Security, Safety, and Ethics of AI-driven Innovations for Digital Health - Security, Safety, and Ethics of AI-driven Innovati – Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 12: 1–10

DOI: 10.1177/20552076261419956
journals.sagepub.com/home/dhj

Sage

Generative artificial intelligence in palliative care: A comparative evaluation of ChatGPT-4o and ChatGPT-5 as clinical decision support tools

Emre Vuraloglu[1]

Abstract

Background and objective: Generative artificial intelligence (AI) tools such as ChatGPT are increasingly integrated into healthcare, with potential to support clinical decision-making and improve patient outcomes. In palliative care, where access to multidisciplinary expertise is often limited, these tools may provide support for symptom management. This study aimed to systematically compare ChatGPT-4o and ChatGPT-5 for common palliative care symptoms across four key domains: clinical appropriateness, safety, ethical sensitivity, and understandability.

Methods: Clinical scenarios representing 10 key symptoms (pain, anxiety, pressure ulcer, nausea, delirium, dyspnea, constipation, diarrhea, dry mouth, and sleep disturbance) were presented first to ChatGPT-4o and, 1 week later, to ChatGPT-5. Responses were evaluated independently by two physicians using a five-point Likert scale. Inter-rater agreement was analyzed with weighted Cohen's kappa and Spearman's correlation. The statistical analyses in this study were conducted using the Friedman test, Mann–Whitney U test, and Wilcoxon signed-rank test.

Results: Inter-rater agreement was consistently high across all domains (kappa 0.806–0.886, Spearman's rho 0.813–0.888; all $p < 0.001$). ChatGPT-5 outperformed ChatGPT-4o in clinical appropriateness ($p = 0.010$), safety ($p = 0.002$), and understandability ($p = 0.011$). Ethical sensitivity scores were high for both models, with no significant difference ($p = 0.102$).

Conclusions: ChatGPT-5 demonstrated measurable improvements over ChatGPT-4o in key domains of palliative care symptom management, while maintaining consistently high ethical sensitivity. These findings provide the first systematic evidence of the potential of generative AI, with the updated ChatGPT-5 model released in August 2025, as a complementary and reliable clinical decision support tool in palliative care.

Keywords

Artificial intelligence, ChatGPT, decision support systems, large language model, palliative care

Received: 22 August 2025; accepted: 14 January 2026

Introduction

Palliative care is a holistic model designed to improve the quality of life for individuals with life-threatening illnesses and their families.[1–3] It encompasses both psychosocial support and the relief of physical symptoms.[2,3] The management of symptoms such as pain, nausea, dyspnea, delirium, and anxiety is a cornerstone of palliative care, and effective control is essential for patient well-being.[1–3]

However, as in many countries, the delivery of palliative care services in Turkey faces several challenges.[4,5] There is no dedicated specialty for palliative care, and services are often provided through the support of family medicine, internal medicine, and geriatrics specialists.[4] The number of geriatricians is particularly limited and even absent in some regions. Moreover, multidisciplinary teams exclusively devoted to palliative care are not consistently

[1]Ahi Evran University Faculty of Medicine, Department of Family Medicine, Kirşehir Training and Research Hospital, Kirşehir, Turkey

Corresponding author:
Emre Vuraloglu, Kervansaray, 2019 Street No:1, 40200 Kirşehir City Center/Kirşehir, Turkey.
Email: emrevuraloglu@gmail.com

available across healthcare institutions.[4] These limitations hinder timely and accurate decision-making, especially in the management of acute symptoms, and highlight the need for alternative support solutions.

In this context, generative artificial intelligence (AI) systems based on large language models (LLMs) have recently emerged as transformative tools in healthcare, with applications in diagnostics, personalized treatment planning, and clinical decision support.[6–13] Among these, ChatGPT, developed by OpenAI, has gained particular prominence as one of the most widely accessible generative AI platforms. Initially released as GPT-3.5 and GPT-4, it has since advanced to ChatGPT-5, launched in August 2025, offering enhanced reasoning capacity, broader integration of medical knowledge, and potentially greater clinical relevance.[14–16]

Within palliative care, studies suggest that generative AI can provide accurate information, dispel misconceptions, and clarify terminology. Nonetheless, concerns persist regarding the consistency, quality, safety, and ethical implications of its recommendations.[14–17] To our knowledge, this is the first study to systematically compare the outputs of ChatGPT-4o and ChatGPT-5 in the context of palliative care symptom management. Although newer versions of ChatGPT might be expected to outperform their predecessors,[18,19] this assumption has not been systematically tested in palliative care, a domain in which safety and ethical considerations are particularly critical. While model generation updates may enhance overall reasoning capacity, they also carry the potential to unintentionally alter safety profiles or ethical frameworks.[20] Therefore, each new model iteration should undergo renewed ethical evaluation.[20] In this context, comparing the most recent ChatGPT model with its immediate predecessor reflects the rapid and iterative development of LLMs, as successive versions are frequently adopted in clinical and educational settings shortly after release. Under such conditions, relative performance changes between consecutive model versions are clinically and ethically meaningful, as model updates may lead to substantial improvements or unintended regressions in safety, reasoning processes, or ethical framing. Accordingly, direct comparison with the immediately preceding model provides a pragmatic and methodologically appropriate benchmark, enabling clinicians and researchers to assess whether transitioning to a newer version is likely to confer tangible benefits in high-stakes domains such as palliative care symptom management, where safety and ethical sensitivity are paramount.

Against this background, the aim of the present study is to provide early evidence on the usability of generative AI as a complementary decision support tool in palliative care and to offer valuable insights for clinicians and researchers interested in the responsible adoption of AI in healthcare. Specifically, we evaluate the recommendations of ChatGPT-4o and ChatGPT-5 in terms of clinical appropriateness, safety, ethical sensitivity, and understandability. Accordingly, beyond a simple generational performance comparison, our aim was to examine whether successive ChatGPT models can provide symptom focused support that remains clinically appropriate, safe, and ethically robust in palliative care.

Methods

Study design and model input

This study does not include any real patient data. The study was conducted using a scenario-based simulation design with quantitative expert ratings. Responses that received lower scores were additionally examined descriptively to contextualize the ratings more effectively. To determine the target symptom set, current palliative care guidelines and relevant reviews were examined, and hospitalization patterns observed in tertiary-level palliative care units in Turkey were also considered. In this way, the selection of scenarios was ensured to realistically reflect national clinical priorities. As a result of this process, 10 core symptoms were identified: pain, anxiety, pressure ulcer, nausea, delirium, dyspnea, constipation, diarrhea, dry mouth, and sleep disturbance.

All ten case scenarios were developed by the investigator through a structured and evidence-informed process. Based on the literature review and national clinical patterns, the investigator drafted short and open-ended scenarios for each symptom and refined them iteratively to ensure clinical realism, internal consistency, and clarity. Each scenario included standardized elements such as patient age, diagnosis, symptom duration, relevant clinical or laboratory findings, and current treatments. Considering that palliative care is not yet recognized as a formal specialty in Turkey and access to multidisciplinary teams is often limited, the scenarios were prepared in Turkish to ensure a realistic assessment of the applicability of LLMs for Turkish-speaking healthcare professionals.

Two LLMs were evaluated: ChatGPT-4o and the most recent version, ChatGPT-5, released in August 2025. Identical scenarios were presented to both models, with the timing determined by model availability. ChatGPT-4o was evaluated in late July 2025, and ChatGPT-5 approximately 1 week later, immediately following its release. In both evaluations, scenarios were presented with the prompts: “How would you approach this situation?” and “How should this symptom be managed?” An example scenario with sequential responses from both models is presented in Table 1. All scenarios and model outputs are provided in the Supplementary Material. Reporting of this simulation-based evaluation was informed by relevant items from the STROBE-AI, adapted to the non-interventional nature of the study.

Table 1. Sample symptom scenario with ChatGPT-4o and ChatGPT-5 responses.

Symptom: Constipation	
Scenario: A 47-year-old female with metastatic breast cancer is under follow-up in the palliative care unit for pain management. She has experienced constipation and abdominal bloating for 5 days. She is receiving opioid analgesics. Laboratory findings are unremarkable except for anemia. Her dietary fiber intake is insufficient, and daily fluid intake is approximately 800 mL.	
Prompt: "*How would you approach this case? How should this symptom be managed?*"	
Model	**Response**
ChatGPT-4o	Opioid-induced constipation should be considered. Laxative therapy should be initiated or optimized. Fluid and fiber intake should be increased, and rectal intervention should be considered if necessary. A multimodal approach (laxatives, fluid supplementation, dietary regulation) should be employed to achieve defecation, aiming for symptom control without altering the opioid dosage.
ChatGPT-5	The presentation suggests opioid-induced constipation (OIC), likely related to the absence of prophylactic laxatives at opioid initiation. Management should include increasing fluid intake to 1.5–2 L/day (if no contraindications) and adding fiber-rich foods. Pharmacological treatment should start with a stimulant laxative (senna or bisacodyl) plus a stool softener (docusate sodium), with osmotic agents (lactulose, polyethylene glycol) as needed. In refractory cases, methylnaltrexone may be considered. Opioid dose review and possible rotation are advised. Non-pharmacological measures include regular toileting and increased mobilization. Monitoring should focus on laxative side effects, electrolyte balance, and resolution of abdominal distension.

Evaluation of responses

Responses were independently assessed by two physicians with at least 5 years of clinical experience, both actively practicing in a tertiary-level palliative care unit. The same evaluators participated in both assessment sets to ensure consistency. One evaluator, who was also the study investigator and head of the unit, was not blinded to the origin of the responses due to their involvement in scenario design. The second evaluator was blinded to both the AI origin and the specific model generating the responses. To reduce potential bias related to partial non-blinding, evaluations were conducted using predefined scoring criteria and independent assessments.

Each response was evaluated across four domains: Clinical appropriateness: the extent to which recommendations were evidence-based and relevant. Safety: the avoidance of potentially harmful or unsafe recommendations. Ethical sensitivity: alignment with core ethical principles, including respect for patient autonomy and informed consent, non-maleficence, and appropriate benefit risk balance, respect for dignity and privacy, and attention to patient and family values through shared decision-making. Understandability: clarity, simplicity, and ease of interpretation. Domains were scored on a five-point Likert scale (1 = completely incorrect, 5 = completely correct). For statistical analyses, both individual and averaged rater scores were used.

Statistical analysis

All analyses were performed using IBM SPSS Statistics version 27. Inter-rater agreement was examined using weighted Cohen's kappa and Spearman's rank correlation. Normality was tested using the Shapiro–Wilk test. As data were not normally distributed, non-parametric methods were applied. Domain score differences within each symptom were tested with the Friedman test. Comparisons between psychiatric symptoms (anxiety, delirium, sleep disorder) and non-psychiatric symptoms were analyzed using the Mann–Whitney U test. Comparisons between the two models (ChatGPT-4o and ChatGPT-5) were conducted using the Wilcoxon signed-rank test. A p-value <0.05 was considered statistically significant.

Ethical considerations

As no real patient data were used, ethics committee approval was not required.

Results

A total of 10 palliative care symptom scenarios were evaluated with ChatGPT-4o and, 1 week later, with ChatGPT-5. Each scenario was presented in the same format and concluded with identical open-ended prompts. Responses were assessed across four domains: clinical appropriateness, safety, ethical sensitivity, and understandability. Importantly, none of the responses contained misleading or harmful information.

Inter-rater agreement

Inter-rater reliability was consistently high across all domains for both models. For ChatGPT-4o, weighted

Cohen's kappa values were 0.871 for clinical appropriateness, 0.820 for safety, 0.833 for ethical sensitivity, and 0.806 for understandability, with corresponding Spearman's rho values of 0.876, 0.825, 0.832, and 0.813 (all $p < 0.001$). ChatGPT-5 showed similarly high agreement, with kappa values of 0.886, 0.844, 0.859, and 0.841, and Spearman's rho values of 0.888, 0.849, 0.861, and 0.845 (all $p < 0.001$). These results confirm the reliability and consistency of the expert evaluations. The consistently high inter-rater agreement observed across all evaluated domains supports the robustness of the evaluation process and indicates that any potential bias introduced by partial unblinding was minimal and did not meaningfully influence the results.

Symptom-specific evaluation

As shown in Table 2, ChatGPT-5 consistently achieved higher scores than ChatGPT-4o across nearly all domains. Clinical appropriateness: The largest improvements were observed in anxiety, pressure ulcer, nausea, dry mouth, and sleep disturbance scenarios. Safety: GPT-5 showed notable gains in delirium, constipation, and diarrhea scenarios. Ethical sensitivity: Both models scored highly, though GPT-5 achieved perfect scores (5.0) in pain, delirium, dyspnea, constipation, diarrhea, dry mouth, and sleep disturbance. Understandability: The greatest improvement was seen in the sleep disturbance scenario.

Descriptive analysis of lower-scoring responses

Lower scores in both models were primarily due to insufficient scenario-specific detail, limited articulation of risk–benefit considerations, suboptimal structuring of recommendations, and incomplete integration of ethical sensitivity. In clinical appropriateness, omissions included disease-specific strategies (e.g. bisphosphonates or denosumab for metastatic bone pain, targeted antiemetic regimens, standardized protocols for opioid-induced constipation) or incomplete diagnostic frameworks (e.g. staging in pressure ulcers, systematic etiologic workup in delirium). In safety, common gaps included lack of monitoring parameters, limited specification of contraindications or adverse effects (e.g. benzodiazepine risks in COPD/dementia, QT prolongation, drug–drug interactions), and insufficient escalation/de-escalation pathways. In ethical sensitivity, deductions stemmed from inadequate attention to informed consent, incomplete discussion of risk–benefit balance, and limited incorporation of patient preferences. For example, in the COPD-related anxiety scenario, benzodiazepine risks were not clearly communicated, consent processes were absent, and long-term psychiatric planning did not sufficiently integrate patient choice. In the pressure ulcer scenario, failure to establish accountability mechanisms for repositioning and limited emphasis on patient care rights reduced ethical scores. In understandability, reductions were linked to the absence of stepwise algorithms, unclear sequencing of management, and insufficiently detailed action plans that could hinder clinical implementation. Overall, while both models provided contextually accurate and relevant recommendations, performance declined when precision, structured guidance, explicit safety frameworks, and comprehensive ethical integration were lacking.

Within-symptom domain comparisons

Friedman test results showed that, in ChatGPT-4o, significant domain differences were observed only for nausea ($\chi^2 = 7.82$, $p = 0.049$), with the highest score in clinical appropriateness and the lowest in understandability. In ChatGPT-5, significant differences were observed in anxiety ($\chi^2 = 11.20$, $p = 0.011$) and nausea ($\chi^2 = 8.23$, $p = 0.042$). For anxiety, ethical sensitivity scored highest and safety lowest, whereas for nausea, clinical appropriateness was highest and understandability lowest.

Psychiatric vs. non-psychiatric symptom comparisons

Between-group comparisons revealed that, in ChatGPT-4o, ethical sensitivity was significantly higher for psychiatric symptoms ($U = 3.0$, $p = 0.043$), while clinical appropriateness ($p = 0.999$), safety ($p = 0.161$), and understandability ($p = 0.248$) showed no differences. In ChatGPT-5, understandability was significantly higher for non-psychiatric symptoms ($U = 2.0$, $p = 0.033$). Differences in clinical appropriateness ($p = 0.999$) were non-significant, while safety ($p = 0.080$) and ethical sensitivity ($p = 0.073$) approached significance (Figure 1).

Wilcoxon signed-rank tests demonstrated that ChatGPT-5 performed significantly better than ChatGPT-4o in clinical appropriateness ($p = 0.010$), safety ($p = 0.002$), and understandability ($p = 0.011$). The improvement in ethical sensitivity was not statistically significant ($p = 0.102$) (Table 3). Visual analyses supported these results, with radar charts showing a consistent upward shift in GPT-5 scores across all domains, most notably in safety and understandability (Figure 2).

Discussion

In this study, 10 symptom scenarios related to palliative care were presented to ChatGPT models. According to expert evaluations, the ChatGPT-5 model, launched in August 2025, was found to perform better than the ChatGPT-4o model in palliative care symptom management in terms of clinical appropriateness, safety, and understandability. However, in both models, certain symptom responses exhibited limitations regarding clinical appropriateness, safety, ethical sensitivity, and understandability.

Most of the existing studies examining the role of ChatGPT in palliative care have been conducted within

Table 2. Comparison of mean domain scores for ChatGPT-4 and GPT-5 across palliative care symptom scenarios.

Symptom	Clinical appropriateness (GPT-4)	Clinical appropriateness (GPT-5)	Safety (GPT-4)	Safety (GPT-5)	Ethical sensitivity (GPT-4)	Ethical sensitivity (GPT-5)	Understandability (GPT-4)	Understandability (GPT-5)
Pain	4.0	4.0	3.0	4.0	5.0	5.0	4.0	4.0
Anxiety	3.0	4.0	2.5	3.5	4.0	4.0	3.0	4.0
Pressure ulcer	3.0	4.0	3.5	4.0	4.0	4.0	3.0	4.0
Nausea	3.0	4.0	3.0	4.0	4.0	4.5	3.0	4.0
Delirium	3.0	3.0	2.5	4.0	4.0	5.0	3.0	3.5
Dyspnea	3.5	4.0	3.0	4.0	5.0	5.0	3.5	4.0
Constipation	3.0	3.5	3.0	4.5	5.0	5.0	3.0	4.0
Diarrhea	3.5	4.0	3.5	5.0	5.0	5.0	4.0	4.0
Dry mounth	3.0	4.0	3.0	4.0	5.0	5.0	3.0	4.5
Sleep disorder	2.5	4.5	3.0	4.0	4.5	5.0	2.5	4.5

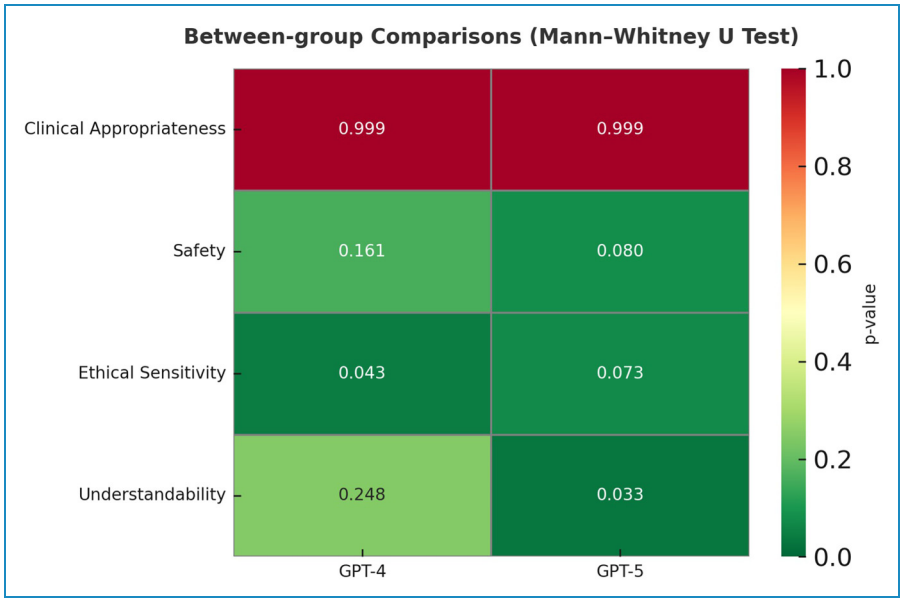

Figure 1. Heatmap showing *p*-values from Mann–Whitney U tests comparing psychiatric and non-psychiatric symptom groups across four evaluation domains for GPT-4 and GPT-5.

Table 3. Direct comparison of ChatGPT-4 and ChatGPT-5 across evaluation domains using Wilcoxon signed-rank tests.

Domain	ChatGPT-4 (Mean, SD)	ChatGPT-5 (Mean, SD)	Inter-rater Kappa (GPT-4o and 5)	Effect Size (*r*)	*P*-value*
Clinical appropriateness	3.15, 0.97	3.90, 0.71	0.871–0.886	0.81	0.010
Safety	3.00, 0.94	4.10, 0.72	0.820–0.844	0.98	0.002
Ethical sensitivity	4.55, 0.65	4.75, 0.55	0.833–0.859	0.52	0.102
Understandability	3.20, 1.01	4.05, 0.78	0.806–0.841	0.81	0.011

*Wilcoxon signed-rank tests.

the last 2 years. To our knowledge, no prior research comprehensively matches this study in terms of scope, design, and methodological rigor. The study findings will make a distinctive and timely contribution to the literature. Importantly, as ChatGPT-5 has only recently been released, this investigation stands among the first to assess its performance in the context of palliative care. Furthermore, the direct, scenario-based comparison between ChatGPT-5 and its predecessor, ChatGPT-4o, offers a rare and valuable perspective on the evolution of LLM capabilities, enabling an evidence-based appraisal of generational advancements relevant to clinical decision support. However, a thematically similar study was conducted by Braun et al. in 2023. In the study published by Braun et al., the quality and guideline concordance of ChatGPT's treatment suggestions for palliative symptoms specific to gynecologic oncology were evaluated.[21] In this study, ChatGPT's responses to 10 different gynecologic symptom scenarios were analyzed by medical experts.[21] The findings of the study indicate that language models such as ChatGPT can provide general treatment recommendations for patient symptoms that are largely consistent with clinical guidelines.[21] However, it was noted that ChatGPT occasionally omitted important treatment options and was unable to provide a personalized treatment plan.[21] Like the findings of Braun et al., this study also revealed that ChatGPT's treatment responses were insufficient for certain symptoms. Braun et al. emphasized in their study that treatment recommendations should always be supported by expert physician input to be complete and individualized.[21]

Although not conducted directly in a palliative patient group, the study by Bianco et al. focused on chronic pain patients receiving long-term opioid therapy, a population indirectly related to palliative care practice. Frequently asked questions by these patients were submitted to

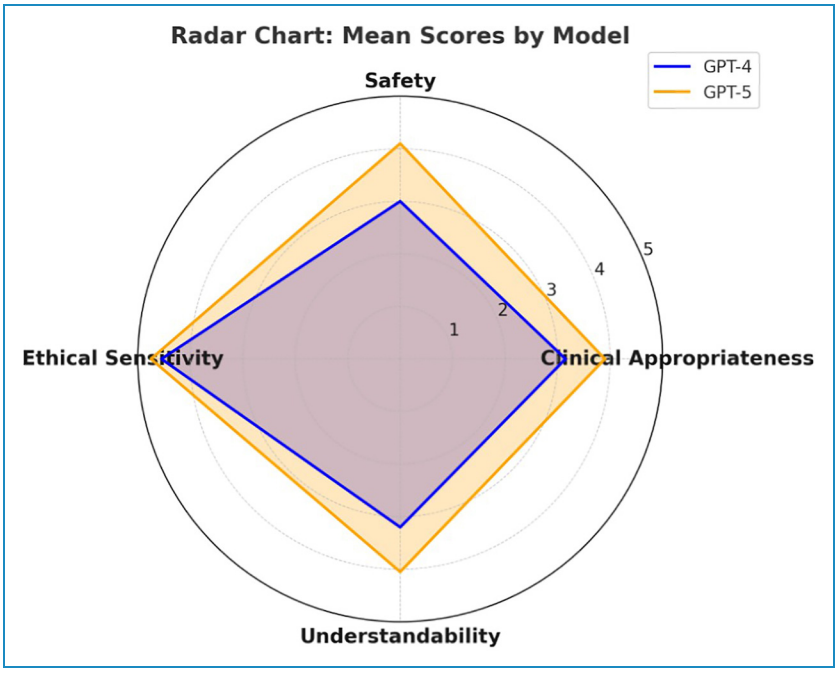

Figure 2. Radar chart illustrating the mean scores of ChatGPT-4o and ChatGPT-5 across the domains of clinical appropriateness, safety, ethical sensitivity, and understandability.

ChatGPT.[22] The reliability, accuracy, and comprehensibility of ChatGPT's responses were evaluated.[22] Overall, ChatGPT's responses showed high reliability and good comprehensibility. Similarly, in this study, ChatGPT demonstrated good understandability in pain symptom management. Bianco et al. also showed that AI systems have significant potential as complementary tools in patient education.[22] However, it was emphasized that performance may be limited in questions requiring high technical knowledge or influenced by clinical context. Therefore, integrating AI applications into clinical practice should involve collaboration with healthcare professionals to ensure accuracy, personalization, and up-to-date responses.[22]

In the study conducted by Jimenez et al. in 2024, the researchers investigated the response quality of the AI tool ChatGPT in various domains of geriatric medicine and the degree of agreement between its responses and physicians' evaluations.[23] Designed similarly to this study, the research involved presenting ChatGPT with 10 clinical scenarios related to geriatrics and requesting physicians to assess the AI generated responses using a 5-point Likert scale.[23] Among these scenarios, two were directly related to delirium. ChatGPT received average scores of 2.9/5 and 3.0/5 for these scenarios, respectively, making them among the lowest-rated cases in the study.[23] These findings are consistent with this study's results, in which ChatGPT's responses to delirium symptom-related questions also received lower scores compared to other symptom categories. These low ratings highlight the limitations of AI in managing multifaceted conditions like delirium, which often require rapid clinical assessment and medical intuition. Despite its stronger performance in theoretical or general knowledge questions, these findings suggest that ChatGPT is not yet sufficiently reliable to support clinical decision-making in complex scenarios such as delirium, particularly in geriatric and palliative care settings.

Although this study focuses on symptom management, which remains a relatively underexplored area in the context of palliative care and ChatGPT, several studies in the existing literature have examined the use of ChatGPT in other aspects of palliative care. In an observational and cross-sectional study conducted by Hanci et al. in 2024, the quality, reliability, and readability of information provided by ChatGPT and other chatbots on palliative care were comparatively evaluated.[14] Hanci et al. reported that, overall, the readability levels and content quality of the texts were found to be inadequate for patient education purposes.[14] Therefore, it is emphasized that AI systems used in sensitive fields such as palliative care should be improved to provide more understandable and higher-quality information.[14] In contrast to the approach of Hanci et al., our study focused on ChatGPT's responses to symptoms commonly encountered in the palliative care setting rather than general patient-generated questions.

In 2024, Kim et al. evaluated the ability of chatbots to define and distinguish the terms palliative care, supportive care, and hospice care.[16] Kim et al. showed that although AI platforms generally have high accuracy, they still have significant limitations in terms of reliability and comprehensiveness.[16] In 2025, Admane et al. asked AI models to

define the terms "terminally ill," "end of life," "transitions of care," and "actively dying."[24] The study showed that although chatbots could define some terms with high accuracy, they had significant shortcomings in reliability and readability. Therefore, these technologies should be evaluated under clinician supervision before being used as direct information sources in palliative care.[24]

In their 2023 study, Srivastava et al. examined the role of AI technologies in communication, particularly at the end of life.[25] They evaluated whether ChatGPT could contribute to communication processes in palliative care. They suggested that ChatGPT could be considered a complementary tool in therapeutic communication.[25] However, these technologies cannot yet fully replace human interaction; human-specific qualities such as empathy, intuition, and contextual awareness remain beyond the scope of AI.[25] In their 2024 study, Burry et al. stated that using AI as a direct communication tool in highly sensitive fields like palliative care is not appropriate. However, they suggested that it could serve as a valuable support tool for training physicians in language and approach for serious illness communication.[26]

A key methodological limitation of the present study is its reliance on standardized, scenario-based inputs rather than real-time clinical encounters. Although this limits the external validity and may not fully capture the complexity and variability of real-world patient care, the approach was deliberately chosen to address several important considerations. First, scenario-based testing allows for ethically safe and controlled evaluation of AI-generated recommendations without exposing identifiable patient data. This is particularly relevant in palliative care, where patient vulnerability and privacy concerns are paramount. Second, by standardizing clinical inputs, this design reduces uncontrolled variability, enabling fair and consistent comparison across multiple symptom types and evaluation domains. In addition to these methodological considerations, the study has other limitations. First, the scenarios were prepared only in Turkish. Future studies incorporating multilingual scenarios and cross-cultural evaluator panels are needed to assess generalizability. Considering that the performance of LLMs may vary by language, our findings mainly reflect the Turkish user experience, which may limit the generalizability of the results to other languages. Second, although the physician evaluations involve a high level of expertise, they may inherently include subjective judgments. In particular, the ethical sensitivity domain inherently involves interpretive judgment, as ethical reasoning in palliative care is context-dependent and value-laden. Another limitation is that the responses were evaluated by only two physicians from the same palliative care unit. Although inter-rater agreement was excellent, the lack of a more diverse rating panel may limit the generalizability of the findings.

Despite these limitations, the study also has several notable strengths. To our knowledge, it is one of the first to evaluate ChatGPT-5 in the context of palliative care and to directly compare its performance with ChatGPT-4, offering valuable insights into generational improvements in LLMs. The use of expert raters with specialized experience in palliative care ensured high-quality and clinically relevant assessments. Additionally, the structured and symptom-specific scenario design allowed for a focused evaluation across multiple critical domains clinical appropriateness, safety, ethical sensitivity, and understandability providing a comprehensive view of the model's potential and limitations in clinical decision support.

The release of a new version of an AI model does not automatically mean that its clinical decision support will be better. This assumption is especially unreliable in fields such as palliative care, where safety and ethical principles are critically important. Therefore, each new model must be systematically re-evaluated within the relevant domain to determine whether it offers genuine improvement or whether there are regressions in safety or ethical performance.[20,27] The findings of this study carry important implications for both clinical practice and future research in palliative care. The measurable improvements observed with ChatGPT-5 highlight its potential as a supplementary decision support tool, particularly in resource-limited settings where timely access to multidisciplinary expertise is often restricted. However, AI-generated recommendations should be regarded as complementary rather than substitutive, ensuring that clinical judgment, ethical oversight, and individualized care remain central to decision-making. This perspective aligns with the broader discourse in the literature, which emphasizes that AI is designed to augment rather than replace healthcare providers.[28]

Future research should include multilingual validation studies, prospective real-world evaluations across diverse patient populations, and investigations on the safe integration of AI tools into clinical workflows and electronic health record systems. In addition, comparative studies assessing model performance against early-career clinicians may help clarify the practical utility and potential educational value of generative AI in palliative care. Furthermore, research directly comparing generative AI outputs with the information sources traditionally used by patients and family caregivers may offer valuable insights into the real-world relevance, accessibility, and limitations of these tools. Such efforts will be essential to determine feasibility, safety, long-term clinical impact, and to guide the responsible adoption of generative AI in clinical practice.

Conclusions

This study is, to our knowledge, among the first to evaluate ChatGPT-5 in palliative care and to directly compare it with ChatGPT-4o across several clinically important domains. Based on 10 standardized symptom scenarios assessed by experienced physicians, ChatGPT-5 showed significant

improvements in clinical appropriateness, safety, and understandability, while ethical sensitivity remained high in both models without a meaningful difference.

The strong agreement between raters supports the credibility of these results. Overall, the findings suggest that LLMs particularly ChatGPT-5 may serve as useful complementary tools for clinical decision-making in palliative symptom management. At the same time, the study underlines the need for further refinement to improve accuracy, ensure robust safety mechanisms, and strengthen the integration of patient-centered ethical principles. Future work should aim to validate performance in real-world clinical settings, across different languages, and within existing healthcare workflows. Ultimately, in a sensitive field such as palliative care, generative AI should be used only as a supportive aid, with final responsibility and decisions remaining in the hands of clinicians.

Guarantor

EV.

Acknowledgments

The author would like to express sincere gratitude to the palliative care team at Kirsehir Training and Research Hospital for their valuable contributions throughout this study.

ORCID iD

Emre Vuraloglu https://orcid.org/0000-0001-6634-0671

Ethical considerations

As no real patient data were used, ethics committee approval was not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Author contributions

EV contributed to conceptualization; methodology; software; validation; formal analysis; investigation; resources; data curation; writing—original draft preparation; writing—review and editing.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Guarantor

EV.

Supplemental material

Supplemental material for this article is available online.

References

1. Pourghazian N, Mahmoud L and Hammerich A. Strengthening palliative care in the WHO Eastern Mediterranean region. *East Mediterr Health J* 2022; 28: 553–554.
2. World Health Organization. *Long-term care for older people: package for universal health coverage*. Geneva: WHO, 2023.
3. Radbruch L, De Lima L, Knaul F, et al. Redefining palliative care-A new consensus-based definition. *J Pain Symptom Manage* 2020; 60: 754–764.
4. Ahmed F, Kutluk T, Yurduşen S, et al. Palliative care in Turkey: insights from experts through key informant interviews. *J Cancer Policy* 2024; 42: 100506.
5. Kutluk T, Ahmed F, Cemaloğlu M, et al. Progress in palliative care for cancer in Turkey: a review of the literature. *Ecancermedicalscience* 2021; 15: 1321.
6. Yu E, Chu X, Zhang W, et al. Large language models in medicine: applications, challenges, and future directions. *Int J Med Sci* 2025; 22: 2792–2801.
7. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. *J Med Internet Res* 2025; 27: e59069.
8. Omiye JA, Gui H, Rezaei SJ, et al. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med* 2024; 177: 210–220.
9. Shah NH, Entwistle D and Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA* 2023; 330: 866–869.
10. Choi J. Large language models in medicine. *Healthc Inform Res* 2025; 31: 111–113.
11. Yip HF, Li Z, Zhang L, et al. Large language models in integrative medicine: progress, challenges, and opportunities. *J Evid Based Med* 2025; 18: e70031.
12. Kantor J. ChatGPT, large language models, and artificial intelligence in medicine and health care: a primer for clinicians and researchers. *JAAD Int* 2023; 13: 168–169.
13. Yang R, Tan TF, Lu W, et al. Large language models in health care: development, applications, and challenges. *Health Care Sci* 2023; 2: 255–263.
14. Hanci V, Ergün B, Gül Ş, et al. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine (Baltimore)* 2024; 103: e39305.

15. Gondode PG, Mahor V, Rani D, et al. Debunking palliative care myths: assessing the performance of artificial intelligence chatbots (ChatGPT vs. Google Gemini). *Indian J Palliat Care* 2024; 30: 284–287.
16. Kim MJ, Admane S, Chang YK, et al. Chatbot performance in defining and differentiating palliative care, supportive care, hospice care. *J Pain Symptom Manage* 2024; 67: e381–e391.
17. Barreda M, Cantarero-Prieto D, Coca D, et al. Transforming healthcare with chatbots: uses and applications—A scoping review. *DIGITAL HEALTH* 2025; 11: 20552076251319174.
18. Sanli AN, Turan B and Tekcan Sanli DE. Advances in large language model performance: a comparative study of ChatGPT-4 and ChatGPT-5 on ABSITE questions. *Am Surg* 2025: 31348251390958.
19. Tabanli A and Demirkiran ND. Comparing ChatGPT 3.5 and 4.0 in low back pain patient education: addressing strengths, limitations, and psychosocial challenges. *World Neurosurg* 2025; 196: 123755.
20. Haltaufderheide J and Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). *NPJ Digit Med* 2024; 7: 183.
21. Braun EM, Juhasz-Böss I, Solomayer EF, et al. Will I soon be out of my job? Quality and guideline conformity of ChatGPT therapy suggestions to patient inquiries with gynecologic symptoms in a palliative setting. *Arch Gynecol Obstet* 2024; 309: 1543–1549.
22. Lo Bianco G, Robinson CL, D'Angelo FP, et al. Effectiveness of generative artificial intelligence-driven responses to patient concerns in long-term opioid therapy: Cross-model assessment. *Biomedicines* 2025; 13: 636.
23. Rosselló-Jiménez D, Docampo S, Collado Y, et al. Geriatrics and artificial intelligence in Spain (Ger-IA project): talking to ChatGPT, a nationwide survey. *Eur Geriatr Med* 2024; 15: 1129–1136.
24. Admane S, Kim MJ, Reddy A, et al. Performance of three conversational artificial intelligence agents in defining end-of-life care terms. *J Palliat Med* 2025; 28: 1102–1107.
25. Srivastava R and Srivastava S. Can artificial intelligence aid communication? Considering the possibilities of GPT-3 in palliative care. *Indian J Palliat Care* 2023; 29: 418–425.
26. Burry N, Nakagawa S and Blinderman CD. "You are not alone": The allure and limitations of artificial intelligence in serious illness communication. *J Palliat Med* 2024; 27: 7–9.
27. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc* 2024; 31: 2730–2739.
28. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *DIGITAL HEALTH* 2023; 9: 20552076231186520.