npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02274-x

ARTEMIS: a pilot study comparing AI-based and expert therapeutic decisions in simulated clinical cases of neuroendocrine neoplasms

Giuseppe Lamberti[1], Francesco Panzuto[2] ✉, Sara Massironi[3], Mauro Cives[4,5], Anna La Salvia[6], Francesca Spada[7], Antongiulio Faggiano[8], Sara Pusceddu[9], Manuela Albertelli[10,11], Salvatore Tafuto[12], Elisa Andrini[1], Claudio Ricci[1] & Davide Campana[1]

Neuroendocrine neoplasms (NENs) are rare and heterogeneous malignancies requiring multidisciplinary management. Large language models (LLMs) are emerging as decision-support tools, but their role in therapeutic decision-making is largely unexplored. ARTEMIS was a pilot cross-sectional study comparing three configurations—a baseline GPT, a customised GPT with static domain knowledge (GPTs), and a retrieval-augmented GPT (RAG)—against a panel of nine Italian NEN experts using twenty simulated, non-surgical cases. The primary endpoint was non-inferiority for systemic therapy recommendations; secondary endpoints included completeness, explicit uncertainty, parsimony of additional tests, costs, and variability metrics. RAG and GPTs achieved 70.0% agreement versus the expert benchmark (63.8%), meeting the exploratory −10% non-inferiority margin but not the stricter −5% threshold. Baseline GPT reached 60.0% and was not non-inferior. All AI systems consistently produced complete recommendations and expressed uncertainty more often than experts; RAG tended to propose fewer additional tests and lower associated costs. Experts showed greater variability than AI systems, and Ki-67 correlated with disagreement, indicating biological aggressiveness as a source of uncertainty. This exploratory study suggests that LLMs can approximate expert therapeutic reasoning under controlled conditions, but concordance remains limited and external validation in real-world settings is needed before clinical use.

Neuroendocrine neoplasms (NENs) are uncommon but increasingly frequent malignancies. U.S. SEER data show a > 5-fold rise in incidence from 1.64 per 100,000 in 1975 to 8.52 per 100,000 in 2021, underscoring their growing burden[1]. Their biological heterogeneity—across grade/Ki-67, site, receptor status, and metastatic pattern—makes management complex. Guidelines from ENETS, ESMO, and NANETS emphasise multidisciplinary evaluation and evidence-based systemic strategies including somatostatin analogues, targeted agents, peptide receptor radionuclide therapy, and cytotoxic chemotherapy[2–10].

AI has rapidly expanded in oncology decision support. Large language models (LLMs) have progressed from knowledge-testing to retrieval-augmented generation (RAG) and domain-specific fine-tuning. Systematic reviews show dozens of LLM applications in cancer care, but designs and endpoints remain heterogeneous, often limited to feasibility or exam-style tasks[11,12].

In NENs, AI research has focused almost exclusively on diagnostics: segmentation of tumour burden on PET/CT, radiomics-based grade prediction, and digital Ki-67 scoring[13–15]. Additional modalities include AI applied to endoscopic ultrasound (EUS) images for grading pancreatic neuroendocrine tumours, and deep learning on digital pathology slides for pulmonary NEN subtyping[16,17]. Multimodal approaches integrating computational pathology and radiomics have improved prediction of post-operative liver metastases[18]. Despite promise, these studies are retrospective, small, and rarely externally validated.

Therapeutic decision support remains scarcely studied. An Italian study compared LLMs for patient and clinician queries, finding good clarity

A full list of affiliations appears at the end of the paper. ✉e-mail: francesco.panzuto@uniroma1.it

but variable accuracy and poor performance on chemotherapy questions[19]. No NEN studies have benchmarked AI-generated therapy recommendations against expert consensus using predefined non-inferiority margins.

Research gap. No published studies have systematically tested whether LLMs can achieve therapeutic decisions comparable with expert consensus under prespecified non-inferiority thresholds. Key constructs such as explicit uncertainty, completeness, and parsimony of resource use remain underexplored. The influence of biological aggressiveness (grade) and site of origin on human–AI divergence is also unknown.

Objective. ARTEMIS evaluated whether three LLM configurations (baseline GPT, GPTs, and RAG) could deliver systemic therapy recommendations comparable to a multidisciplinary expert panel, with secondary aims to characterise performance across completeness, tests, uncertainty, cost, and variability.

Results

Baseline characteristics

Twenty simulated case vignettes of patients with NENs were included. Median age was 66 years (range 53–83). ECOG performance status was 0 in 17 cases (85%), 1 in two (10%), and 2 in one (5%). Primary tumour sites were gastroenteropancreatic (GEP, $n = 15$, 75%), lung ($n = 4$, 20%), and carcinoma of unknown primary (CUP; $n = 1$, 5%). Most cases were stage IV with distant metastases at evaluation ($n = 17$, 85%), with three stage III cases (15%). Grade distribution included NET G1 ($n = 7$, 35%), NET G2 ($n = 10$, 50%), NEC ($n = 2$, 10%), and one case with missing data (5%). Treatment history was balanced across first ($n = 7$, 35%), second ($n = 7$, 35%), and third or subsequent lines ($n = 6$, 30%). Somatostatin receptor positivity was present in 15 cases (75%). Table 1 summarises baseline characteristics of virtual patients.

Table 1 | Baseline characteristics of the 20 simulated clinical cases

Characteristic	Category	N (%) / Median (range)
Age, years	–	66 (53–83)
ECOG performance status	0	17 (85%)
	1	2 (10%)
	2	1 (5%)
Primary tumour site	GEP	15 (75%)
	Lung	4 (20%)
	CUP	1 (5%)
Stage	III	3 (15%)
	IV	17 (85%)
Grade	NET G1	7 (35%)
	NET G2	10 (50%)
	NEC	2 (10%)
	ND	1 (5%)
Treatment line	First	7 (35%)
	Second	7 (35%)
	Third or subsequent	6 (30%)
Somatostatin receptor status	Positive	15 (75%)
	Negative	5 (25%)

GEP gastro-entero-pancreatic, *CUP* carcinoma of unknown primary, *NET* Neuroendocrine Tumour, *NEC* Neuroendocrine Carcinoma.

Primary endpoint

The primary endpoint was to demonstrate the non-inferiority of AI-based recommendations compared with human experts in systemic therapy decisions. As shown in Table 2, the mean agreement of individual experts with the panel reference (human benchmark) was 63.8% (95% CI 56.5–71.0). Against this benchmark, baseline GPT achieved 60.0% (95% CI 44.4–76.9), GPTs 70.0% (95% CI 54.5–85.7), and the retrieval-augmented model (RAG) 70.0% (95% CI 54.5–84.6). The estimated differences (Δ = p_AI − p_H) were −3.8 pp for GPT, and +6.2 pp for both GPTs and RAG. Based on the lower bound of the one-sided 95% bootstrap confidence interval, GPT did not meet non-inferiority, whereas GPTs and RAG satisfied the –10% but not the –5% margin (Fig. 1).

Secondary endpoints

Secondary outcomes assessed parsimony of diagnostic work-up, explicit uncertainty, completeness of recommendations, and estimated costs. Summary results are presented in Table 3.

Parsimony, defined as agreement with the expert benchmark on not recommending additional tests, revealed marked differences across models. Experts themselves showed an AC1 of 0.443 (95% CI 0.313–0.575), indicating residual variability despite the modal response being "no additional test" in most cases. Among AI systems, RAG demonstrated the highest intra-run consistency (AC1 0.946, 95% CI 0.817–1.000), maintaining identical outputs across repeated generations and achieving significantly higher agreement than experts (ΔAC1 + 0.50, 95% CI 0.30–0.62; $p < 0.001$). GPTs more frequently introduced additional investigations (AC1 0.370, 95% CI 0.163–0.574), whereas baseline GPT achieved a slightly higher level of agreement (AC1 0.482, 95% CI 0.277–0.682). Overall, intra-LLM reproducibility ranked RAG > GPT > GPTs, whereas intra-expert agreement remained lower, confirming that RAG not only aligned with expert parsimony but did so with greater stability across repeated runs.

Costs, defined as concordance with the expert benchmark regarding the absence of additional resource use, showed clear divergence between models. Experts displayed low internal agreement (AC1 0.323, 95% CI 0.247–0.403), indicating variability in test-related cost assessments. Among AI systems, RAG achieved the highest intra-run consistency with the expert benchmark (AC1 0.925, 95% CI 0.811–1.000), maintaining stable low-cost outputs across repeated generations and reaching significantly higher agreement than experts (ΔAC1 + 0.60, 95% CI 0.45–0.70; $p < 0.001$). In contrast, both baseline GPT (AC1 0.287, 95% CI 0.126–0.451) and GPTs (AC1 0.309, 95% CI 0.151–0.471) frequently introduced additional investigations, with AC1 values indicating partial agreement with experts. Intra-LLM reproducibility confirmed RAG's stability, with consistently narrow bootstrap confidence intervals, whereas GPT and GPTs showed greater variability. Overall, reproducibility ranked RAG > GPT ≈ GPTs, while

Table 2 | Agreement between AI systems and human experts for systemic therapy recommendations, with bootstrap-based non-inferiority analysis

Model	p_H (95% CI)	p_AI (95% CI)	Δ	LB95 (1-sided)	NI δ = 5%	NI δ = 10%
GPT	0.638 (0.565–0.711)	0.600 (0.444–0.769)	−0.038	−0.174	No	No
GPTs	0.638 (0.565–0.710)	0.700 (0.545–0.857)	0.062	−0.062	No	**Yes**
RAG	0.638 (0.567–0.714)	0.700 (0.545–0.846)	0.062	−0.058	No	**Yes**

p_H leave-one-out agreement among human experts (benchmark), *p_AI* agreement between AI system and expert consensus, Δ difference (*p_AI* – *p_H*), *LB95* lower bound of the one-sided 95% bootstrap confidence interval for Δ, *NI δ* predefined non-inferiority margin (5% or 10%). Non-inferiority is concluded when $LB95 \geq -\delta$.

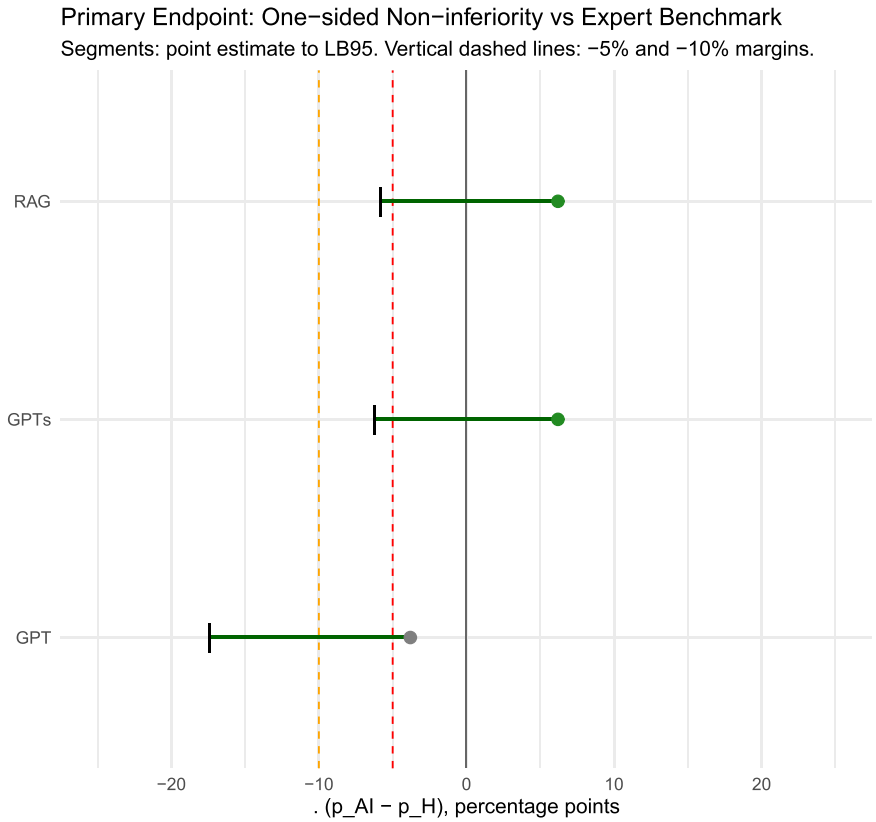

Fig. 1 | Primary endpoint: one-sided non-inferiority of AI systems versus human experts in systemic therapy recommendations. The figure shows the difference in agreement (Δ = p_AI − p_H) between each AI model and the expert benchmark. Segments represent the point estimate and the lower bound of the one-sided 95% bootstrap confidence interval (LB95). Vertical dashed lines indicate the prespecified non-inferiority margins of –5% (red) and –10% (orange). Non-inferiority is concluded when the LB95 lies to the right of the chosen margin.

intra-expert agreement remained lower, underscoring RAG's ability to replicate expert parsimony in resource use with greater stability.

Uncertainty was more frequently acknowledged by AI than experts. Expert responses demonstrated limited overlap, with low agreement on uncertainty assessments (AC1 0.239, 95% CI 0.103–0.394). By contrast, GPTs consistently expressed uncertainty across repeated runs (AC1 1.000, 95% CI 1.000–1.000), reflecting perfect intra-run reproducibility and significantly higher concordance than experts (ΔAC1 + 0.76, 95% CI 0.59–0.90; $p < 0.001$). RAG also achieved high reproducibility (AC1 0.875, 95% CI 0.670–1.000), while baseline GPT showed intermediate stability (AC1 0.835, 95% CI 0.642–0.966). Overall, intra-LLM reproducibility ranked GPTs > RAG > GPT, underscoring that both retrieval augmentation and static knowledge integration enhanced the explicit and consistent expression of uncertainty compared with expert assessments.

Completeness scores reached the maximum value (3/3: therapy, follow-up, and tests) in all AI generations. GPTs, RAG, and baseline GPT each achieved perfect intra-run reproducibility (AC1 1.000, 95% CI 1.000–1.000). By contrast, experts rarely provided fully comprehensive recommendations (AC1 0.164, 95% CI 0.095–0.231). The difference in concordance between AI systems and experts was highly significant (ΔAC1 + 0.84, 95% CI 0.70–0.90; $p < 0.001$). Overall, intra-LLM reproducibility was uniformly maximal and clearly superior to intra-expert agreement, confirming that completeness was an intrinsic feature of the structured AI outputs rather than of human recommendations.

Entropy distributions for secondary outcomes are visualised in Fig. 2, illustrating higher dispersion in expert decisions compared with the stable outputs of RAG.

Subgroup analyses

Subgroup analyses explored associations between case-level clinical features and measures of decision variability (entropy), agreement (Gwet's AC1), and predictors of AI–expert discordance in systemic therapy decisions (Supplementary Information, Table 4, and Fig. 3). In univariate analyses, no significant associations were observed between entropy and demographic or clinical variables such as age, ECOG status, disease stage, treatment line, or receptor status (all $p > 0.3$). Grade showed a positive correlation with entropy ($r = 0.510$, 95% CI 0.073–0.783; $p = 0.026$) and a negative correlation with AC1 ($r = –0.573$, 95% CI –0.815 to –0.161; $p = 0.010$). Given the small subgroup size, these findings should be regarded as exploratory and interpreted with caution.

Higher reported uncertainty among experts was also associated with lower agreement ($r = –0.595$; $p = 0.006$), suggesting that cases perceived as ambiguous were those with the greatest divergence in therapeutic choices. In multivariable logistic regression models (Fig. 4), higher perceived uncertainty remained a significant predictor of disagreement for baseline GPT (OR 13.99; $p = 0.032$), while grade showed a trend towards increased discordance (OR 7.32; $p = 0.068$) and lung primary site approached significance (OR 8.26; $p = 0.103$). For GPTs, discordance was primarily related to lung primaries (OR 11.90; $p = 0.060$). For RAG, lung primary was significantly associated with discordance (OR 19.50; $p = 0.032$) and showed a trend toward higher costs (OR 1.005; $p = 0.095$). Given the very small subgroup ($n = 4$), these findings should be regarded as hypothesis-generating and not overgeneralised.

Sensitivity analyses assessed intra-run reproducibility across repeated generations on the same cases (Supplementary Information). For systemic therapy, RAG achieved perfect stability (AC1 1.000, 95% CI 1.000–1.000; 100% agreement). Baseline GPT was more reproducible than GPTs (AC1 0.843, 95% CI 0.687–0.961; 80.1%, 95% CI 60–95 vs 69.9%, 95% CI 50–90), yielding the ranking RAG > GPT > GPTs. For secondary endpoints, intra-run reproducibility of AI models remained consistently high, particularly for RAG (tests: AC1 0.946, 95% CI 0.817–1.000; costs: AC1 0.925, 95% CI 0.811–1.000), whereas GPT and GPTs showed intermediate stability. In contrast, inter-expert agreement was low across all dimensions (all AC1 < 0.45). These findings indicate that LLMs were substantially more reproducible across repeated runs than human experts evaluating the same cases.

Table 3 | Secondary endpoints: agreement between AI models and human experts across clinical dimensions (Bootstrap analysis, 95% CI)

Outcome	Experts AC1 (95% CI)	GPT AC1 (95% CI)	GPTs AC1 (95% CI)	RAG AC1 (95% CI)	ΔAC1 vs Experts (95% CI)	p-value (bootstrap)
Parsimony	0.443 (0.313–0.575)	0.482 (0.277–0.682)	0.370 (0.163–0.574)	0.946 (0.817–1.000)	RAG–Experts: +0.50 (0.30–0.62)	<0.001
Costs	0.323 (0.247–0.403)	0.287 (0.126–0.451)	0.309 (0.151–0.471)	0.925 (0.811–1.000)	RAG–Experts: +0.60 (0.45–0.70)	<0.001
Uncertainty	0.239 (0.103–0.394)	0.835 (0.642–0.966)	1.000 (1.000–1.000)	0.875 (0.670–1.000)	GPT–Experts: +0.60 (0.40–0.75)	<0.001
Completeness	0.164 (0.095–0.231)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	All AI–Experts: +0.84 (0.70–0.90)	<0.001

Gwet's AC1 with 95% bootstrap confidence intervals. "ΔAC1 vs Experts" shows the difference between AI and expert agreement with bias-corrected 95% CI from 1000 bootstrap resamples. p-values from bootstrap distribution of ΔAC1.

Discussion

In ARTEMIS, the primary endpoint showed that retrieval-augmented and customised GPT configurations achieved agreement rates of 70.0%, meeting the prespecified −10% non-inferiority margin but not the stricter −5% threshold. The baseline GPT reached 60.0% and did not achieve non-inferiority. This study provides the first structured evidence that large language models can approximate expert reasoning in systemic therapy recommendations for NENs.

In the broader oncology context, our agreement rates (~70%) are substantially higher than those reported in real-world breast cancer tumour boards (16%) and the moderate ratings documented in glioma adjuvant therapy decision-making (median overall agreement 5/10)[20,21]. Other experiences confirm a wide performance spectrum: in colorectal cancer, one MDT-based evaluation reported an agreement rate of approximately 87%, although within a single-centre setting and using permissive definitions of concordance[22], while in precision oncology tasks LLMs showed markedly low accuracy (F1 0.04–0.19 vs expert benchmark) when confronted with complex biomarker-driven decisions[23].

Systematic reviews and meta-analyses place pooled LLM accuracy in oncology around 70–76% but highlight substantial heterogeneity and a lack of evaluation of safety, clarity, or clinical harm[24]. Cautionary reports also emphasise risks of misinformation in patient-facing contexts, where treatment suggestions were sometimes inaccurate or guideline-incongruent.

Rare tumours represent a particularly challenging setting for AI-assisted decision-making. In sarcoma, LLMs provided highly variable outputs, with frequent inaccuracies in treatment recommendations and only partial alignment with expert standards, underscoring the difficulties of applying general-purpose models to low-incidence and heterogeneous malignancies[25]. Specifically, in NENs, Panzuto et al. systematically evaluated responses from three LLMs to clinically relevant queries. While the clarity of responses was often judged positively, accuracy and completeness were inconsistent, and chemotherapy-related questions yielded the lowest scores[19]. These findings highlight both the promise and the limitations of current LLMs in rare cancers: they can serve as accessible informational resources but remain unreliable for complex therapeutic decisions without expert oversight. By contrast, ARTEMIS demonstrated non-inferiority in systemic therapy recommendations for NENs, a result that may be explained by the preparatory adaptation of the GPTs and RAG models. Unlike general-purpose configurations, these systems were either pre-loaded with guideline-based content and real-world clinical cases or dynamically retrieved validated sources, thereby reducing the risk of hallucinations and improving concordance with expert reasoning. None of these previous studies applied a formal non-inferiority framework, making ARTEMIS the first systematic benchmarking of LLMs against expert consensus with predefined statistical margins ($\Delta = 10\%$ exploratory, unmet at $\Delta = 5\%$). This is particularly relevant in rare tumours, including NENs, where therapeutic choices are often specialist-dependent due to less stringent guidelines and limited evidence. In this setting, LLMs may be simultaneously more challenged and more valuable, especially for non-expert clinicians.

No previous LLM studies have formally assessed intra-run reproducibility in clinical decision-making tasks. ARTEMIS therefore provides the first evidence that repeated generations from the same model yield highly consistent therapeutic recommendations, particularly for the RAG configuration (AC1 1.000) and, to a slightly lesser extent, for GPT and GPTs (AC1 0.843 and 0.766, respectively).

By contrast, concordance among human experts on the same cases was much lower (all AC1 < 0.45), reflecting heterogeneity of clinical judgement in rare and ambiguous scenarios rather than inconsistency within individuals. This variability should not be interpreted as error but as an inherent feature of complex oncology decision-making, where evidence is limited and guidelines allow discretion.

To provide a unique and reproducible comparator for non-inferiority testing, ARTEMIS adopted the per-case modal expert response as the

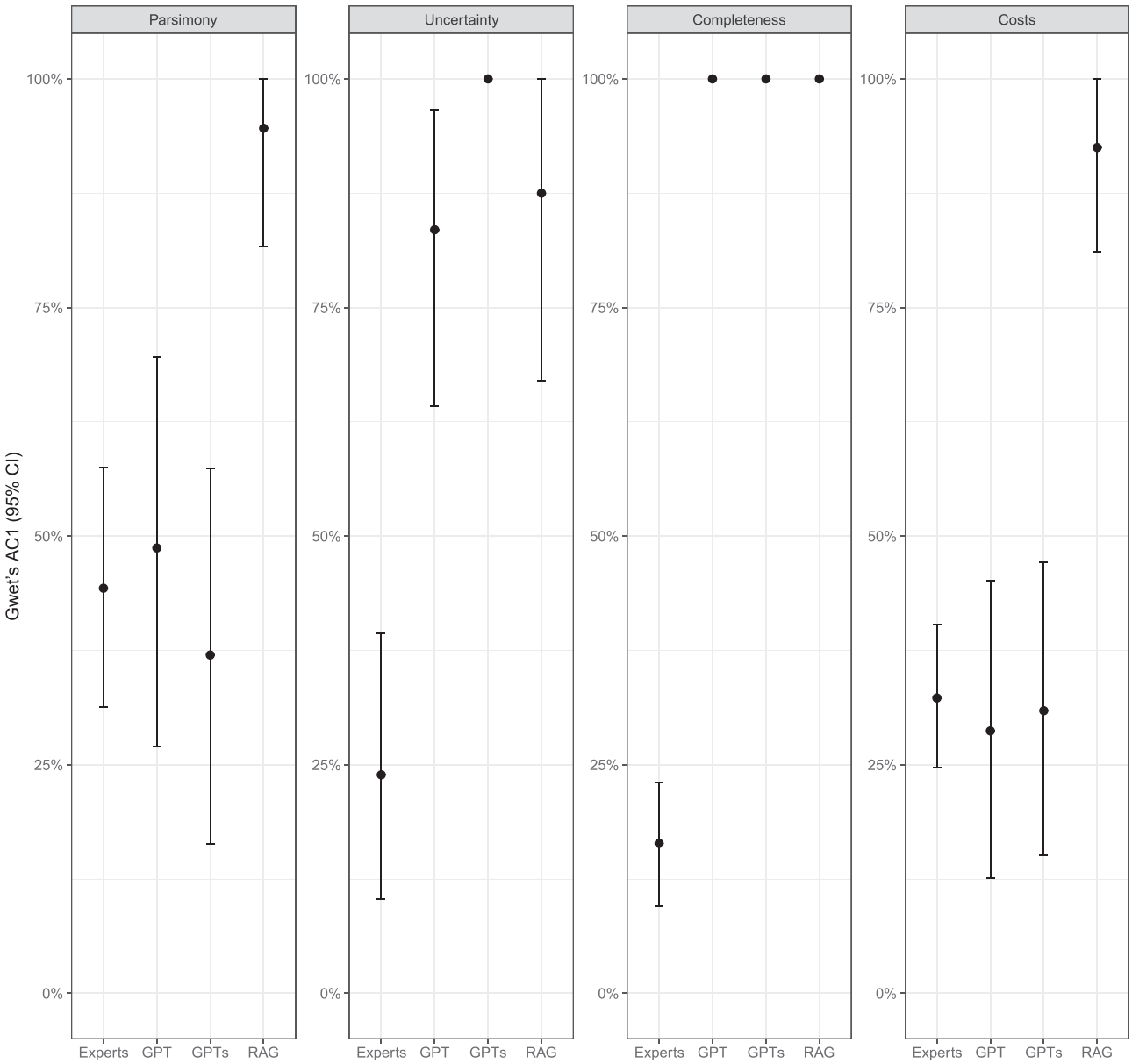

Fig. 2 | Agreement of AI systems and experts across clinical dimensions. Gwet's AC1 with 95% bootstrap confidence intervals for experts, baseline GPT, customised static GPT (GPTs), and retrieval-augmented GPT (RAG). Outcomes assessed were parsimony (agreement on not recommending additional tests), explicit uncertainty, completeness (therapy, follow-up, and tests), and costs (absence of additional resource use). AI systems consistently achieved higher reproducibility than experts, with RAG showing the greatest stability across most domains.

Table 4 | Multivariable logistic regression models for predictors of disagreement between AI systems and experts

Variable	GPT OR (95% CI), p	GPTs OR (95% CI), p	RAG OR (95% CI), p
Grade	7.32 (0.90–59.4), $p = 0.068$	NR	NR
Lung	8.26 (0.67–101.7), $p = 0.103$	11.90 (0.95–149.5), $p = 0.060$	19.50 (1.32–287.3), $p = 0.032$
Uncertainty	13.99 (1.25–156.5), $p = 0.032$	NR	NR
Cost	NR	NR	1.005 (0.999–1.012), $p = 0.095$

Odds ratios (OR) with 95% CI and p-values from multivariable logistic regression models. Variables were retained if $p \leq 0.15$ in univariate screening; "NR" indicates exclusion above model-specific thresholds.

operational benchmark, thereby capturing real-world variability while avoiding an unattainable 100% consensus. This likely reflects the tendency of specialists to rely on personal experience and clinical reasoning when data are scarce or ambiguous. Agreement metrics therefore, assess concordance rather than correctness; variability among experts reflects legitimate divergence in complex cases, consistent with prior reports of limited inter-rater agreement in rare tumour management.

These findings imply that, for customised and retrieval-augmented systems, repeating the same query is unlikely to add value, since a single generation already provides a stable and reproducible output—a practical

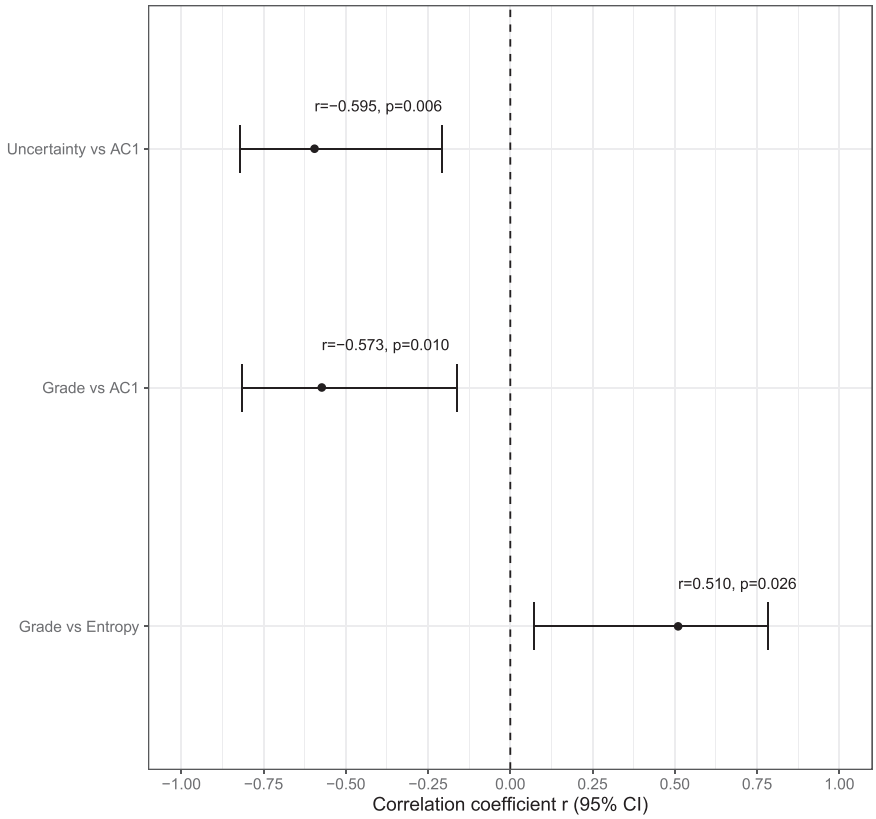

Fig. 3 | Correlations between clinical variables and expert–AI agreement. Pearson correlation coefficients (r) with 95% confidence intervals are shown for associations between Ki-67 proliferation index or expert-reported uncertainty and concordance metrics. Higher Ki-67 was significantly associated with lower agreement (AC1; $r = -0.573$, $p = 0.010$) and greater variability (entropy; $r = 0.510$, $p = 0.026$). Expert-reported uncertainty also correlated negatively with agreement (AC1; $r = -0.595$, $p = 0.006$).

advantage that may reduce redundancy and enhance confidence in AI-assisted clinical workflows. In this respect, ARTEMIS contrasts with prior NEN-specific work by Panzuto et al., who compared general-purpose LLMs (ChatGPT Plus, Copilot, Perplexity) for patient and clinician queries and reported heterogeneous accuracy with poor performance on systemic therapy questions[19]. That study did not include reproducibility analyses or retrieval augmentation, underscoring the inconsistent quality of outputs in rare tumours. Against this background, ARTEMIS provides a structured framework showing that customised and retrieval-augmented systems can mitigate variability and achieve reproducible concordance with expert consensus. Among AI systems, RAG showed the highest reproducibility and closer alignment with expert benchmarks, while GPTs was most consistent in explicitly expressing uncertainty. All AI models systematically achieved complete recommendations, contrasting with frequent omissions by experts. These results highlight retrieval augmentation as a determinant of stability and static knowledge integration as a driver of uncertainty disclosure.

Exploratory analyses indicated that higher Ki-67 was associated with greater variability and lower concordance, while lung primaries predicted disagreement, especially for RAG. The observed correlation between grade and entropy should be considered exploratory, reflecting limited sample size rather than a validated biological association. No significant effects were observed for age, ECOG, stage, treatment line, or receptor status. Expert-reported uncertainty also correlated with lower concordance, suggesting that divergence arises particularly in intrinsically ambiguous cases. To our knowledge, no previous studies have provided quantitative benchmarks for intra-run reproducibility, completeness, explicit uncertainty, or parsimony of therapeutic recommendations in NENs. Accordingly, the secondary endpoints of ARTEMIS cannot be directly compared with existing literature, underscoring the novelty of this analysis and its role in defining methodological standards for future research.

Taken together, these findings show that ARTEMIS is the first to quantify reproducibility, completeness, and uncertainty in therapeutic decision support. While AI outputs can exceed human performance in consistency and coverage, they remain sensitive to biological aggressiveness and poorly standardised clinical contexts, underscoring their role as complementary rather than substitutive tools in rare cancers.

These results also illustrate the potential of ARTEMIS as an expert-in-the-loop clinical decision-support framework. Rather than replacing human judgement, retrieval-augmented and customised LLMs can assist multidisciplinary teams by generating consistent, guideline-based recommendations, enhancing completeness, and reducing redundant investigations. This approach aligns with contemporary perspectives on AI-assisted decision support, which emphasise collaboration between clinicians and adaptive algorithms to improve quality and safety of care[26,27].

This study has several limitations. First, its exploratory design relied on simulated clinical cases rather than real patient data, and case selection, although validated by a multidisciplinary team, may still introduce bias. Second, the small expert panel meant that consensus was sometimes based on fewer than 55% of participants. Low inter-expert completeness (AC1 < 0.2) reflected optional reporting fields rather than omissions; experts could omit components considered clinically irrelevant. No post-hoc questionnaire was administered. Third, outputs were generated under controlled prompting and domain-constrained conditions, which may not reflect real-world clinical use. Fourth, the study was not preregistered, and external validation was not performed. These factors limit generalisability and call for caution in interpreting the results. Nevertheless, ARTEMIS establishes a reproducible framework for benchmarking LLMs against experts with prespecified non-inferiority margins and multidimensional endpoints. Future research should validate these findings in prospective, multicentre cohorts of real patients, incorporate diverse LLM architectures, and assess the clinical impact of AI assistance on therapeutic outcomes.

In conclusion, ARTEMIS provides the first structured evidence that LLMs can approximate expert reasoning in systemic therapy decisions for NENs. RAG and GPTs configurations achieved non-inferiority at the prespecified −10% margin, while the baseline GPT did not. These findings indicate that adapted LLMs can deliver guideline-concordant therapeutic

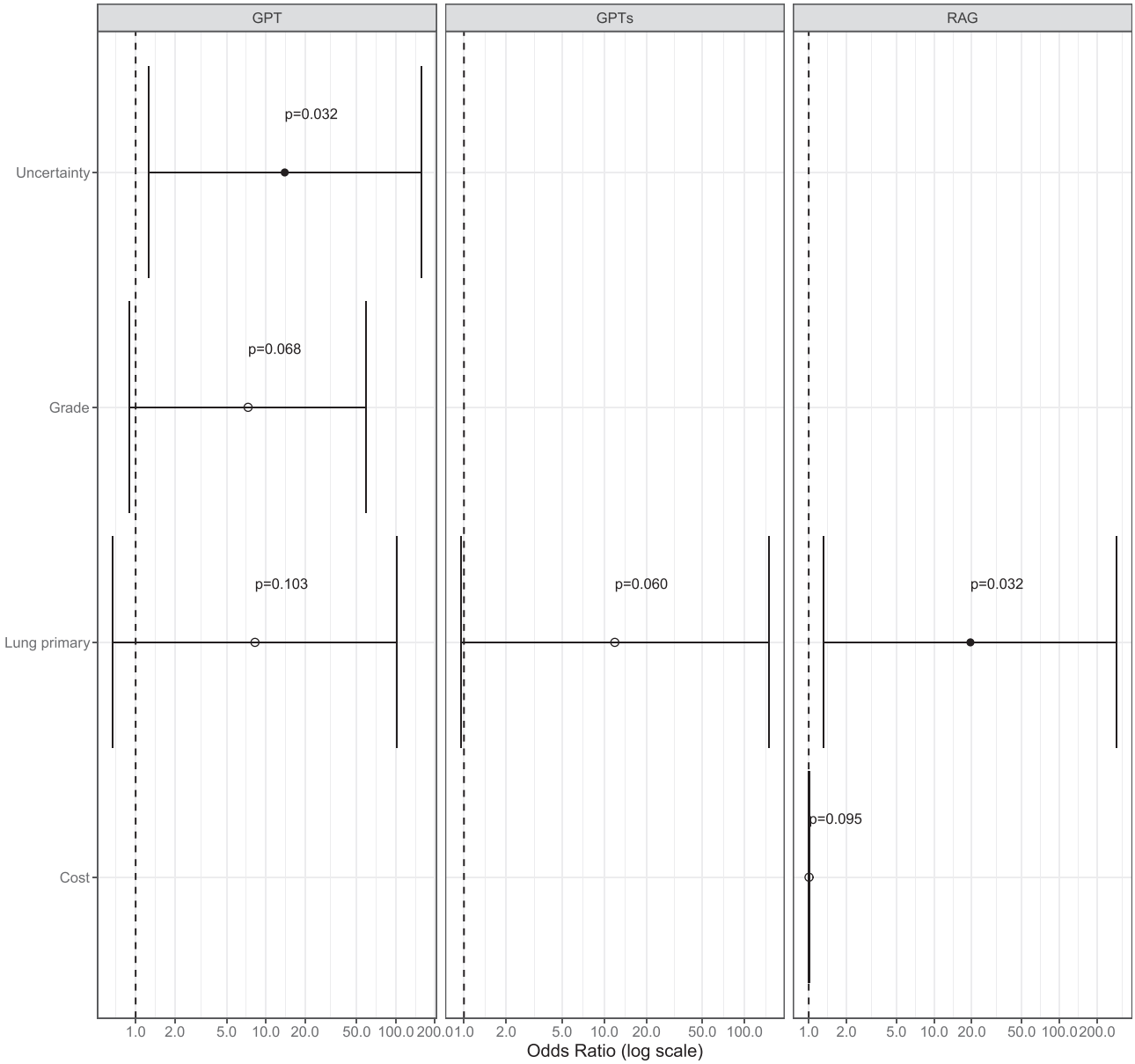

Fig. 4 | Multivariable predictors of disagreement between AI models and experts. Forest plot showing odds ratios (log scale) with 95% confidence intervals from multivariable logistic regression models. For baseline GPT, higher perceived uncertainty predicted disagreement (OR 13.99, $p = 0.032$), with trends for Ki-67 (OR 7.32, $p = 0.068$) and lung primaries (OR 8.26, $p = 0.103$). For GPTs, lung primaries approached significance (OR 11.90, $p = 0.060$). For RAG, lung primaries significantly predicted disagreement (OR 19.50, $p = 0.032$), with a trend for higher costs (OR 1.005, $p = 0.095$).

recommendations close to expert consensus, although validation in prospective, real-world studies remains essential before clinical use.

Methods

Study design and setting

ARTEMIS was a methodological, cross-sectional evaluation of LLMs for decision support in NEN. Standardised case vignettes were developed by the study team and subsequently evaluated independently by clinical experts, who provided the human benchmark. Three AI configurations were compared: baseline GPT-o3 (GPT), customised GPT with static domain knowledge (GPTs), and GPT with RAG. The primary endpoint was prespecified as non-inferiority of AI compared with human experts for systemic therapy recommendations. Development of AI systems occurred from March to June 2025, vignettes from May to June 2025, expert ratings in June–July 2025, and AI generations in August 2025.

Participants and cases

Nine Italian NEN specialists (five oncologists, two endocrinologists, two gastroenterologists) provided blinded and independent recommendations. Twenty simulated NEN cases were synthetically generated and validated by a multidisciplinary tumour board (MDT) comprising oncologists, surgeons, pathologists, radiologists, and nuclear medicine physicians. Each case followed predefined clinical rules derived from ENETS and AIOM guidelines to ensure realism in tumour grade, stage, and receptor profile (Table 1). The final cohort reproduced typical distributions in advanced NENs (stage IV ≈ 85%, G1/G2 ≈ 85%, SSTR-positive ≈ 75%). No external patient data were used, and no modelling of real-world predictors was performed. Each vignette included demographics, ECOG performance status, stage, prior therapies, imaging, biochemistry, grade, receptor status, and site of origin, and concluded with the same open therapeutic question: "the systemic treatment to be proposed". All cases were independently validated (ie, reviewed for clinical plausibility and completeness) by a multidisciplinary

team at the NENs Outpatient Clinic of IRCCS Azienda Ospedaliero-Universitaria di Bologna - Policlinico di Sant'Orsola (EURACAN–ERN reference center). For each case and endpoint, the human benchmark was defined as the modal expert response, even when it did not reach absolute majority consensus. This approach ensured that all cases had a unique reference for AI–expert comparisons and non-inferiority testing.

AI systems and prompting

No gradient-based fine-tuning was performed. GPT and GPTs used OpenAI o3, whereas RAG employed o3-mini, which was appropriate given its retrieval-constrained, guideline-anchored outputs. GPTs operated in closed-book mode with a static domain corpus comprising ENETS and AIOM guidelines, drug monographs, and 200 de-identified MDT cases, distinct from the 20 evaluation vignettes. RAG accessed the same corpus dynamically through an EU-hosted retrieval pipeline, injecting only verified documents into context. All configurations received an identical structured prompt requesting a single guideline-aligned recommendation (therapy, follow-up, additional tests, rationale, and citation). For each of the 20 simulated cases, every AI system was prompted three independent times, and the modal recommendation across runs was used as the representative output for comparison with the expert benchmark. All generations were executed with temperature = 0.2, ensuring low stochasticity and near-deterministic outputs. Governance details are provided in the Supplement, including corpus versioning, encrypted embeddings with hash-based indexing, role-based access controls, and confirmation of complete de-identification for all cases.

Endpoints

The primary endpoint was systemic therapy recommendation, tested for non-inferiority of AI compared with human experts. Two margins were pre-specified (–5% conservative, –10% exploratory), reflecting thresholds previously applied in comparable AI–clinician evaluations and consistent with EMA/FDA/CONSORT guidance, which does not define fixed cut-offs but requires margins to be clinically and statistically justified a priori[28,29]. Secondary endpoints included completeness (0–3 scale: 0 = incoherent/absent; 1 = therapy only; 2 = therapy + follow-up; 3 = therapy + follow-up + tests), explicit uncertainty, parsimony of additional tests, incremental cost of tests, and variability metrics (intra-run, inter-expert, inter-model). Incremental cost was defined as the sum of regional reimbursement tariffs for all additional tests proposed beyond the expert benchmark, according to the official *Nomenclatore Tariffario Regionale*. Concordance was assessed against the modal expert response, which most often corresponded to no additional test.

Data collection and quality control

Experts completed structured Google Forms combining fixed-choice and free-text fields. Responses were exported into an Excel database with double data entry and reconciliation. AI and expert outputs were anonymised, coded into a unified grid, and checked for internal consistency.

Statistical analysis

Analyses were prespecified and conducted in R (version 4.4.2, R Foundation for Statistical Computing, Vienna, Austria) with packages including dplyr, tidyr, purrr, tibble, ggplot2/tidyverse, irr, irrCAC, entropy, psych, ordinal, performance, boot, and lme4, using a fixed random seed (20250808). For each vignette, the human benchmark was defined as the modal expert response (the option with the highest number of votes among the nine panellists), applied even when fewer than five experts agreed, to guarantee a consistent reference across all cases. Baseline characteristics were summarised as counts and percentages for categorical variables, and as medians with ranges for continuous variables. Percentages were calculated over the full cohort ($N = 20$). No imputation was performed. Agreement was quantified using percent agreement (reported in the Supplement) and Gwet's AC1 with 95% bootstrap CIs. Dispersion was summarised by Shannon entropy. We preferred AC1 to Cohen's κ given known κ instabilities under high agreement or prevalence imbalance[30,31]. For agreement metrics, bootstrap analyses (1000 iterations) resampled cases with replacement, preserving all responses for each case and system. Bias-corrected 95% confidence intervals were derived. The wide ranges observed are attributable to N = 20 and to the inherent variability of therapeutic decision-making. Exploratory analyses assessed correlations between entropy and clinical covariates (Pearson for continuous variables, Kruskal–Wallis for categorical factors) and multivariable linear regression models with entropy as outcome. Univariate logistic regression models were fitted for each AI system with therapy disagreement as outcome and predictors considered individually (uncertainty, tests, cost, completeness, age, ECOG, stage, treatment line, grade, receptor status, organ). For the primary non-inferiority analysis, the reference standard was the human benchmark defined above. For each AI system, the proportion of cases in agreement with this benchmark was denoted p_AI, and the average proportion of individual experts agreeing with the same benchmark was denoted p_H. The difference $\Delta = p\_AI - p\_H$ was estimated for each model using 2000 bootstrap resamples. The lower bound of the one-sided 95% bootstrap confidence interval (LB95) for Δ was compared with predefined non-inferiority margins ($\delta = 5\%$ and 10%). AI was considered non-inferior when $LB95 \geq -\delta$, meaning that the lower limit of the estimated difference did not fall more than δ below the average human performance.

The analysis codes will be openly available upon publication on Zenodo.

Ethics approval and consent to participate

The study used exclusively simulated clinical cases and did not involve human participants, patient data, or identifiable information. According to the policies of the Ethics Committee of IRCCS Azienda Ospedaliero-Universitaria di Bologna, studies based solely on fully synthetic, non-patient-derived scenarios do not require ethics committee approval or informed consent. No ethics approval or consent was therefore required.

Data availability

This study used exclusively synthetic, fully de-identified clinical vignettes, which are not intended to reproduce real patient data and are not required for analytical reproducibility. The materials necessary to reproduce the statistical analyses and AI system configuration will be openly available upon publication in a public repository (Zenodo), including: (i) the complete R code for data processing, statistical analyses, and figure generation; (ii) the full prompt templates and model configuration files used for all AI systems; and (iii) the extended documentation of the retrieval-augmented generation (RAG) pipeline, including corpus structure, preprocessing steps, and embedding/metadata specifications. All resources will be deposited under an open license and assigned a permanent DOI (https://doi.org/10.5281/zenodo.17965643). No patient data were used in this study.

Code availability

All analysis scripts were written in R (version 4.4.2). The complete and annotated code will be released on Zenodo together with the supplementary materials upon publication.

Received: 14 October 2025; Accepted: 10 December 2025;
Published online: 23 December 2025

References

1. Dasari, A. et al. Epidemiology of neuroendocrine neoplasms in the US. *JAMA Netw. Open* **8**, e2515798 (2025).
2. Lamarca, A. et al. European Neuroendocrine Tumor Society (ENETS) 2024 guidance paper for the management of well-differentiated small intestine neuroendocrine tumours. *J. Neuroendocrinol.* **36**, e13423 (2024).

3. Rinke, A. et al. European Neuroendocrine Tumor Society (ENETS) 2023 guidance paper for colorectal neuroendocrine tumours. *J. Neuroendocrinol.* **35**, e3309 (2023).
4. Kos-Kudła, B. et al. European Neuroendocrine Tumour Society (ENETS) 2023 guidance paper for nonfunctioning pancreatic neuroendocrine tumours. *J. Neuroendocrinol.* **35**, e1334 (2023).
5. Sorbye, H. et al. European Neuroendocrine Tumor Society (ENETS) 2023 guidance paper for digestive neuroendocrine carcinoma. *J. Neuroendocrinol.* **35**, e13249 (2023).
6. Hofland, J. et al. European Neuroendocrine Tumor Society 2023 guidance paper for functioning pancreatic neuroendocrine tumour syndromes. *J. Neuroendocrinol.* **35**, e13318 (2023).
7. Panzuto, F. et al. European Neuroendocrine Tumor Society (ENETS) 2023 guidance paper for gastroduodenal neuroendocrine tumours (NETs) G1–G3. *J. Neuroendocrinol.* **35**, e13306 (2023).
8. Kaltsas, G. et al. European Neuroendocrine Tumor Society (ENETS) 2023 guidance paper for appendiceal neuroendocrine tumours (aNET). *J. Neuroendocrinol.* **35**, e13332 (2023).
9. Grozinsky-Glasberg, S. et al. European Neuroendocrine Tumor Society (ENETS) 2022 Guidance Paper for carcinoid syndrome and carcinoid heart disease. *J. Neuroendocrinol.* **34**, e13146 (2022).
10. Del Rivero, J., Kennedy, E. B. & Perez, K. Systemic therapy for tumor control in well-differentiated metastatic gastroenteropancreatic neuroendocrine tumors: ASCO Guideline Q and A. *JCO Oncol. Pract.* **19**, 955–958 (2023).
11. Hao, Y. et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digit. Med.* **8**, 450 (2025).
12. Carl, N. et al. Large language model use in clinical oncology. *NPJ Precis. Oncol.* **8**, 240 (2024).
13. Carlsen, E. A. et al. A convolutional neural network for total tumor segmentation in [64Cu]Cu-DOTATATE PET/CT of patients with neuroendocrine neoplasms. *EJNMMI Res.* **12**, 30 (2022).
14. Yan, Q. et al. Predicting histologic grades for pancreatic neuroendocrine tumors by radiologic image-based artificial intelligence: a systematic review and meta-analysis. *Front. Oncol.* **14**, 1332387 (2024).
15. Luo, Y. et al. Preoperative prediction of pancreatic neuroendocrine neoplasms grading based on enhanced computed tomography imaging: validation of deep learning with a convolutional neural network. *Neuroendocrinology* **110**, 338–350 (2020).
16. Kiremitci, S. et al. The role of artificial intelligence and deep learning in determining the histopathological grade of pancreatic neuroendocrine tumors by using EUS images. *Endosc. Ultrasound* **14**, 48–56 (2025).
17. Ilié, M. et al. Deep learning facilitates distinguishing histologic subtypes of pulmonary neuroendocrine tumors on digital whole-slide images. *Cancers* **14**, 1740 (2022).
18. Ma, M. et al. A novel model for predicting postoperative liver metastasis in R0 resected pancreatic neuroendocrine tumors: integrating computational pathology and deep learning-radiomics. *J. Transl. Med.* **22**, 768 (2024).
19. Panzuto, F. et al. Enhancing patient-centered care with AI: a study of responses to neuroendocrine neoplasms queries. *Endocrine* **89**, 921–929 (2025).
20. Lukac, S. et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch. Gynecol. Obstet.* **308**, 1831–1844 (2023).
21. Haemmerli, J. et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board?. *BMJ Health Care Inform.* **30**, e100775 (2023).
22. Choo, J. M. et al. Conversational artificial intelligence (chatGPT™) in the management of complex colorectal cancer patients: early experience. *ANZ J. Surg.* **94**, 356–361 (2024).
23. Benary, M. et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **6**, E2343689 (2023).
24. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
25. Valentini, M. et al. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients?. *Front. Public Health* **12**, 1303319 (2024).
26. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
27. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Dig. Med.* **3**, 41 (2020).
28. Oberije, C. J. G. et al. Assessing artificial intelligence in breast screening with stratified results on 306 839 mammograms across geographic regions, age, breast density and ethnicity: a Retrospective Investigation Evaluating Screening (ARIES) study. *BMJ Health Care Inform.* **32**, e101318 (2025).
29. Repici, A. et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* **159**, 512–520.e7 (2020).
30. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* **61**, 29–48 (2008).
31. Shannon, C. E. A. Mathematical Theory of Communication. *Bell Syst. Technical J* **27**, 379–423 (1948).

Acknowledgements

Generative AI (OpenAI GPT-5) was used exclusively to assist with English grammar and readability. No scientific content, analyses, or conclusions were generated by AI. No medical writer was involved. This research received no external funding. Publication costs were covered by Sapienza University Ateneo grant 2023 (RM123188F2A0536B). The funder had no role in study design; data collection, analysis, or interpretation; writing of the manuscript; or the decision to submit for publication.

Author contributions

All authors meet ICMJE authorship criteria. Concept and design: D.C., C.R., and G.L. Data access and verification: D.C. and C.R. Data curation: D.C. Formal analysis and visualisation: D.C. Investigation: D.C., G.L., F.P., S.M., M.C., A.S., F.S., A.F., S.P., M.A., S.T., and E.A. Methodology and software: D.C., C.R., and G.L. Supervision: D.C. Drafting of the manuscript: D.C. and G.L. Critical revision for important intellectual content: all authors. All authors had full access to all data and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02274-x.

Correspondence and requests for materials should be addressed to Francesco Panzuto.

[1]Department of Medical and Surgical Sciences (DIMEC), Alma Mater Studiorum - University of Bologna, Bologna, Italy. [2]Department of Medical-Surgical Sciences and Translational Medicine, Sapienza Università di Roma, Digestive Disease Unit, ENETS Center of Excellence, Sant'Andrea University Hospital, Rome, Italy. [3]Vita-Salute San Raffaele University, Faculty of Medicine and Surgery, Milan, Italy. [4]Interdisciplinary Department of Medicine, University of Bari "Aldo Moro", Bari, Italy. [5]Division of Medical Oncology, A.O.U. Consorziale Policlinico di Bari, Bari, Italy. [6]National Center for Drug Research and Evaluation, Istituto Superiore di Sanità, Rome, Italy. [7]Divisione di Oncologia Medica Gastrointestinale e Tumori Neuroendocrini, Istituto Europeo di Oncologia, IEO, IRCCS, Milano, Italy. [8]Endocrinology Unit, Department of Clinical and Molecular Medicine, Sant'Andrea University Hospital, ENETS Center of Excellence, Sapienza University of Rome, Rome, Italy. [9]Department of Medical Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, ENETS Center of Excellence, Milan, Italy. [10]Endocrinology Unit, Department of Internal Medicine and Medical Specialties (DiMI), University of Genova, Genova, Italy. [11]Endocrinology Unit, IRCCS Ospedale Policlinico San Martino, Genova, Italy. [12]Department of Sarcoma and Rare Tumors, Istituto Nazionale Tumori I.R.C.C.S. ENETS Center of Excellence Fondazione "G.Pascale", Naples, Italy.
✉e-mail: francesco.panzuto@uniroma1.it