Oncology, Biology, Physics* **116** (2023) 977–983.
16. Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, et al. Quality of large language model responses to radiation oncology patient care questions. *JAMA network open* **7** (2024) e244630–e244630.
17. Chen D, Parsa R, Swanson K, Nunez JJ, Critch A, Bitterman DS, et al. Large language models in oncology: a review. *BMJ oncology* **4** (2025) e000759.
18. Hao Y, Qiu Z, Holmes J, Löckenhoff CE, Liu W, Ghassemi M, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digital Medicine* **8** (2025) 450.
19. Huang Y, Gomaa A, Hoefler D, Schubert P, Gaip U, Frey B, et al. Principles of artificial intelligence in radiooncology. *Strahlentherapie und Onkologie* **201** (2025) 210–235.
20. Gomaa A, Huang Y, Stephan P, Breininger K, Frey B, Dörfler A, et al. A self-supervised multimodal deep learning approach to differentiate post-radiotherapy progression from pseudoprogression in glioblastoma. *Scientific Reports* **15** (2025) 17133.
21. Huang Y, Bert C, Sommer P, Frey B, Gaip U, Distel LV, et al. Deep learning for brain metastasis detection and segmentation in longitudinal mri data. *Medical Physics* **49** (2022) 5773–5786.
22. Weissmann T, Huang Y, Fischer S, Roesch J, Mansoorian S, Ayala Gaona H, et al. Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. *Frontiers in Oncology* **13** (2023) 1115258.
23. Wang H, Liu X, Kong L, Huang Y, Chen H, Ma X, et al. Improving cbct image quality to the ct level using reggan in esophageal cancer adaptive radiotherapy. *Strahlentherapie und Onkologie* **199** (2023) 485–497.
24. Xing Y, Nguyen D, Lu W, Yang M, Jiang S. A feasibility study on deep learning-based radiotherapy dose calculation. *Medical physics* **47** (2020) 753–758.
25. Erdur AC, Rusche D, Scholz D, Kiechle J, Fischer S, Llorian-Salvador O, et al. Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. *Strahlentherapie und Onkologie* **201** (2025) 236–254.
26. Hou Y, Bert C, Gomaa A, Lahmer G, Höfler D, Weissmann T, et al. Fine-tuning a local llama-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *Frontiers in Artificial Intelligence* **7** (2025) 1493716.
27. Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering* **3** (2024) 133.

28. Liu S, Pastor-Serrano O, Chen Y, Gopaulchan M, Liang W, Buyyounouski M, et al. Automated radiotherapy treatment planning guided by gpt-4-vision. *Physics in Medicine and Biology* (2024).
29. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology* **13** (2023) 1219326.
30. Wang P, Holmes J, Liu Z, Chen D, Liu T, Shen J, et al. A recent evaluation on the performance of llms on radiation oncology physics using questions of randomly shuffled options. *Frontiers in Oncology* **15** (2025) 1557064.
31. Maruyama H, Toyama Y, Takanami K, Takase K, Kamei T. Role of artificial intelligence in surgical training by assessing gpt-4 and gpt-4o on the japan surgical board examination with text-only and image-accompanied questions: Performance evaluation study. *JMIR Medical Education* **11** (2025) e69313.
32. Krumsvik RJ. Gpt-4's capabilities for formative and summative assessments in norwegian medicine exams—an intrinsic case study in the early phase of intervention. *Frontiers in Medicine* **12** (2025) 1441747.
33. Palepu A, Dhillon V, Niravath P, Weng WH, Prasad P, Saab K, et al. Exploring large language models for specialist-level oncology care. *arXiv preprint arXiv:2411.03395* (2024).
34. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. *Nature Medicine* **31** (2025) 943–950.
35. Putz F, Haderlein M, Lettmaier S, Semrau S, Fietkau R, Huang Y. Exploring the capabilities and limitations of large language models for radiation oncology decision support. *International Journal of Radiation Oncology, Biology, Physics* **118** (2024) 900–904.
36. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).
37. [Dataset] OpenAI. Gpt-5 system card. <https://openai.com/index/gpt-5-system-card/> (2025). Accessed 2025-08-10.
38. [Dataset] OpenAI. Gpt-5. <https://openai.com/research/index/> (2025). Accessed 2025-08-10.
39. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking chatgpt-4 on a radiation oncology in-training exam and red journal gray zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Frontiers in Oncology* **13** (2023) 1265024. doi:10.3389/fonc.2023.1265024.
40. Rogacki K, Gutiontov S, Goodman C, Jeans E, Hasan Y, Golden DW. Analysis of the radiation oncology in-training examination content using a clinical care path conceptual framework. *Appl Radiat Oncol* **10** (2021) 41–51.
41. de Boer SM, Powell ME, Mileshekin L, Katsaros D, Bessette P, Haie-Meder C, et al. Adjuvant chemoradiotherapy versus radiotherapy alone in women with high-risk endometrial cancer (portec-3): patterns of recurrence and post-hoc survival analysis of a randomised phase 3 trial. *The lancet oncology* **20** (2019) 1273–1285.
42. Phillips R, Shi WY, Deek M, Radwan N, Lim SJ, Antonarakis ES, et al. Outcomes of observation vs stereotactic ablative radiation for oligometastatic prostate cancer: the oriole phase 2 randomized clinical trial. *JAMA oncology* **6** (2020) 650–659.
43. Amin MB, Greene FL, Edge SB, Compton CC, Gershengwald JE, Brookland RK, et al. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA: a cancer journal for clinicians* **67** (2017) 93–99.

44. Trapp C, Schmidt-Hegemann N, Keilholz M, Brose SF, Marschner SN, Schönecker S, et al. Patient-and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. *Strahlentherapie und Onkologie* **201** (2025) 333–342.
45. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model ai chatbots require approval as medical devices. *Nature Medicine* **29** (2023) 2396–2398.
46. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using ai-generated suggestions from chatgpt to optimize clinical decision support. *Journal of the American Medical Informatics Association* **30** (2023) 1237–1245.

FIGURE CAPTIONS

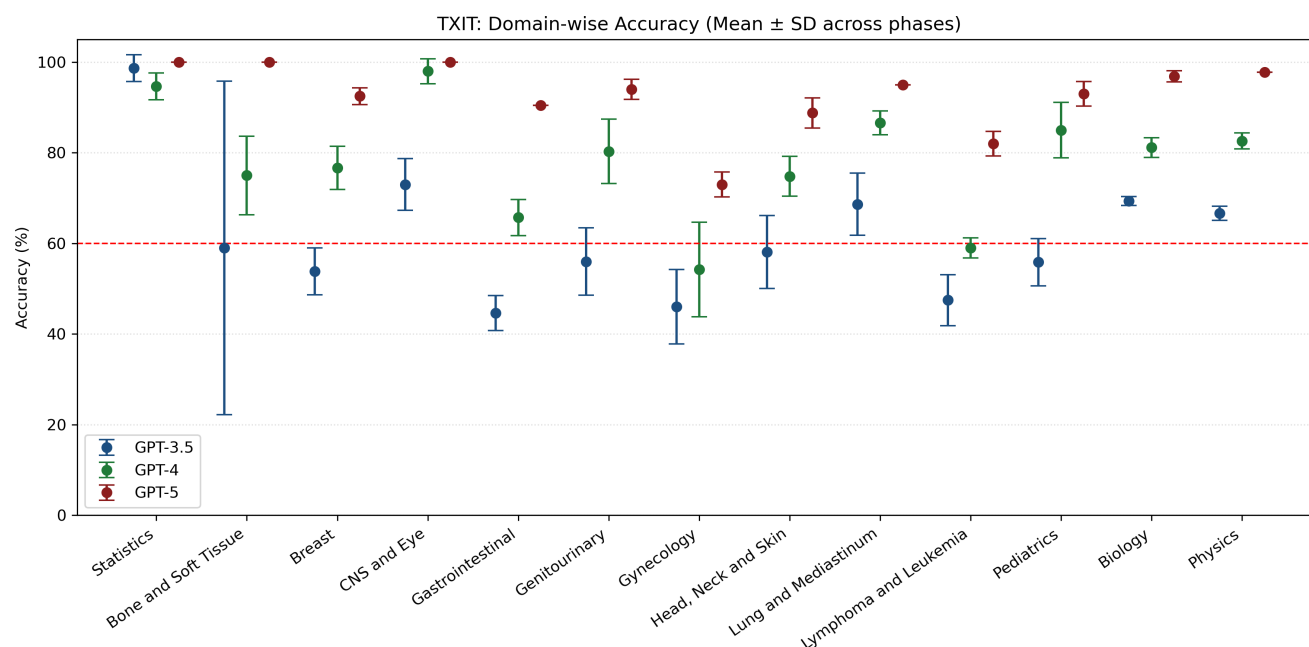


Figure 1. TXIT accuracy by model. Symbols show mean accuracy and error bars indicate the standard deviation (SD) across five runs for GPT-3.5, GPT-4, and GPT-5.

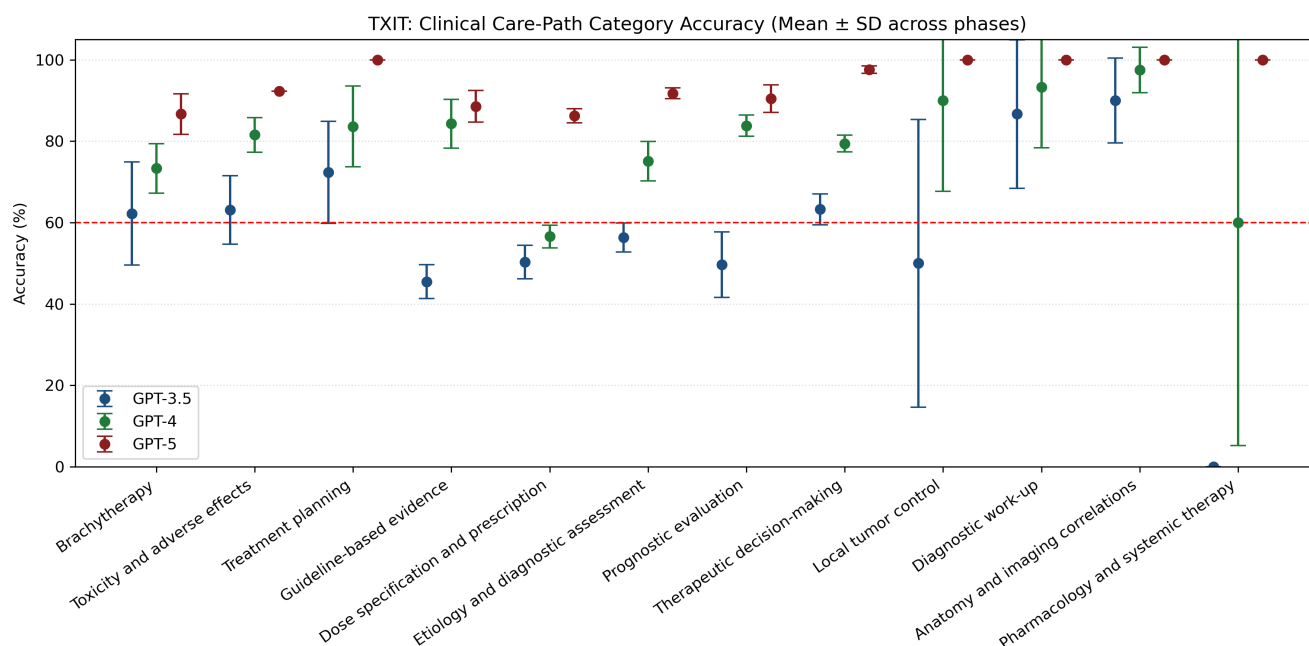


Figure 2. Domain-wise accuracy across models. Symbols show mean accuracy and error bars indicate the SD across five runs for GPT-3.5, GPT-4, and GPT-5.

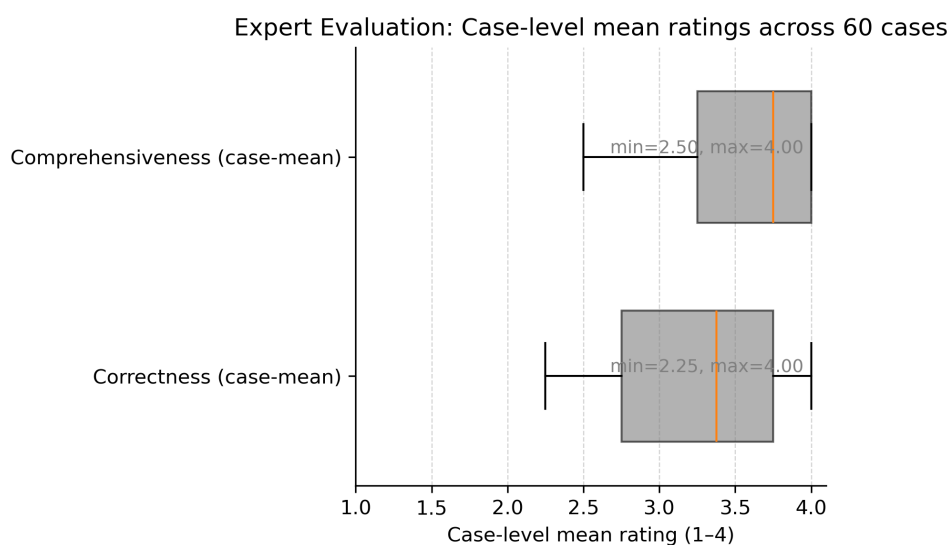


Figure 3. Distribution of case-level mean expert ratings for correctness and comprehensiveness across 60 cases. Each box represents the inter-quartile range (IQR) with whiskers indicating outliers, summarizing the distribution of ratings. Case-level mean correctness ranged from 2.25 to 4.00, while case-level mean comprehensiveness ranged from 2.50 to 4.00.

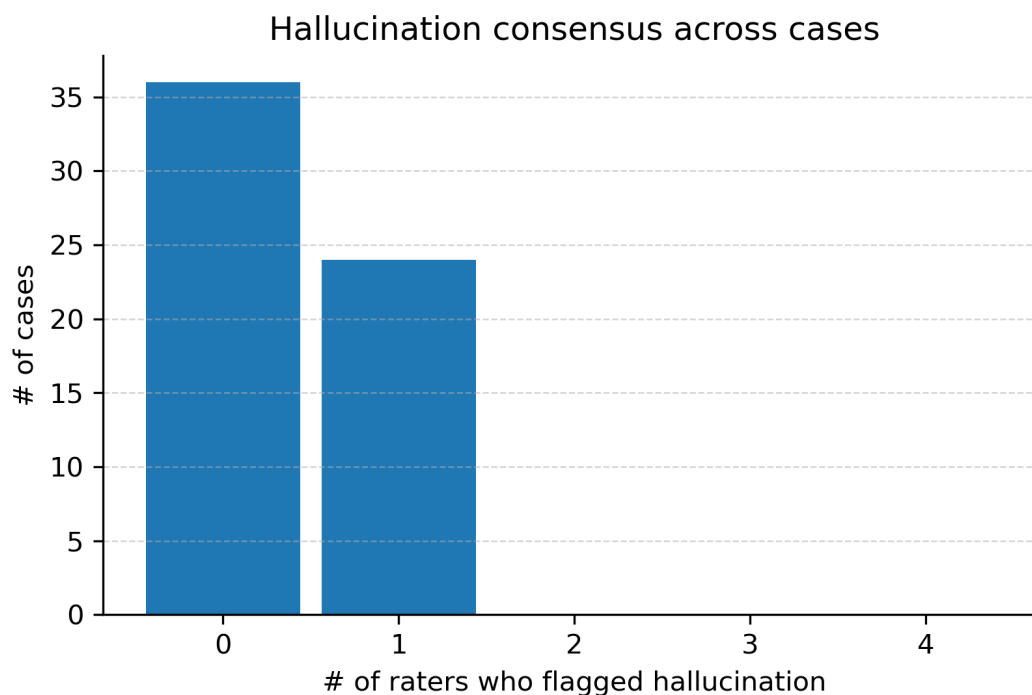


Figure 4. Hallucination consensus across cases. Bars show the number of cases with 0, 1, 2, 3, or 4 raters flagging hallucination. In this cohort, 36/60 cases had 0 flags and 24/60 had exactly 1 flag; no case reached majority ($\geq 2/4$), strong ($\geq 3/4$), or unanimous ($4/4$) consensus.

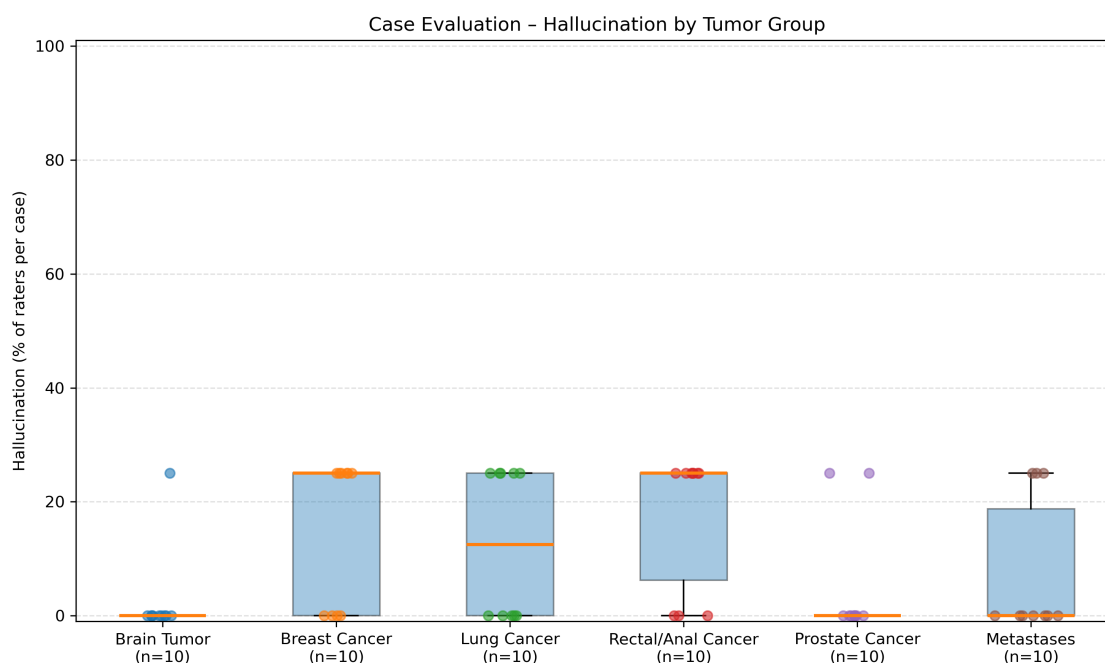


Figure 5. Hallucination rates by tumor group (10 cases each). Hallucinations were rare overall. Prostate and brain tumor cases showed almost no hallucinations, while breast, rectal/anal, lung, and metastasis cases exhibited higher variability, with some reaching up to 25%.

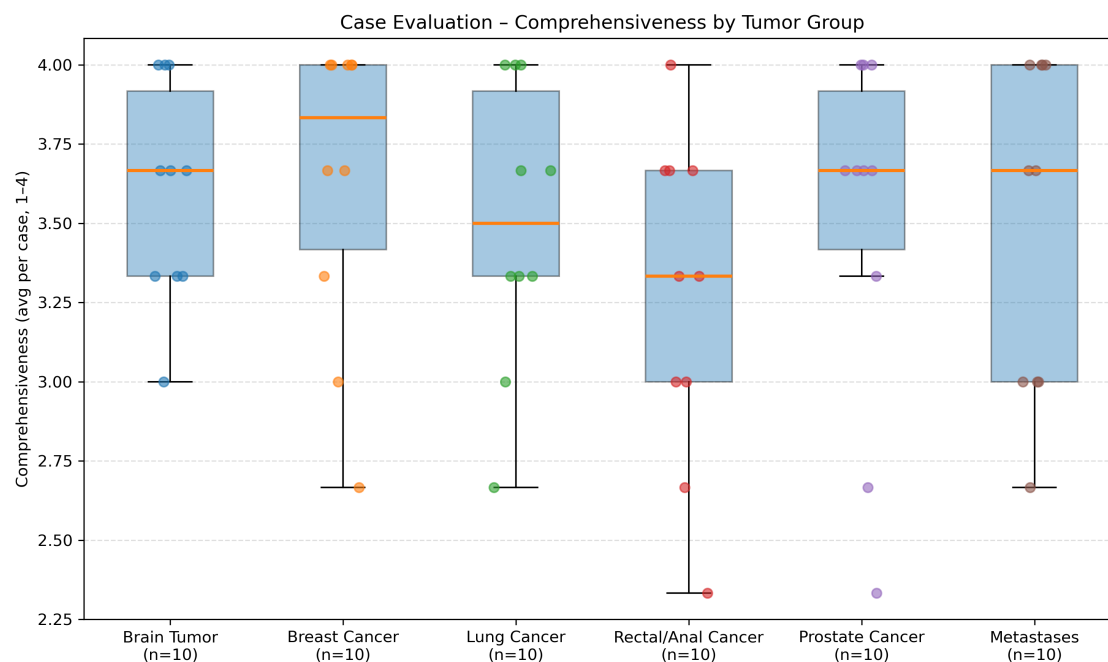


Figure 6. Comprehensiveness ratings (1–4 scale) by tumor group. All groups except for rectal/anal achieved high median scores (≥ 3.5). Breast, prostate, and brain tumor cases were most consistently rated as highly comprehensive, while rectal/anal and lung cancers showed broader variability.

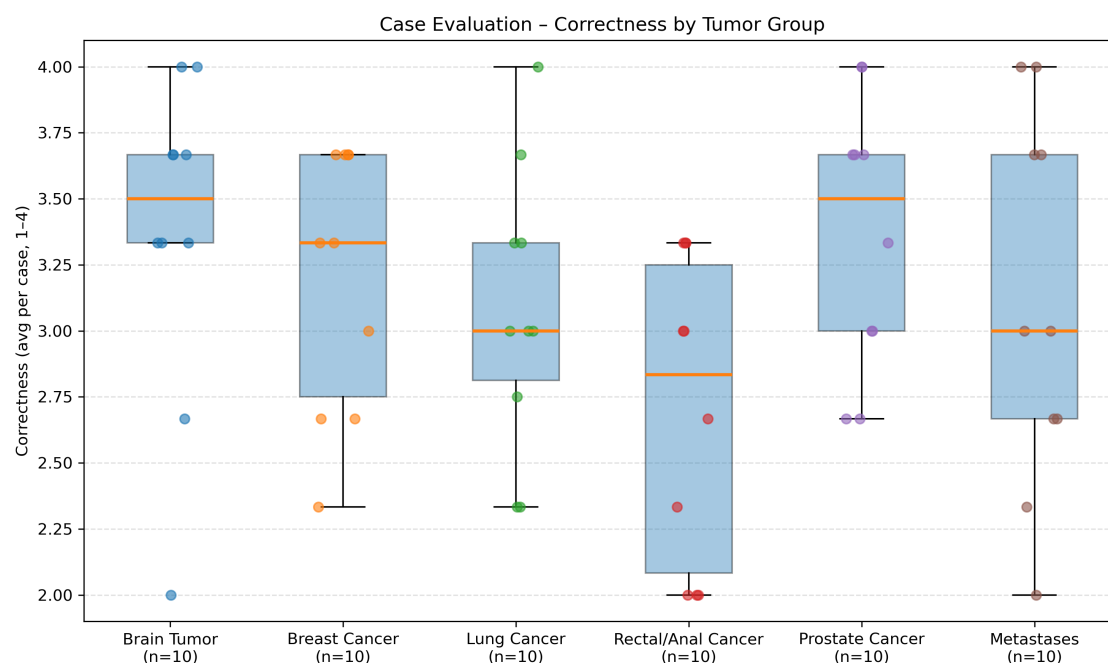


Figure 7. Correctness ratings (1–4 scale) by tumor group. Prostate and brain tumor cases scored highest for correctness (median ≥ 3.5). Lung and rectal/anal cancer cases showed lower and more variable correctness, whereas breast cancer and metastasis cases performed intermediately.

Table 1. Sampling frame for the 60 clinical cases set. Category totals were balanced to support stratified analyses by site and radiooncologic treatment indication.

Category	Subcategory	Number of cases
Brain Tumor	Glioblastoma (grade 4)	2
	Glioma (grade 2/3)	2
	Meningioma	2
	Vestibular schwannoma	2
	Paranglioma	2
Breast Cancer	Adjuvant—nodal positive	2
	Adjuvant—nodal negative	2
	Adjuvant—low risk	2
	Loco-regional recurrence	2
	DCIS	2
Lung Cancer	NSCLC—definitive (stage III)	3
	NSCLC—SBRT	3
	NSCLC—reirradiation	2
	SCLC	2
Rectal/Anal Cancer	Neoadjuvant—rectal cancer	4
	Definitive—anal cancer	3
	Local recurrence	3
Prostate Cancer	Definitive—low risk	2
	Definitive—intermediate risk	2
	Definitive—high risk	2
	Biochemical recurrence after RPE	2
	Local recurrence after RPE + EBRT	2
Metastases	Brain metastases	4
	Palliative—bone metastases	3
	SBRT	3

Abbreviations: DCIS, ductal carcinoma in situ; NSCLC, non-small cell lung cancer; SCLC, small-cell lung cancer; SBRT, stereotactic body radiotherapy; RPE, radical prostatectomy; EBRT, external beam radiotherapy.

Table 2. Clinical cases : case-level outcomes across raters.

Metric	Estimate	95% CI	Notes
Correctness (mean/4)	3.24	3.11–3.38	mean across cases
Comprehensiveness (mean/4)	3.59	3.49–3.69	mean across cases
Hallucination (mean % per case)	10.0%	6.8–13.2%	proportion of raters per case
Any hallucination (per case)	40.0%	28.6–52.6%	24/60 cases
Majority/Strong/Unanimous	0%/0%/0%	–	no case $\geq 2/4$, $\geq 3/4$, or $4/4$
Inter-rater reliability (Fleiss' κ)	0.083 / -0.016 / -0.111	–	correctness / compreh. / hallucination

SUPPLEMENTARY MATERIAL

f"""

You are a tumor board assistant in Germany (radiation oncology, medical oncology). Cite German S3 guidelines first with exact recommendation numbers and short code. Use secondary sources (NCCN/ESMO/ESTRO/ICRU) only as supportive.

Output ONE SINGLE LINE of JSON with EXACTLY these keys:

- "diagnosis_compact": three short lines, each very concise:
 - Line 1: cancer type + side/site (if known) + ED:MM/YY (Erstdiagnose), and year
 - e.g., "Mammakarzinom links ED:05/2023, Rez 03/2025"
 - Line 2: TNM (c/p + T,N,M), key biology (e.g., ER/PR, HER2, p16/HPV, RAS/P53)
 - e.g., "cT2 cN1 cM0, ER/PR+, HER2+, G2"
 - Line 3: starts with "Bisherige onkologische Therapie:" + last relevant point
 - e.g., "Bisherige onkologische Therapie: Zn. OP (BET) 06/2025" or "Bis"
- "therapy_compact": one concise line (German abbreviations) like:
 - "adj. RCT: 50 Gy/25 Fr (ED 2.0 Gy) + Boost 10 Gy/5 Fr; Chemo: CAPOX q21d"
- "tumorformel": concise TNM/tumor formula if inferable; otherwise "Unklar".
- "suggested_therapieplan": concrete, guideline-aligned plan. If listing multiple points, put each point on its own line starting with a number and a closing parenthesis.
 - e.g., "1) First thing\\n2) Second thing\\n3) Third thing".
 - Cover: chemotherapy (drug(s), schema, dose or range), radiotherapy (technique), surgery (if/when appropriate), tumorboard recommendation.
- "notes_to_clinician": practical next steps (further diagnostics before/after therapy, renal/hepatic clearance, dental eval, PEG, toxicity considerations). Use numbers.
- "guideline_primary": German S3 priority list with each item on its own line.
 - "1) S3 [Disease], Empfehlung Nr. X.Y: \\\"short quote\\\""
- "guideline_secondary": secondary brief support with each item on its own line.
 - "1) NCCN v.2025.1: short point" or "2) ESTRO/ICRU: short point"
- "key_characteristics": numbered list (each on its own line) stating:
 - [Therapieelement] { Indikation: [kurze Begründung basierend auf Patientendaten]
 - e.g., "1) RT 70 Gy { Indikation: definitive Behandlung bei lokal fortgeschrittenem Tumor"
- "self_score": integer 0{100 reflecting confidence in the suggestions given (100 = guideline-clear with complete info; lower if key data are missing/ambiguous)

Rules:

- Output JSON ONLY (no prose before/after).
- If critical info is missing, say what's needed in "notes_to_clinician" and "suggested_therapieplan".

Patient INPUT (verbatim):

randID: {rand_id}

Age: {age}

Diagnose: {diagnose}

Nebendiagnosen: {nebendiagnosen}

```
Anamnese: {anamnese}
"".strip()

def build_user_content(question_text: str, image_path: str | None):
    parts = [
        {"type": "input_text", "text": (
            "Please answer the following ACR multiple-choice exam question. "
            "First answer comprehensively deriving from your expert knowledge "
            'then give the final answer in the following form: "Final answer: '
            "where X is A, B, C, or D.\n\n"
            f"ACR question:\n{question_text}\n"
        )}
    ]
    if image_path:
        with open(image_path, "rb") as f:
            b64 = base64.b64encode(f.read()).decode("ascii")
        parts.append({
            "type": "input_image",
            "image_url": f"data:image/png;base64,{b64}",
        })
    return parts
```

Table 2. Supplementary Table S1: Full list of benchmark cases included in the real-world oncologic decision-support evaluation.

Case #	Tumor Site	Clinical Scenario
1	Rectum-Anal	Neoadjuvant – Rectal Cancer
2	Rectum-Anal	Neoadjuvant – Rectal Cancer
3	Lung	NSCLC – SBRT
4	Breast	Adjuvant – nodal positive
5	Brain	Meningioma
6	Breast	Adjuvant – nodal positive
7	Prostate	Definitive – low risk
8	Rectum-Anal	Neoadjuvant – Rectal Cancer
9	Lung	SCLC
10	Rectum-Anal	Neoadjuvant – Rectal Cancer
11	Metastases	Brain metastases
12	Lung	NSCLC – Definitive – Stage III
13	Breast	DCIS
14	Prostate	Definitive – high risk
15	Brain	Glioma grade 2 / 3
16	Brain	Vestibular Schwannoma
17	Breast	DCIS
18	Brain	Pituitary Adenoma
19	Lung	NSCLC – Definitive – Stage III
20	Rectum-Anal	Anal Cancer
21	Breast	Loco-regional recurrence
22	Brain	Glioblastoma grade 4
23	Rectum-Anal	Local Recurrence
24	Brain	Vestibular Schwannoma
25	Prostate	Definitive – high risk
26	Lung	NSCLC – Re-RT
27	Prostate	Local recurrence after RPE + EBRT
28	Rectum-Anal	Local Recurrence
29	Lung	NSCLC – SBRT
30	Prostate	Local recurrence after RPE + EBRT
31	Brain	Glioblastoma grade 4
32	Prostate	Definitive – low risk
33	Lung	NSCLC – SBRT
34	Prostate	Definitive – intermediate risk
35	Brain	Pituitary Adenoma
36	Metastases	SBRT
37	Lung	NSCLC – Re-RT
38	Metastases	Brain metastases
39	Metastases	SBRT
40	Lung	NSCLC – Definitive – Stage III
41	Prostate	Biochemical recurrence after RPE
42	Metastases	Brain metastases
43	Breast	Adjuvant – low-risk
44	Lung	SCLC
45	Metastases	Palliative – bone metastases
46	Prostate	Biochemical recurrence after RPE
47	Rectum-Anal	Local Recurrence
48	Metastases	Palliative – bone metastases
49	Metastases	Palliative – bone metastases
50	Metastases	Brain metastases
51	Rectum-Anal	Anal Cancer
52	Prostate	Definitive – intermediate risk
53	Breast	Adjuvant – nodal negative
54	Breast	Adjuvant – nodal negative

Table 2. Supplementary Table S1: Full list of benchmark cases included in the real-world oncologic decision-support evaluation.

55	Rectum-Anal	Anal Cancer
56	Breast	Adjuvant – low-risk
57	Brain	Glioma grade 2 / 3
58	Metastases	SBRT
59	Brain	Meningioma
60	Breast	Loco-regional recurrence