Article

Not peer-reviewed version

Evaluating 11 Large Language Models in Answering Key Questions on Ovarian Cancer

Michela Quaranta , Yong Sheng Tan , Areti Karamanou , Evangelos Kalampokis , Nicolas M Orsi , Diederick DeJong , Alexandros Laios *

Posted Date: 11 April 2026

doi: 10.20944/preprints202604.0800.v1

Keywords: large language models; ovarian cancer; patient communication; artificial intelligence; natural language processing; empathy; readability; GPT-4o; Claude

 doi:10.20944/preprints202604.0800.v1

Article

Evaluating 11 Large Language Models in Answering Key Questions on Ovarian Cancer

Michela Quaranta [1,2], Yong Tan [1], Areti Karamanou [3], Evangelos Kalampokis [3], Nicolas Orsi [1], Diederick DeJong [2] and Alexandros Laios [2]

[1] University of Leeds, Leeds Institute of Medical Research, St. James's University Hospital, Beckett Street, Leeds LS9 7TF, United Kingdom
[2] Department of Gynaecological Oncology, ESGO Centre of Excellence in Advanced Ovarian Cancer Surgery, St. James's University Hospital, Beckett Street, Leeds LS9 7TF, United Kingdom
[3] Information Systems Lab, Department of Business Administration, University of Macedonia, 54636 Thessaloniki, Greece
* Correspondence: a.laios@nhs.net; Tel.: +44-113-206-8438

Abstract

Background: The release of Large Language Models (LLMs) has introduced numerous benefits across the healthcare domain. This study evaluated the responses of 11 LLMs from the Claude, Mistral, Llama, and GPT families to Frequently Asked Questions (FAQs) regarding ovarian cancer with regards to three domains: (a) ease of understanding, (b) accuracy, and (c) empathy. **Methods:** Fifteen FAQs were sourced from the Ovarian Cancer Action (OCA) website comprising (a) anticipated questions and (b) actual questions. Responses from each of the 11 LLMs were blinded and then evaluated by three Gynaecological Oncology Surgical Fellows using a 5-point Likert scale. Inter-observer agreement was calculated for each response, and LLMs were compared across the three domains using Friedman's test ($p < 0.05$). Finally, all LLM responses were compared with the ones from the OCA website using the same evaluation criteria. **Results:** Varying levels of inter-observer agreement were observed. Claude 3 Opus produced the easiest-to-understand answers (average score 4.38), followed by Mistral Large (4.36) and GPT-4o (4.33). GPT-4o scored highest for accuracy (average score 4.24) and showed strongest performance in empathy (average of 3.87). Compared with the OCA responses, GPT-4o outperformed all models in accuracy (4.24) and empathy (3.87), with 50% of its responses rated more accurate and 70% more empathetic than the OCA content. Claude 3 Opus, Mistral Large, and Mixtral 8x7B surpassed OCA in clarity for one-third of responses, while Claude 3 Sonnet achieved the highest readability gains (40%). **Conclusion:** The study informs the development of LLMs suitable for patient-facing ovarian cancer communication with Claude 3 Opus and GPT-4o excelling in different metrics. While improvements in emotional intelligence remain necessary, our findings pave the way for developing a specialized LLM for ovarian cancer, using domain-specific text to provide comprehensive and empathetic information.

Keywords: large language models; ovarian cancer; patient communication; artificial intelligence; natural language processing; empathy; readability; GPT-4o; Claude

1. Introduction

Ovarian cancer remains a major global health challenge and is among the most common malignancies affecting women, accounting for approximately 4% of cancer diagnoses and nearly 5% of cancer-related mortality worldwide in 2020, with substantial variation in incidence between regions [1]. In recent decades, medical and scientific advances have led to better diagnostic tools, more refined surgical strategies, and novel treatments such as targeted therapies. However, despite these developments, ovarian cancer remains challenging to diagnose early, mainly due to non-specific presenting symptoms [2]. Consequently, patients often seek information online, turning to reputable sources or general internet searches to learn about symptoms, risks, and treatments. In this evolving digital

 doi:10.20944/preprints202604.0800.v1

landscape, large language models (LLMs) have emerged as critical tools in facilitating cancer healthcare information delivery [3].

In healthcare, these AI-driven innovations range from algorithms that predict patient risk to virtual nurse assistants that provide guidance [4]. More recently, the release of sophisticated LLMs, exemplified by ChatGPT, has gained interest in their use for complex tasks such as summarising research findings, generating patient information, and offering empathetic advice [5,6]. Traditional question-and-answer portals on healthcare websites or via chatbots have often been constrained by limited domain knowledge. Current-generation LLMs employ large datasets to produce coherent and context-sensitive responses [7]. Nevertheless, questions remain regarding their reliability, consistency, and empathetic tone when handling sensitive medical topics such as ovarian cancer.

Ovarian cancer information requires factual accuracy because of the disease's complexity and the emotional burden associated with this devastating diagnosis. Patients and their families often want more than just technical or statistical insights; they seek reassurance and clear guidance. The Ovarian Cancer Action (OCA) website is a key resource for anyone seeking up-to-date information on ovarian cancer [8]. It offers clear guidance on symptoms, diagnosis, and treatment, as well as practical advice for those living with the disease. The website also reflects the organisation's research initiatives, aiming to advance scientific understanding and improve patient outcomes. In addition, users can access personal narratives, awareness campaigns, and opportunities to support the charity through fundraising, volunteering, or advocacy. In a digital environment where internet-based health resources are abundant but highly variable in quality, and where inaccurate or overly generalised information may be harmful, OCA reasonably serves as a benchmark against which other online resources can be evaluated. Against this backdrop, there is a clear need to assess emerging technologies including LLMs.

Current LLMs vary in both their architecture and training data [9]. Some may have been trained primarily on scientific literature, providing precise and accurate answers. Other models may have been exposed to broader web-based datasets, offering more conversational and empathetic responses. In the context of cancer care where accuracy, clarity, and compassion are vital, these distinctions become crucial. A key challenge in assessing LLMs for clinical use is the complexity of patient queries [10]. Questions can range from straightforward inquiries about symptoms to deeper concerns about prognosis, treatment side effects, and emotional wellbeing. Anticipated questions reflect clinician's priorities in risk factors, treatment modalities, and preventive measures while actual questions come from real users. Evaluating both offers a more comprehensive picture of model performance across patient needs. The task is not as simple as reviewing a single metric like accuracy. In patient communication, the perception of clarity and empathy significantly affects comprehension, adherence to medical advice, and overall satisfaction. If language is too technical, it may be confusing, whereas overly informal explanations risk lacking the necessary medical rigour. Empathy is a key factor in oncology, where fear, uncertainty, and emotional vulnerability are common [11]. A purely factual response may be accurate but can fail to reassure or emotionally support a patient.

Given this landscape, we undertook a study to systematically evaluate 11 LLMs in answering 15 frequently asked questions (FAQs) about ovarian cancer. These questions, derived from the OCA website, captured both anticipated and actual queries that patients often pose via search engines or through the organisation's communication channels. The primary goal of this study was to identify which LLMs excel at understanding and addressing ovarian cancer queries in a manner that is both medically accurate and empathic. We also aimed to explore whether certain models consistently outperformed others across all metrics, or if specific models shone in individual categories.

2. Methods

2.1. Data Sources and Question Selection

This cross-sectional study aimed to assess the performance of 11 LLMs in responding to FAQs related to ovarian cancer. Fifteen FAQs were sourced from the Ovarian Cancer Action (OCA) website

 doi:10.20944/preprints202604.0800.v1

(Supplementary Table 1). These FAQs encompassed both (a) anticipated questions, which are what people generally seek to know about ovarian cancer that health professionals consider crucial for patient education, and (b) actual questions, frequently encountered through user interactions, which are commonly asked by users via search engines or the OCA's communication channels including emails and social media. As this study did not involve the collection or analysis of patient- identifiable data, formal ethical approval was not required. All researchers involved are members of the core gynaecological oncology team.

2.2. Large Language Models Evaluated

The 15 questions were given as prompts to 11 LLMs from four wider families (Supplementary Table 2): Claude (Claude 3 Haiku v1.0, Claude 3 Opus v1.0, Claude 3 Sonnet v1.0, and Claude Instant v1.2), Mistral (Mistral Large 24.02 v1.0, Mixtral 8x7b Instruct v1.0, and Mistral 7b Instruct v0.2), Llama (Llama 3.1 70b Instruct v1.0 and Llama 3.1 8b Instruct v1.0), and GPT (GPT 4 Turbo, and GPT4o). These LLMs were selected to represent a diversity of architectural approaches, openess, and model sizes. Openness refers to the degree to which the model components including architecture, pretrained weights, training codes and datasets, are published under licenses that allow independent use and sharing. Claude and GPT models are proprietary since their license do not allow access to their components, while Llama models are open and freely accessible. Pseudonyms or version identifiers were attached to the models to preserve anonymity.

The required LLM inferences were obtained through the SageMaker or Bedrock Amazon Web Services (AWS). For all experiments, the temperature, top-P, and max output token were set to 0, 0.9, and 512, respectively [12]. Temperature controls how random the output is. A zero value makes the model always choose the most likely word, ensuring consistent results. A higher temperature value would reduce the randomness in the selection of the output. Top P sets a threshold, keeping only the most likely words eligible for selection. Setting top-P to 0.9 limits the word selection to 90% of the total probability, enhancing the coherence and relevance of the generated text [12].

2.3. Study Design

A study flowchart is shown in Figure 1.

Each of the 15 FAQs was imputed into each LLM under identical conditions to standardise the testing procedure. We instructed the models to provide answers suitable for a lay audience while focusing on factual correctness. The outputs were collected in a random order, with no identifiers indicating which LLM had generated a particular response. This randomisation sought to reduce potential bias and prevent assessors from associating responses with known model characteristics or reputations.

2.4. Assessment Criteria

We utilised a 5-point Likert scale to evaluate three key domains:

Ease of Understanding: The degree to which the answer was readily comprehensible to a typical layperson. This category focused on clarity, logical flow, and minimal jargon.

Accuracy: The correctness of the content based on established medical knowledge and best practice guidelines. Assessors were asked to pay particular attention to information, consistency with recognised standards, and avoidance of misinformation.

Empathy: The extent to which the response acknowledged the emotional impact of an ovarian cancer diagnosis and provided reassurance. This domain measured the "human touch" and sensitivity displayed in discussing prognosis, treatment, or supportive care.

 doi:10.20944/preprints202604.0800.v1

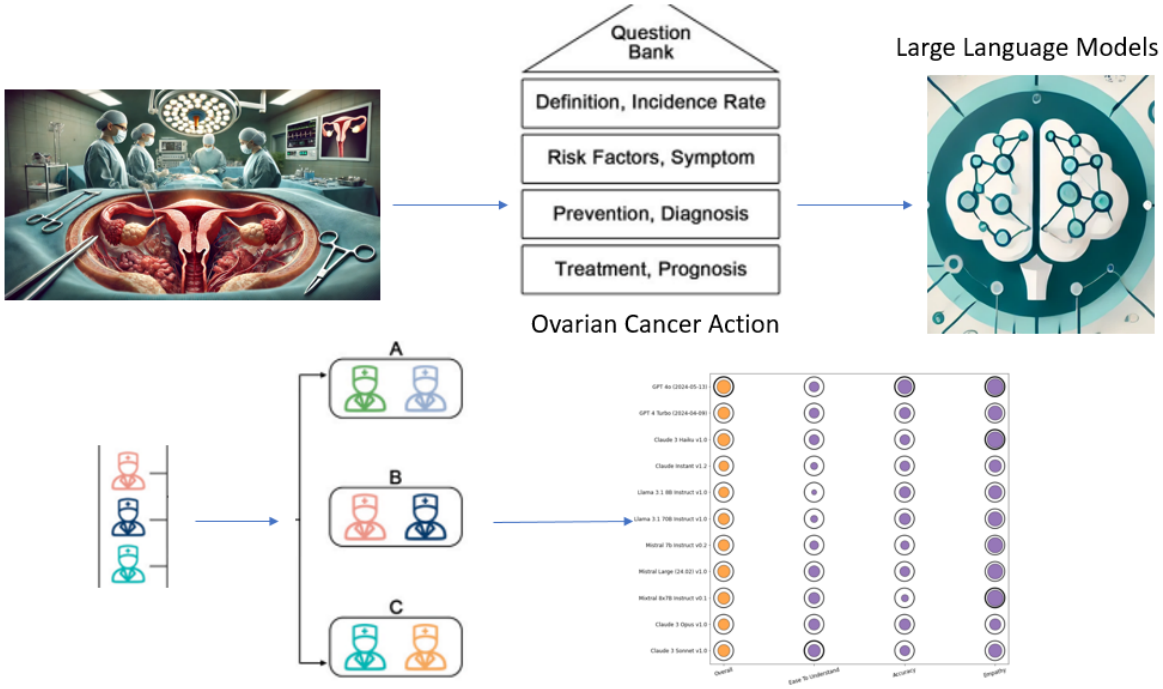

Figure 1. Study workflow

The responses were blinded and independently evaluated by three Gynaecological Oncology Surgical Fellows (MQ, YC, AL) using the above criteria. In addition, answers to the same questions extracted from the OCA website were also evaluated using the same criteria by the three evaluators. Before scoring, they underwent a training session to align their interpretations of each Likert point. Their professional experience in gynaecological oncology albeit variable, was crucial for judging both the accuracy of content and the appropriateness of empathetic language.

2.5. Inter-Observer Agreement

We assessed pairwise percentage agreement among the Gynaecological Oncology Surgical Fellows across the three domains on the 15 responses from each LLM. initial analysis revealed varying degrees of inter-observer agreement, with more substantial divergence occurring around borderline cases; for instance, whether a response should be rated 3 or 4 on the 5-point scale. To reduce ambiguity and improve reliability, the initial 5-point Likert scale was collapsed into 3-points. The initial range of 1 to 5 was condensed by merging: scores 1 and 2 into "1 (negative)"; score 3 into "2 (neutral)"; scores 4 and 5 into "3 (positive)". The mean pairwise percent agreement for each LLM and metric was calculated by comparing the ratings assigned by each pair of evaluators across the 15 responses and calculating the proportion of identical scores.

2.6. Performance Evaluation

To assess the overall ease of understanding, accuracy, and empathy, the mean scores from the three evaluators were calculated across each metric. The mean scores for responses extracted from the OCA website were calculated using the same criteria, allowing for direct comparison. The results were analysed alongside the pairwise percentage agreement. In addition, the proportion of answers rated with higher mean scores related to the mean evaluation of the OCA answers was also calculated across each metric.

The Friedman test was applied to compare the median ratings across the 11 LLMs with regards to the three evaluation criteria. A p-value of less than 0.05 was considered statistically significant. This was used to account for repeated responses from each LLM.

 doi:10.20944/preprints202604.0800.v1

3. Results

3.1. General Observations and Inter-Observer Agreement

Inter-observer agreement was highly variable across the three domains. Figure 2 summarises the 3-point Likert scale agreement levels.

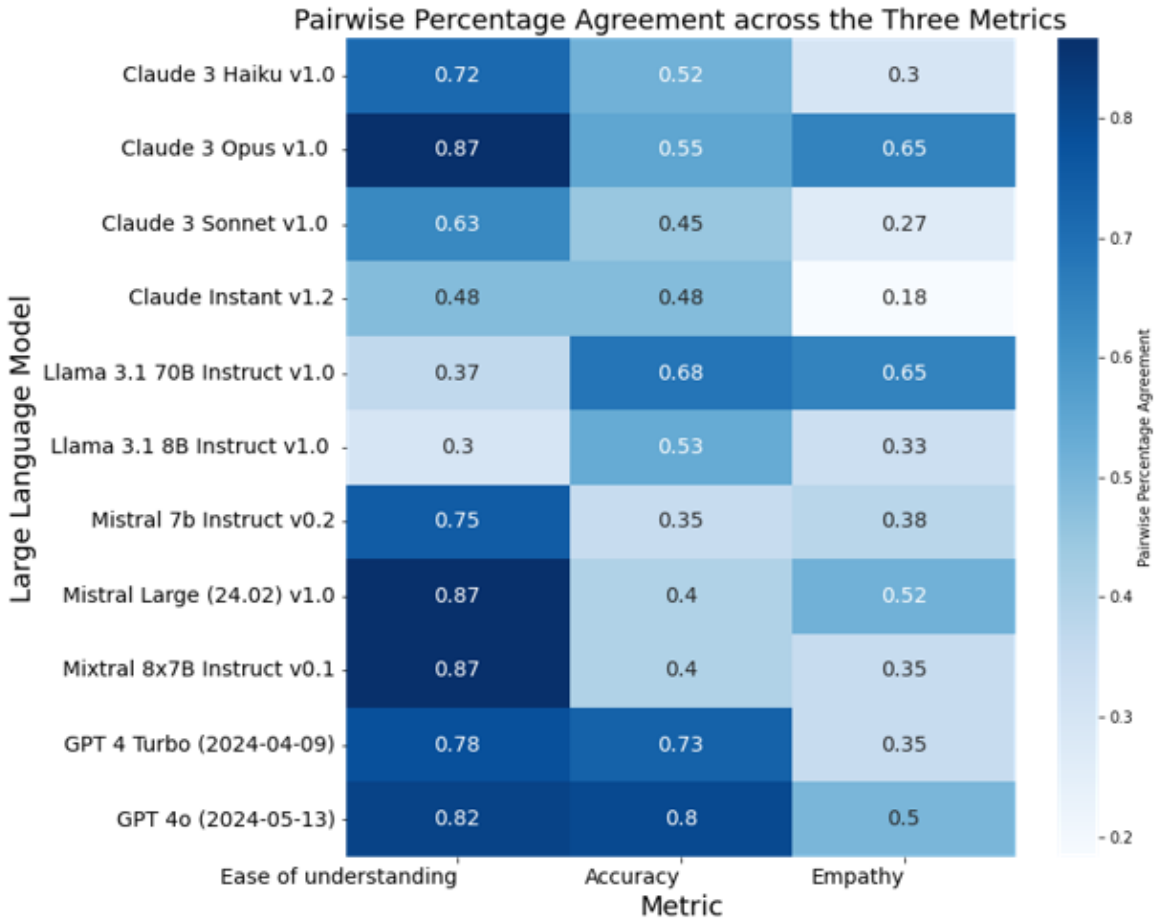

Figure 2. Heatmap showing the percent agreement of the three assessors for 11 LLMs across three domains (columns); ease of understanding, accuracy and empathy. The pairwise agreement between assessors for each response was computed. Colour intensity corresponds to the agreement percentage. Varying levels of inter-observer agreement were revealed.

With regards to ease of understanding, Claude 3 Opus, Mistral Large and Mistral 8x7B Instruct were scored with an average 87%, indicating a strong consensus among evaluators, followed by the LLMs of the GPT family (GPT4o and GPT 4 Turbo with 82% and 78% mean percentage agreement respectively). By contrast, Llama 8B (30%), Llama 3.1 70B (30%), and Claude Instant (48%) scored a lower percentage agreement in ease of understanding. Regarding accuracy, GPT 4o (80%) and GPT 4 Turbo (73%) showed the highest evaluator agreement. However, the percentage agreement for most models (8 out of 11) was in the range between 40% and 55%. Finally, empathy was the challenging criterion, with percentage agreement often below 50%. Specifically, apart from Claude 3 Opus and Llama 3.1 70B (both 65%), the percentage agreement among the evaluators was 52% or lower.

3.2. Ease of Understanding

Figure 3a shows the mean ease of understanding scores related to the percentage agreement among evaluators across all LLMs.

 doi:10.20944/preprints202604.0800.v1

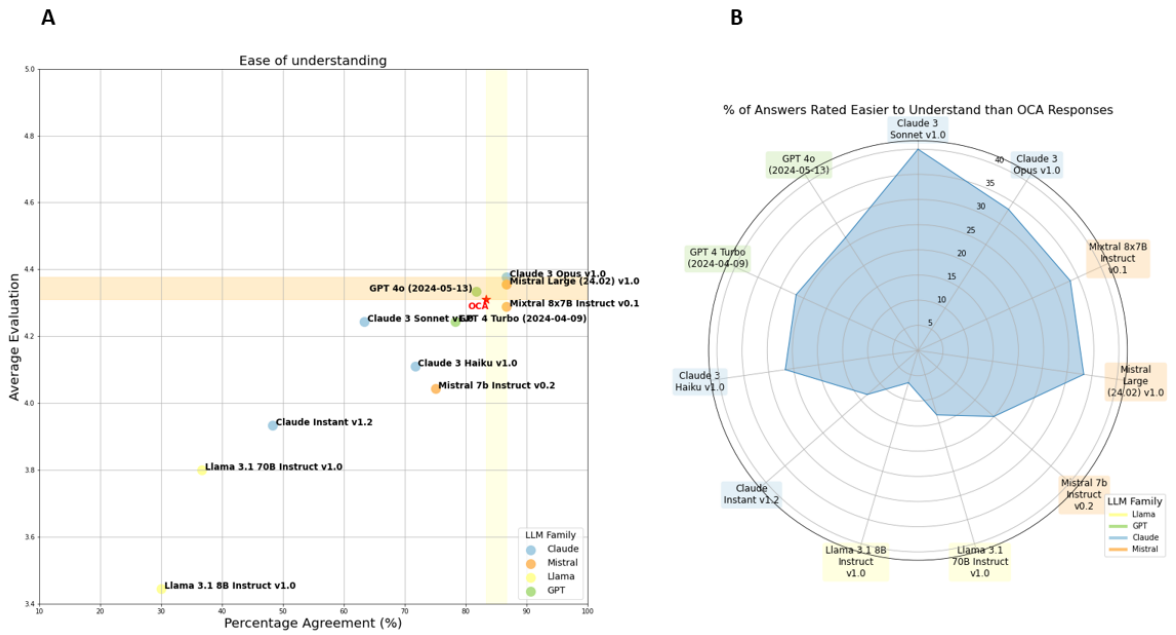

Figure 3. A) Scatter plot percentage agreement with average evaluation (easy to understand) B) Spider chart (radar plot) with % of Answers Evaluated as easier to understand related to OCA site. The percentage has been calculated based on the pairwise agreement. The yellow shaded region indicates a threshold for desirable performance. Only the top-performing models exceed both the evaluation score and agreement percentage highlighted by this region. The blue shaded region in the radar plot represents the performance of the LLMs in relation to the percentage of answers evaluated as easier to understand.

Overall, the evaluators tended to agree more on the models that provided easier to understand responses. The evaluators strongly agreed (87%) that Claude 3 Opus, Mistral Large, and GPT 4o provided the easiest to understand responses related to ovarian cancer with average scores of 4.38 (95% CI: 4.11, 4.65), 4.36 (95% CI: 3.98, 4.73), and 4.33 (95% CI: 4.05, 4.62), respectively. While GPT-4o generally provided coherent answers, it offered more detail than strictly necessary, occasionally drifting into technical descriptions. GPT-4 Turbo (4.24, 95% CI: 3.93, 4.56) followed close behind GPT 4o in ease of understanding. Mistral Large (4.35, 95% CI: 3.98, 4.73) tended to opt for accessible language yet retained accuracy. The layout of its responses appeared straightforward, often employing short paragraphs and bulleted lists when addressing complex points. Most models produced overly long or convoluted texts, referencing information without sufficiently explaining its relevance. The responses of the Llama models were the most difficult to understand. Specifically, Llama 3.1 8B (3.44, 95% CI: 2.86–4.03) had the lowest mean score for ease of understanding, followed by Llama 3.1 70B (3.8, 95% CI: 3.31–4.29), and Claude Instant (3.93, 95% CI: 3.56–4.30). However, there was a notable disagreement among the evaluators on the difficulty to understand these responses ($p < 0.05$).

3.3. Accuracy

OpenAI's models performed best in factual accuracy and evaluator alignment (Figure 4a).

 doi:10.20944/preprints202604.0800.v1

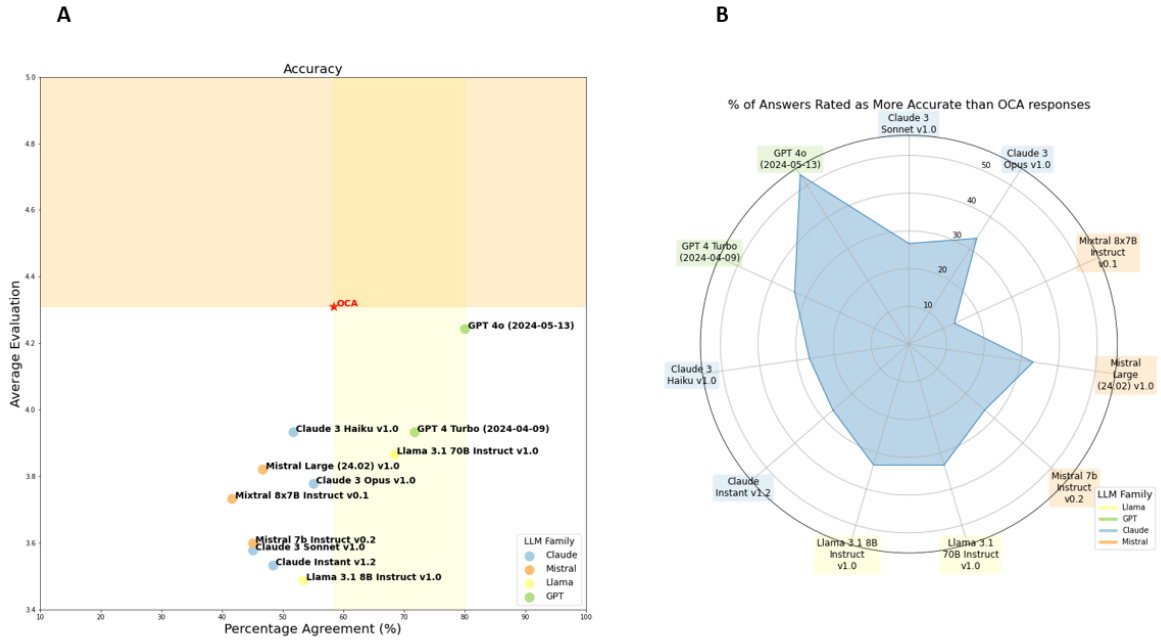

Figure 4. A) Scatter Plot percentage agreement with average evaluation (accuracy) B) Spider chart (radar plot) with % of Answers Evaluated as more accurate related to OCA site. The percentage has been calculated based on the pairwise agreement. The yellow shaded region indicates a threshold for desirable performance. Only the top-performing models exceed both the evaluation score and agreement percentage highlighted by this region. The blue shaded region in the radar plot represents the performance of the LLMs in relation to the percentage of answers evaluated as easier to understand.

Specifically, evaluators agreed that GPT 4o led with an average rating of 4.24 (95%, CI: 3.97–4.52) on the 5-point scale followed by GPT 4 Turbo (3.93, 95% CI: 3.62–4.24), albeit with slightly lower agreement. Claude 3 Haiku's accuracy (3.93, 95% CI: 3.56–4.30) was identical to that of GPT 4 Turbo, but with a lower degree of inter-evaluator agreement (52% compared to 73%). By contrast, Claude 3 Opus (3.78, 95% CI: 3.47–4.09) and Claude 3 Sonnet (3.58, 95% CI: 3.15–4.00) received slightly lower ratings for both accuracy and evaluator agreement. In addition, evaluators agreed on the lower accuracy of Llama 3.1 70B (3.87, 95% CI: 3.56–4.13). Mistral Large (3.82, 95% CI: 3.42–4.22) and Mixtral 8x7B (3.73, 95% CI: 3.35–4.11) featured near the centre of the scatterplot with moderate accuracy and evaluator agreement compared to the other models. Finally, the lowest accuracy scores were achieved by Llama 3.1 8B (3.49, 95% CI: 3.03–3.95), Claude Instant (3.53, 95% CI: 3.20–3.86), and Mistral 7B (3.6, 95% CI: 3.23–3.97), but the agreement of the evaluators in these ratings was low ($> 54\%$). Although statistically significant differences were observed between models ($p < 0.05$), the absolute differences in performance were small, indicating stable accuracy across models.

3.4. Empathy

Empathy emerged as the most challenging domain for all models (Figure 5a).

GPT 4o featured the highest mean score (3.87, 95% CI: 3.45–4.29) but with moderate evaluator agreement (50%). It was followed by Mistral Large (3.65, 95% CI: 3.45–4.29) and GPT-4 Turbo (3.58, 95% CI: 3.09–4.06), albeit the latter with much lower agreement (35%). By contrast, models from the Claude family, Claude 3 Opus (3.24, 95% CI: 3.05–3.44) and Claude Instant (3.29, 95% CI: 2.67–3.91), underperformed, both in low empathy score and evaluator agreement. Llama models showed mixed performance, with Llama 3.1 70b achieving a very good average empathy score of 3.58 (95% CI: 3.43–3.73) on the 5-point scale ($p < 0.05$). Responses frequently recognised patient anxiety prior to discussing treatment options and encouraged engagement with support networks. Those features were less consistently observed in other models, which often prioritised clinical information without empathetic framing. The Llama 8B ranked among the lowest performing models (3.23, 95% CI:

 doi:10.20944/preprints202604.0800.v1

2.85–3.73). Empathy was variably expressed across the remaining models, often limited to generic statements. We observed used disclaimers or messages urging patients to seek professional help without offering genuine warmth.

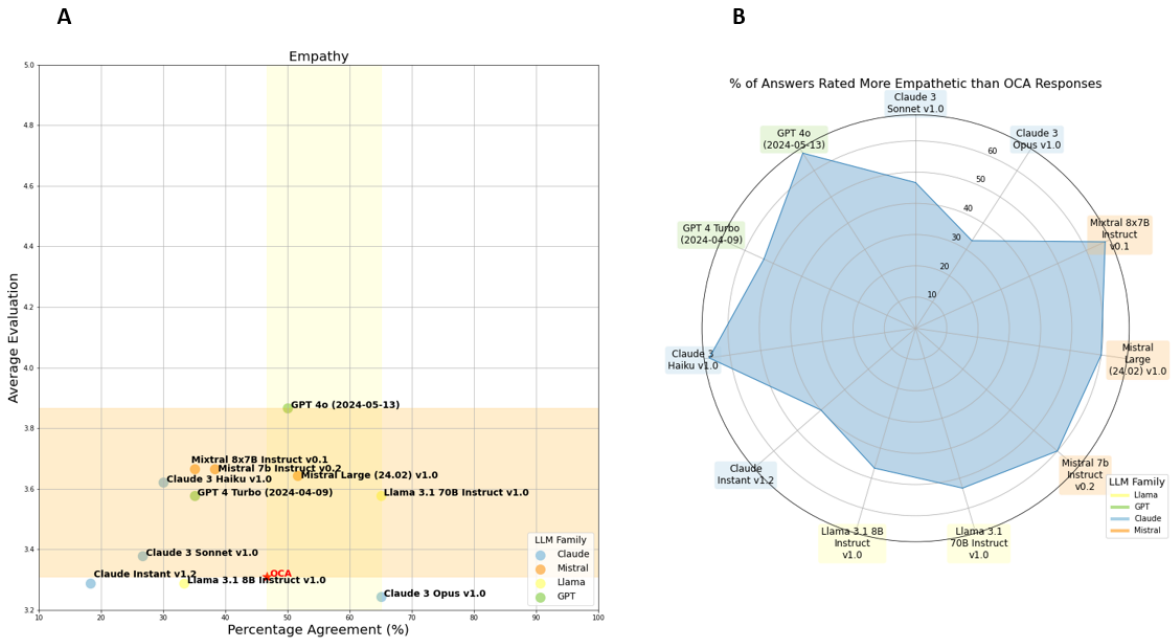

Figure 5. A) Scatter plot percentage agreement with average evaluation (empathy) B) Spider chart (radar plot) with % of answers evaluated as more empathetic related to OCA site. The percentage has been calculated based on the pairwise agreement. The yellow shaded region indicates a threshold for desirable performance. Only the top-performing models exceed both the evaluation score and agreement percentage highlighted by this region. The blue shaded region in the radar plot represents the performance of the LLMs in relation to the percentage of answers evaluated as easier to understand.

3.5. Comparison with OCA Responses

3.5.1. Benchmarking Against OCA

Evaluators demonstrated strong agreement (83.34%) on a high ease of understanding score (4.31, 95% CI: 4.08–4.59) for the responses extracted from the OCA site (Figure 6a).

Only the advanced models Claude 3 Opus, GPT 4 Turbo, and Mistral achieved both higher mean scores and evaluator agreement than OCA, placing them in the top right quadrant of Figure 6a. Notably GPT 4 Turbo scored higher but comparatively lower than inter-evaluator agreement. Figure 6b shows the percentage of responses that were rated as easier to understand than the corresponding OCA answers. Claude 3 Sonnet showed the highest performance with 40% of responses judged easier to understand than OCA. Claude 3 Opus, Mixtral 8x7B, and Mistral Large followed exceeding OCA clarity in a third (33%) of questions. Claude 3 Haiku, GPT 4 Turbo, and GPT 4o performed competitively with 25% responses rated clearer. By contrast, Claude Instant, Llama 3.1 8B, and Llama 3.1 70B performed poorly rarely surpassing the OCA baseline with only 13% of responses judged clearer, consistent with their overall lower ease of understanding scores and greater variability.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 April 2026 doi:10.20944/preprints202604.0800.v1

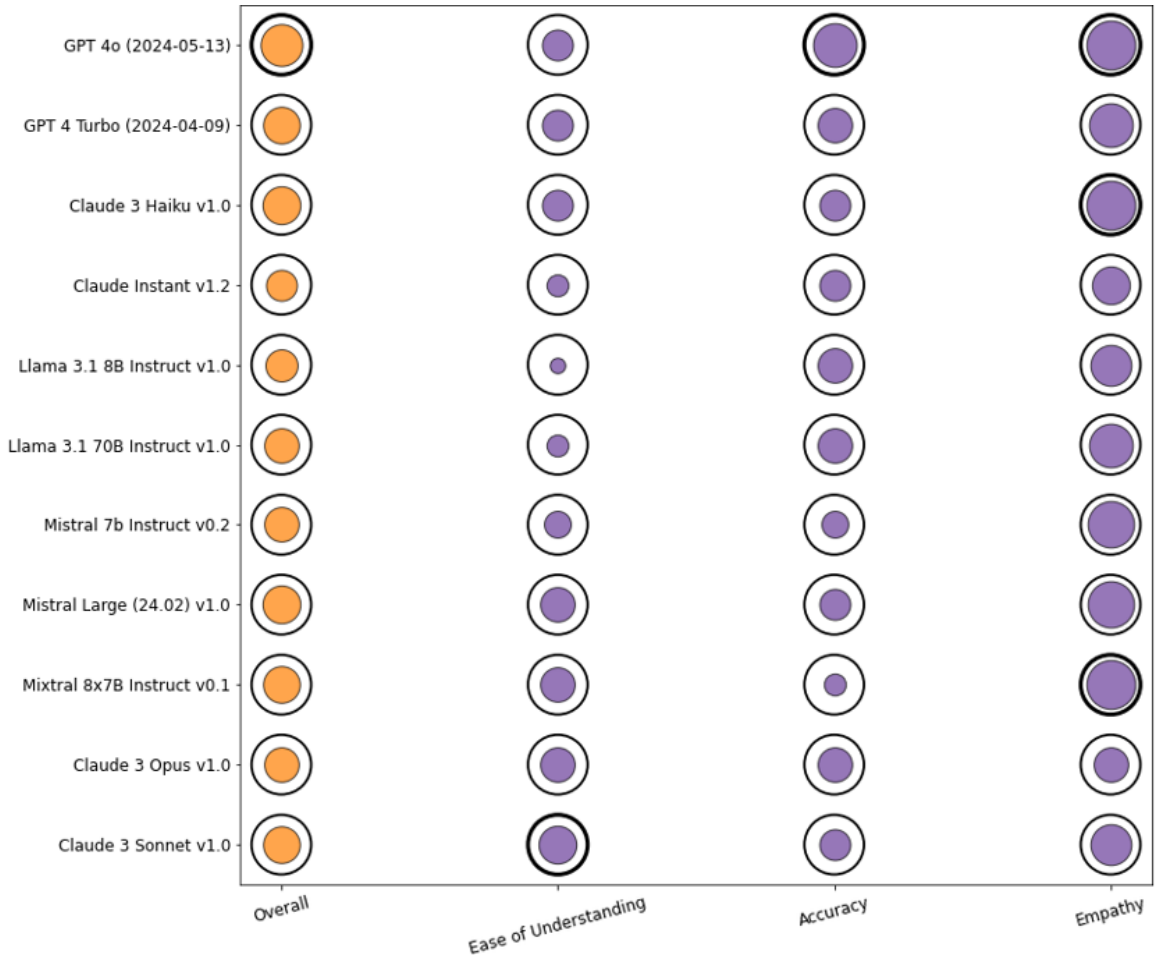

Figure 6. Spot matrix of the percentages of the answered questions judged by each LLM across the three criteria (ease of understanding, accuracy, and empathy) using the OCA response as a reference.

Accuracy

The responses of the OCA website achieved a good accuracy score (3.99, 95% CI: 3.69–4.18). However, inter-evaluator agreement was moderate (58.34%) suggesting some variability in how accuracy was interpreted (Figure 4a). OCA's accuracy score was the best among all models albeit GPT 4o followed closely (4.24, 95%, CI: 3.97–4.52), but with much stronger evaluator agreement (80%). In addition, 50% of GPT 4o's responses were rated to be more accurate than the OCA website's answers (Figure 4b). For all other models apart from Mixtral 8x7B, the percentage of questions that were rated as more accurate than the OCA answers was in the range of 26% to 34%. Mixtral performed poorly in this respect, only rarely exceeding the OCA baseline ($<$ 14% responses were rated as more accurate).

Empathy

The OCA website performed relatively poorly on both the mean empathy score (3.31, 95% CI: 2.78–3.48) and inter-evaluator agreement (46.67%) (Figure 6a). This suggests that the content whilst being perceived as correct and official, was not regarded as very empathetic. Consistent with this observation, most language models produced a higher proportion of responses rated as more empathetic than those from the OCA website (Figure 6b). Except for Claude 3 Opus, for which 33.33% of responses were rated as more empathetic than OCA, all other models exceeded the OCA reference, with 40%–70% of responses judged to demonstrate greater empathy.

4. Discussion

Our study highlights both the potential and the limitations of current LLMs in responding to key questions about ovarian cancer. As LLMs increasingly integrate into healthcare, assessing their

 doi:10.20944/preprints202604.0800.v1

performance is key to ensuring patients receive clear, accurate, and supportive information. Inaccurate responses may reflect some variability in their training data, which relies on heterogenous internet content rather than data sourced from scientific societies [13,14]. This may also reflect the intrinsic complexity of certain questions or gaps in the training data where relevant domain-specific information is limited or absent [15].

4.1. Consolidated Findings

Our results showed that no single LLM excelled uniformly across all three domains. This observation agrees with broader trends in AI language generation, where different models often excel in distinct domains [16]. The Friedman test confirmed statistically significant differences in at least one domain for each model. The evaluators agreed that, while Claude 3 Opus scored the highest for ease of understanding, GPT 4o was the most accurate and demonstrated the greatest empathy. This pattern suggests that the strengths of individual models might stem from their underlying training objectives and data sources, each primed for differing priorities. Some models prioritise clarity, others accuracy, and some aim for a more conversational and supportive tone.

4.2. Interpretation of Key Findings

By comparing LLMs with different data sources and architectural features, we provide a broad overview of how LLMs handle complex, sensitive medical content akin to similar efforts [17]. Readability and ease of understanding are related but distinct concepts. Readability reflects linguistic simplicity, while ease of understanding includes clarity of ideas and context. Some LLMs prioritise clarity and concision through training and design. GPT-4o's excellent accuracy likely derives from its advanced language and reasoning capacity, likely shaped by extensive fine-tuning on reputable medical content [18,19]. Similar patterns have been observed in ophthalmology studies [20]. Ratings for ease of understanding were relatively consistent among assessors albeit accuracy scores occasionally diverged, especially when answers incorporated complex statistics or referenced unverified data. Empathy ratings showed the greatest variability. Some evaluators found certain responses warm and compassionate, whereas others perceived the very same text as overly scripted or generic.

5. Assessing Empathy

Llama 3.1 70b's relative strength in empathy suggests that acknowledging emotional distress or offering reassurance can be learned and deployed effectively, even if the responses can feel somewhat scripted [21]. Indeed, empathy was the most difficult domain to quantitively assess. Medical communication research emphasises that empathy can mitigate anxiety, improve patient satisfaction, and enhance treatment adherence [22]. However, poorly framed or inaccurate information may exacerbate emotional distress [23]. Empathy requires more than polite phrases or cautionary disclaimers; it depends on recognising and responding to a patient's emotional state using predefined markers such as acknowledgement of uncertainty, reassurance, and patient-centred language. Although Llama 3.1 70b performed better in this regard, assessor feedback noted occasional reliance on repetitive patterns, suggesting a lack of context-specific emotional response. While aspirations to develop artificial empathy are commendable, they should complement genuine human empathy to avoid further patient isolation [24]. Future improvements may involve curated datasets or refined prompts focusing on empathetic dialogues or integrated frameworks from mental health counselling.

Scatterplot analyses (Figure 2, Figure 3a, Figure 4a) indicated that OCA underperformed in empathy compared to GPT-4o but performs reasonably well in ease of understanding and accuracy. This likely reflects its clinician-led development, which prioritises objective and factually correct responses over emotional engagement [25]. The frequent use of scientific terms may also increase complexity and hinder ease of understanding. In contrast, LLMs generate responses based on probabilistic patterns in large training datasets to mimic human-like communication, favouring a conversational tone, empathy, and simplicity [26].

 doi:10.20944/preprints202604.0800.v1

5.1. Implications for Clinical Practice

The use of LLMs to answer FAQs could streamline patient education, but even minor inaccuracies highlight the need for careful validation and regular updates to stay aligned with best practices [27]. In fast-evolving fields like oncology, an LLM's "knowledge base" can rapidly become outdated. The role of LLMs in direct patient communication should supplement rather than substitute professional advice. Patients seeking answers about symptoms may value immediate and readily accessible responses. However, urgent matters will still require direct consultation with healthcare professionals. The empathy shortfall observed in many models necessitates human oversight, especially in emotionally sensitive situations.

5.2. Methodological Considerations

The main stronghold of our study was the evaluation of a substantial number of state-of-art LLMs. Our analysis was enriched with performance visualisation using scatter plots, radar plots and spot matrices that can be used for model selection based on specific use cases requiring high agreement or output quality. Another strength of our study was the use of accredited gynaecological oncology experts to assess model outputs, thus reducing reliance on a single perspective. Despite their subjective ratings, they acted as domain experts, refining the responses provided by the LLMs and the OCA website. Their percentage agreement was used to determine an "acceptance rate" that could be used heuristically to build quality models.

6. Limitations and Future Directions

Several limitations must be acknowledged. First, although 11 LLMs were evaluated, newer or less publicised models were not included. In a global context, a focus on English language models may be a limiting factor. Previous research has shown that chatbots, when operating within an English language context, possess the capability to provide accurate responses [28]. Nevertheless, ovarian cancer affects populations worldwide and exploring multilingual LLMs to provide accurate and empathetic information is a critical avenue for future research. Second, the 15 questions may not capture the entire spectrum of domain-oriented ovarian cancer questionnaires. We relied on a standard question-and-answer format without exploring interactive or follow-up queries that might reveal how models adapt to evolving user input. This format does not address many important elements in human communication. In real-world settings, patients usually prefer to acquire information through continuous dialogues rather than isolated inquiries [29]. Third, all prompts were deterministic and run at zero temperature which may not reflect how these models perform under real-world prompting conditions. Fourth, the use of a Likert scale for readability analysis may limit the generalisability of these data. Literacy and comprehension are closely linked to socioeconomic status, factors not accounted for [30]. Although expert evaluation strengthened domain validity, the small number of assessors meant that other gynaecologic oncologists may hold differing opinions regarding accuracy. Patient and public involvement were not incorporated into the study design, interpretation or analysis of findings. These limitations highlight the need for large-scale, multi-centre, prospective studies to better evaluate the accuracy, generalizability and clinical utility of LLMs, and to develop standardised frameworks for their assessment.

Ultimately, these findings can support the development of specialised medical LLMs for oncology, where accuracy and and empathetic communication are critical. Our future aim is to develop a bespoke LLM, trained exclusively on curated up-to-date ovarian cancer content, ensuring both accuracy and a suitable bedside empathy with ongoing input from oncologists, psychologists, and AI specialists. We are currently engaging with local patient advocacy groups to provide valuable perspectives on the language and emotional tone that best supports those affected by ovarian cancer. That said, concerns about data privacy and information misuse remain. Their implementation in healthcare settings must be accompanied by stringent safeguards to protect patient confidentiality and ensure that any advice provided aligns with professional ethical standards.

 doi:10.20944/preprints202604.0800.v1

7. Conclusions

In evaluating 11 LLMs across 15 FAQs about ovarian cancer, we found that each model excelled in at least one domain; ease of understanding, accuracy, or empathy, but no single model delivered consistently strong results across all three. Claude 3 Opus presented the easiest-to-understand responses, GPT-4o offered the greatest accuracy, and Llama 3.1 70b displayed relatively the highest empathy. While certain models show promise in accuracy or readability, the capacity to demonstrate genuine empathy remains an area in need of further refinement. A dedicated ovarian cancer LLM trained in-domain text could deliver accurate, empathetic, and evidence-based information about ovarian cancer, while complementing clinical expertise.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1: Ovarian Cancer Action fundamental questions; Table S2: Overview of access types and model specifications for the investigated Large Language Models (LLMs).

Author Contributions: Conceptualization, M.Q. and A.L.; Methodology, M.Q., Y.T. and E.K.; Software, Y.T. and A.M.; Validation, M.Q., Y.T. and A.L.; Formal Analysis, Y.T. and A.M.; Investigation, M.Q., Y.T. and A.L.; Data Curation, Y.T.; Writing—Original Draft Preparation, M.Q. and A.L.; Writing—Review & Editing, all authors; Visualization, Y.T. and A.M.; Supervision, N.O., D.D. and A.L.; Project Administration, A.L.; Funding Acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

LLM	Large Language Model
FAQ	Frequently Asked Question
OCA	Ovarian Cancer Action
AI	Artificial Intelligence
AWS	Amazon Web Services
RAG	Retrieval-Augmented Generation
CI	Confidence Interval
GPT	Generative Pre-trained Transformer
ESGO	European Society of Gynaecological Oncology

References

1. Webb, P.M.; Jordan, S.J. Global epidemiology of epithelial ovarian cancer. *Nat. Rev. Clin. Oncol.* **2024**, *21*, 389–400. doi:10.1038/s41571-024-00881-3.
2. Ledermann, J.A.; Raja, F.A.; Fotopoulou, C.; et al. Newly diagnosed and relapsed epithelial ovarian carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2013**, *24 Suppl 6*, vi24–vi32.
3. Iannantuono, G.M.; Bracken-Clarke, D.; Floudas, C.S.; et al. Applications of large language models in cancer care: current evidence and future perspectives. *Front. Oncol.* **2023**, *13*, 1268915. doi:10.3389/fonc.2023.1268915.
4. Hirani, R.; Noruzi, K.; Khuram, H.; Hussaini, A.S.; Aifuwa, E.I.; Ely, K.E.; Lewis, J.M.; Gabr, A.E.; Smiley, A.; Tiwari, R.K.; Etienne, M. Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life (Basel)* **2024**, *14*, 557. doi:10.3390/life14050557.

 doi:10.20944/preprints202604.0800.v1

5. Klug, K.; Beckh, K.; Antweiler, D.; Chakraborty, N.; Baldini, G.; Laue, K.; Hosch, R.; Nensa, F.; Schuler, M.; Giesselbach, S. From Admission to Discharge: A Systematic Review of Clinical Natural Language Processing along the Patient Journey. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 238. doi:10.1186/s12911-024-02641-w.
6. Brown, T.; Mann, B.; Ryder, N.; et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
7. Clarke, M.; Gregg, J.; Day, S.; et al. Natural language processing in oncology: exploring challenges and opportunities. *J. Oncol. Inform.* **2021**, *5*, 27–34.
8. Ovarian Cancer Action. Available online: https://ovarian.org.uk/ (accessed on 24 January 2025).
9. Benary, M.; Wang, X.D.; Schmidt, M.; Soll, D.; Hilfenhaus, G.; Nassir, M.; Sigler, C.; Knödler, M.; Keller, U.; Beule, D.; Keilholz, U.; Leser, U.; Rieke, D.T. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw. Open* **2023**, *6*, e2343689. doi:10.1001/jamanetworkopen.2023.43689.
10. Huh, W.K.; Blackwell, K.; Merino, M.J.; et al. A prospective analysis of the communication of supportive care in advanced cancer. *Cancer* **2008**, *113*, 550–558.
11. Tanaka, E.; O'Connor, A.E.; Roston, A.; McClements, M.E. Evaluating empathy in AI-driven chatbots: a systematic review. *Health Informatics J.* **2023**, *29*, 146045822311482.
12. Eriksen, A.V.; Möller, S.; Ryg, J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* **2024**, *1*.
13. Querleu, D.; Planchamp, F.; Chiva, L.; Fotopoulou, C.; Barton, D.; Cibula, D.; Aletti, G.; Carinelli, S.; Creutzberg, C.; Davidson, B.; Harter, P.; Lundvall, L.; Marth, C.; Morice, P.; Rafii, A.; Ray-Coquard, I.; Rockall, A.; Sessa, C.; van der Zee, A.; Vergote, I.; duBois, A. European Society of Gynaecological Oncology (ESGO) Guidelines for Ovarian Cancer Surgery. *Int. J. Gynecol. Cancer* **2017**, *27*, 1534–1542. doi:10.1097/IGC.0000000000001041.
14. Moss, E.; Taylor, A.; Andreou, A.; Ang, C.; Arora, R.; Attygalle, A.; Banerjee, S.; Bowen, R.; Buckley, L.; Burbos, N.; Coleridge, S.; Edmondson, R.; El-Bahrawy, M.; Fotopoulou, C.; Frost, J.; Ganesan, R.; George, A.; Hanna, L.; Kaur, B.; Manchanda, R.; Maxwell, H.; Michael, A.; Miles, T.; Newton, C.; Nicum, S.; Ratnavelu, N.; Ryan, N.; Sundar, S.; Vroobel, K.; Walther, A.; Wong, J.; Morrison, J. British Gynaecological Cancer Society (BGCS) ovarian, tubal and primary peritoneal cancer guidelines: Recommendations for practice update 2024. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2024**, *300*, 69–123. doi:10.1016/j.ejogrb.2024.06.025.
15. Rahsepar, A.A.; Tavakoli, N.; Kim, G.H.J.; Hassani, C.; Abtin, F.; Bedayat, A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* **2023**, *307*, e230922. doi:10.1148/radiol.230922.
16. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; et al. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. *arXiv* **2022**, arXiv:2203.03540.
17. Liu, M.; Okuhara, T.; Dai, Z.; Huang, W.; Gu, L.; Okada, H.; Furukawa, E.; Kiuchi, T. Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. *Int. J. Med. Inform.* **2025**, *193*, 105673. doi:10.1016/j.ijmedinf.2024.105673.
18. Guo, Y.; Li, T.; Xie, J.; Luo, M.; Zheng, C. Evaluating the accuracy, time and cost of GPT-4 and GPT-4o in liver disease diagnoses using cases from "What is Your Diagnosis". *J. Hepatol.* **2025**, *82*, e15–e17. doi:10.1016/j.jhep.2024.09.016.
19. Wang, H.; Lan, J. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists. *Am. J. Emerg. Med.* **2025**, *87*, 197. doi:10.1016/j.ajem.2024.09.041.
20. Shi, R.; Liu, S.; Xu, X.; Ye, Z.; Yang, J.; Le, Q.; Qiu, J.; Tian, L.; Wei, A.; Shan, K.; Zhao, C.; Sun, X.; Zhou, X.; Hong, J. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. *Heliyon* **2024**, *10*, e34391. doi:10.1016/j.heliyon.2024.e34391.
21. Klang, E.; Tessler, I.; Apakama, D.U.; Abbott, E.; Glicksberg, B.S.; Arnold, M.; Moses, A.; Sakhuja, A.; Soroush, A.; Charney, A.W.; et al. Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders. *medRxiv* **2024**. doi:10.1101/2024.10.15.24315526.
22. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; et al. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. *arXiv* **2022**, arXiv:2203.03540. doi:10.48550/arXiv.2203.03540.
23. Danaher, S.; Berry, L.; Howard, C.; Moore, G.; Attai, J. Improving How Clinicians Communicate With Patients: An Integrative Review and Framework. *J. Serv. Res.* **2023**, *26*, 493–510. doi:10.1177/10946705231190018.
24. Lu, D.; Fall, K.; Sparén, P.; Ye, W.; Adami, H.O.; Valdimarsdóttir, U.; Fang, F. Suicide and suicide attempt after a cancer diagnosis among young individuals. *Ann. Oncol.* **2013**, *24*, 3112–3117. doi:10.1093/annonc/mdt415.

 doi:10.20944/preprints202604.0800.v1

25. Koranteng, E.; Rao, A.; Flores, E.; Lev, M.; Landman, A.; Dreyer, K.; Succi, M. Empathy and Equity: Key Considerations for Large Language Model Adoption in Health Care. *JMIR Med. Educ.* **2023**, *9*, e51199. doi:10.2196/51199.
26. Lee, J.E.; Park, K.S.; Kim, Y.H.; Song, H.C.; Park, B.; Jeong, Y.J. Lung Cancer Staging Using Chest CT and FDG PET/CT Free-Text Reports: Comparison Among Three ChatGPT Large Language Models and Six Human Readers of Varying Experience. *AJR Am. J. Roentgenol.* **2024**, *223*, e2431696. doi:10.2214/AJR.24.31696.
27. Calvert, M.J.; Moher, D.; Wiseman, T.; et al. Reporting guidelines for patient and public involvement research: the need for international consensus. *BMJ* **2022**, *378*, e072279. doi:10.1136/bmj-2022-072279.
28. Zhang, Z.; Huang, X. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon* **2024**, *10*, e25370. doi:10.1016/j.heliyon.2024.e25370.
29. Jones, N. How should we test AI for human-level intelligence? OpenAI's o3 electrifies quest. *Nature* **2025**. doi:10.1038/d41586-025-00110-6.
30. Alejos, D.; Tregubenko, P.; Jayarangaiah, A.; Steinberg, L.; Kumar, A. We need to do better: Readability analysis of online patient information on cancer survivorship and fertility preservation. *J. Cancer Policy* **2021**, *28*, 100276. doi:10.1016/j.jcpo.2021.100276.