npj | women's health

Article

https://doi.org/10.1038/s44294-024-00028-w

Exploring the capabilities of ChatGPT in women's health: obstetrics and gynaecology

Magdalena Bachmann[1], Ioana Duta[1], Emily Mazey[1], William Cooke[1], Manu Vatish[1] & Gabriel Davis Jones[1,2] ✉

Artificial Intelligence (AI) is transforming healthcare, with Large Language Models (LLMs) like ChatGPT offering novel capabilities. This study evaluates ChatGPT's performance in interpreting and responding to the UK Royal College of Obstetricians and Gynaecologists MRCOG Part One and Two examinations – international benchmarks for assessing knowledge and clinical reasoning in Obstetrics and Gynaecology. We analysed ChatGPT's domain-specific accuracy, the impact of linguistic complexity, and its self-assessment confidence. A dataset of 1824 MRCOG questions was curated, ensuring minimal prior exposure to ChatGPT. ChatGPT's responses were compared to known correct answers, and linguistic complexity was assessed using token counts and Type-Token ratios. Confidence scores were assigned by ChatGPT and analysed for self-assessment accuracy. ChatGPT achieved 72.2% accuracy on Part One and 50.4% on Part Two, performing better on Single Best Answer (SBA) than Extended Matching (EMQ) Questions. The findings highlight the potential and significant limitations of ChatGPT in clinical decision-making in women's health.

Artificial Intelligence (AI) has emerged as a transformative technology in healthcare. At the forefront of this AI revolution are Large Language Models (LLMs), powerful systems designed to mimic human language processing abilities. In recent years, the utility of text-data based models such as LLMs in health and medicine has garnered significant interest. These models can process and generate human-like text, enabling them to assist in a variety of healthcare applications, from patient education to clinical decision support. Their ability to quickly analyse large datasets and provide coherent, contextually relevant responses holds promise for addressing existing gaps in healthcare, particularly in areas requiring quick and accurate information retrieval and synthesis. These LLMs, trained on vast volumes of data encompassing books, articles, websites and other media possess the potential to drive advancements in medicine, a field where precision and safety are paramount. Chat Generative Pre-trained Transformer (ChatGPT) has recently emerged as the prominent LLM. ChatGPT, first released to the public in November 2022 by OpenAI, represents a significant advancement in the field of Natural Language Processing (NLP)[1]. This large-scale, multimodal model is adept at understanding and generating text that closely resembles human language, making it a potentially valuable tool in medicine and healthcare[2,3].

Women's health, specifically Obstetrics and Gynaecology (O&G), is a medical domain poised to derive significant benefit. O&G, a field with a history of significant diagnostic and treatment gaps[4–8], could leverage LLMs to bridge these disparities. AI could aid in analysing patient histories, imaging, and test results to assist in early and accurate diagnoses. Additionally, AI-driven tools could provide personalised treatment options by processing large datasets to predict the most effective interventions for individual patients. The utilisation of LLMs in O&G not only offers the potential to enhance patient outcomes but also democratise healthcare knowledge, narrowing the existing health inequity gap.

However, the benefits of ChatGPT must be tempered by an acute awareness of its limitations, especially within the complex landscape of healthcare. ChatGPT has been described as a "jack of all trades, master of none"[9]. Nonetheless, it is already being explored by doctors and patients as an adjunct to the traditional medical pathway[10–14]. Ethical concerns regarding this technology are more prevalent than ever, encompassing issues of bias, information governance, patient confidentiality, transparency and accountability[15,16]. ChatGPT's propensity to generate content that is convincing yet factually incorrect, often referred to as "hallucinations," further complicates its potential utility in medical settings. The model's inability to provide a rationale for erroneous decisions further complicates matters, raising concerns about safety, interpretability, reproducibility and the handling of uncertainty, all of which could have profound implications for patients. While ChatGPT holds immense promise, its application in

[1]Nuffield Department of Women's & Reproductive Health, University of Oxford, Women's Centre, John Radcliffe Hospital, Oxford, OX3 9DU, UK. [2]The Alan Turing Institute, London, UK. ✉e-mail: gabriel.jones@wrh.ox.ac.uk

healthcare requires a careful and thorough evaluation to ensure both its reliability and its limitations are understood.

The O&G specialty training programme in the UK is a structured, continuous educational path that spans seven years. It combines both basic and advanced training stages[17]. Training begins after a doctor has completed their initial medical training, gained foundational competencies over two years of work and achieved full registration with the General Medical Council (GMC)[18]. During the programme, trainees are required to pass three key exams (MRCOG Parts One, Two, and Three) at different stages, which assess their clinical knowledge, reasoning and skills in O&G[19]. These exams are also formal requirements in other international O&G training programmes, with over 100 MRCOG test centres outside the UK[20,21].

The MRCOG Part One exam is designed to assess trainees' foundational scientific knowledge. This exam covers four key knowledge domains: cell function, human structure, measurement and manipulation, and understanding illness, encompassing various subjects including physiology, anatomy, biophysics, and clinical management[22]. The MRCOG Part Two exam advances the assessment to a more practical level, testing the application of the knowledge acquired, i.e. clinical reasoning[22]. It comprises a mixture of single best answer (SBA) and extended matching questions (EMQ). These evaluate the trainee's theoretical understanding as well as their ability to apply this knowledge in practical scenarios. The combination of these question types ensures a comprehensive assessment of the trainee's capabilities in O&G, preparing them for advanced practice in the field. The MRCOG exams hold significant international recognition and are widely regarded as a gold standard qualification in O&G. Achieving the MRCOG qualification after a medical degree is regarded as a benchmark of medical competence.

The nature of questions found in the MRCOG examinations, specifically SBAs and EMQs, provide a pertinent framework for gauging the capabilities of LLMs such as ChatGPT. These formats are particularly challenging because they often present multiple answers that could all be considered correct. Clinicians must draw upon not only their knowledge, but also their clinical reasoning and experience to discern the most appropriate answer from among various plausible options. Thus, when ChatGPT is tasked with identifying the single best answer, it undergoes a rigorous test of its clinical reasoning abilities. This goes beyond simple recollection of information, requiring instead the application of knowledge to a defined clinical context, as per the standards established by the RCOG and accepted clinical practice.

The objectives of this study were threefold: Firstly, to assess the efficacy of ChatGPT in interpreting and responding to questions from the MRCOG Part One and Part Two examinations, thus evaluating its domain-specific accuracy in a standardised medical knowledge and reasoning context. Secondly, to determine whether the complexity of the questions influences ChatGPT's performance accuracy, thereby enabling an analysis of its clinical knowledge and reasoning capabilities independent of linguistic difficulty. Thirdly, to investigate ChatGPT's self-assessment of confidence in its responses, providing insight into the reliability and safety of AI in clinical decision-making processes. This self-evaluation aspect is particularly crucial, as it could reflect the model's ability to estimate its certainty and, by extension, its utility in real-world medical applications where the cost of error is potentially high.

Results

1824 MRCOG Part One and Part Two questions from eight sources were extracted and converted into a format readable for ChatGPT. 835 MRCOG Part One single best answer (SBA) questions and 989 MRCOG Part Two questions (589 SBAs and 400 extended matching questions [EMQ]) were identified. The range of answer choices for SBA questions was between A–E (5 options) while the range for EMQs was between 5–18 choices (A–R). 56 questions (3.1%) contained additional tabular data which were converted to JSON format. 4 questions with associated images were omitted. Questions were identified for each of the areas of knowledge prescribed by the RCOG (14 for Part One and 15 for Part Two examinations). The median number of questions in each knowledge area for Parts One and Two were 58 (IQR 32–85) and 45 (IQR 32–64). See Supplementary Tables 1 & 2 for the distribution of knowledge areas.

ChatGPT performance accuracy

ChatGPT achieved an overall accuracy of 72.2% (95% CI 69.2–75.3, 603/835 correct) on Part One and 50.4% (95% CI 47.2–53.5, 534/989 correct) on Part Two of the MRCOG examinations. Across the four domains of understanding for the MRCOG Part One examination (Table 1, Fig. 1), there was a significant difference in the accuracy of ChatGPT ($p = 0.02$, $\chi^2 = 9.85$). ChatGPT performed best in the "Illness" domain with an accuracy of 80.0% (95% CI 73.3–85.7) and worst in the "Measurement and Manipulation" domain with an accuracy of 65.7% (95% CI 58.8–72.7). We then evaluated the accuracy of ChatGPT in the subjects constituting these domains (Table 2, Fig. 1). There was no significant difference between each subject within any domain (Domain-specific p-values: Cell Function, $p = 0.08$; Human Structure, $p = 0.07$; Illness, $p = 0.49$; Measurement and Manipulation, $p = 0.11$, Table 2). For each domain, ChatGPT demonstrated the highest accuracy in Biochemistry (79.8% [95% CI 71.4–88.1], Cell Function), Embryology (80.4% [95% CI 70.0–90.8], Human Structure), Clinical Management (83.3% [95% CI 68.4–98.2], Understanding Illness) and Pharmacology (75.4% [95% CI 64.3–86.6], Measurement and Manipulation). The subjects ChatGPT performed worst in within each domain were Physiology (65.3% [95% CI 56.1–74.6], Illness), Anatomy (63.2% [95% CI 54.0–72.4], Human Structure, Immunology (70.0% [95% CI 53.6–86.4], Illness) and Biophysics (51.4% [95% CI 35.2–67.5], Measurement and Manipulation).

For Part Two, the RCOG does not assign subjects to discrete domains, as subjects and questions can span multiple domains. Therefore ChatGPT's performance was assessed by subject only. The accuracy across subjects did not vary significantly ($p = 0.10$, $\chi^2 = 21.05$, Table 3, Fig. 2). The best performing knowledge area was Urogynaecology & Pelvic Floor Problems (accuracy 63.0% [95% CI 50.1–75.8]) while the worst performing area was Management of Labour (accuracy 35.6% [95% CI 21.6–49.5]. ChatGPT performed better at SBA questions (54.0% accurate [95% CI 50.0-58.0]) than EMQ questions (45.0% accurate [95% CI 40.1–49.9], $p = 0.01$, $\chi^2 = 7.35$, Table 4).

Influence of linguistic complexity on ChatGPT performance

We next evaluated whether the linguistic complexity of the questions given to ChatGPT could influence its performance. Each question was tokenised and the unique token count and type-token ratio (TTR) were calculated (Table 5). For the MRCOG Part One, the median unique token count was marginally higher for correct responses (122 [IQR 114–134]) compared to incorrect responses (120 [IQR 112–131]), with a small effect size of -2 and a p-value of 0.05, indicating a statistically significant but minor difference. In Part Two, no significant difference was found in the unique token count between correct and incorrect responses ($p = 0.60$). A statistically significant

Table 1 | ChatGPT performance accuracy across the four domains of the MRCOG part one examination

Domain	Correct	Incorrect	Total
Cell Function	203 (72.8%)	76 (27.2%)	279
Human Structure	135 (69.9%)	58 (30.1%)	193
Illness	148 (80.0%)	37 (20.0%)	185
Measurement and Manipulation	117 (65.7%)	61 (34.3%)	178
Total	603 (72.2%)	232 (27.8%)	835

The overall accuracy was 72.2% (95% CI 69.2–75.3). There was a significant difference in the accuracy of ChatGPT across the four domains ($p = 0.02$, Chi-squared statistic = 9.85). ChatGPT performed best in the "Illness" domain with an accuracy of 80.0% (95% CI 73.3–85.7) and worst in the "Measurement and Manipulation" domain with an accuracy of 65.7% (95% CI 58.8–72.7). Values in brackets denote the percentage proportion (%).

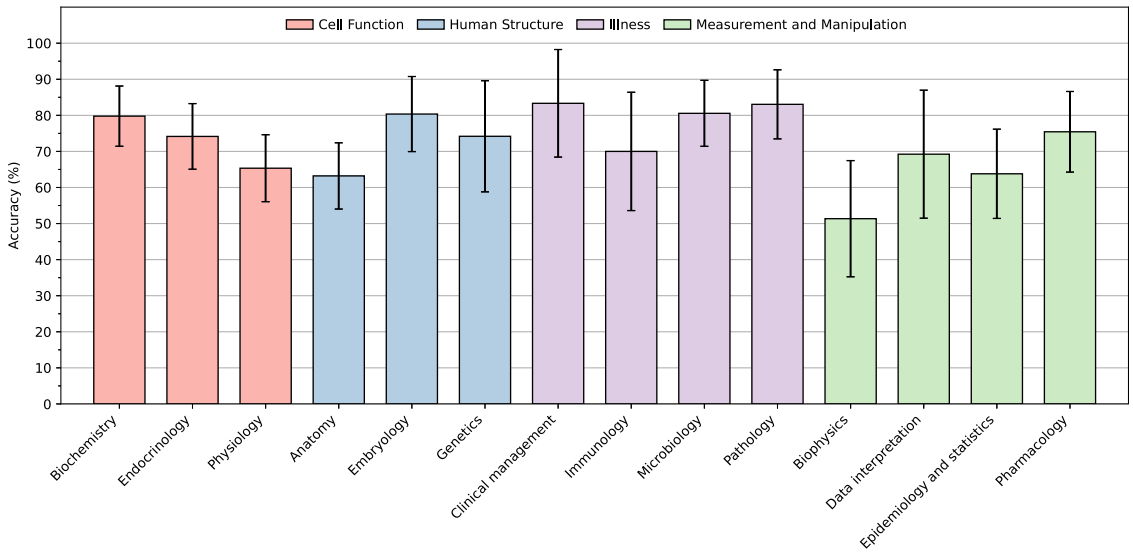

Fig. 1 | Performance of ChatGPT on the MRCOG Part One examination. Significant variance in performance across the four domains was noted ($p = 0.02$; $\chi^2 = 9.85$). The highest accuracy was observed in the domain of "Illness" at 80.0% (95% Confidence Interval [CI]: 73.3–85.7), whereas the lowest was in "Measurement and Manipulation" at 65.7% (95% CI: 58.8–72.7). Analysis of ChatGPT's accuracy within individual subjects corresponding to these domains revealed no substantial differences (Domain-specific p-values: Cell Function, $p = 0.08$; Human Structure, $p = 0.07$; Illness, $p = 0.49$; Measurement and Manipulation, $p = 0.11$). Within each domain, subjects with the highest accuracy were Biochemistry (79.8% [95% CI: 71.4–88.1], Cell Function), Embryology (80.4% [95% CI: 70.0–90.8], Human Structure), Clinical Management (83.3% [95% CI: 68.4–98.2], Illness), and Pharmacology (75.4% [95% CI: 64.3–86.6], Measurement and Manipulation). Subjects with the lowest accuracy were Physiology (65.3% [95% CI: 56.1–74.6], Illness), Anatomy (63.2% [95% CI: 54.0–72.4], Human Structure), Immunology (70.0% [95% CI: 53.6–86.4], Illness), and Biophysics (51.4% [95% CI: 35.2–67.5], Measurement and Manipulation).

Table 2 | ChatGPT performance accuracy in each subject comprising the MRCOG part one domains

Domain	Knowledge area	Correct	Incorrect	Total	*p*-value
Cell Function	Biochemistry	71 (79.8%)	18 (20.2%)	89	0.08
	Endocrinology	66 (74.2%)	23 (25.8%)	89	
	Physiology	66 (65.3%)	35 (34.7%)	101	
HumanStructure	Anatomy	67 (63.2%)	39 (36.8%)	106	0.07
	Embryology	45 (80.4%)	11 (19.6%)	56	
	Genetics	23 (74.2%)	8 (25.8%)	31	
Illness	Clinical management	20 (83.3%)	4 (16.7%)	24	0.49
	Immunology	21 (70.0%)	9 (30.0%)	30	
	Microbiology	58 (80.6%)	14 (19.4%)	72	
	Pathology	49 (83.1%)	10 (16.9%)	59	
Measurement & Manipulation	Biophysics	19 (51.4%)	18 (48.6%)	37	0.11
	Data interpretation	18 (69.2%)	8 (30.8%)	26	
	Epidemiology and statistics	37 (63.8%)	21 (36.2%)	58	
	Pharmacology	43 (75.4%)	14 (24.6%)	57	

There was a significant difference in the accuracy of ChatGPT across the four domains ($p = 0.02$, Chi-squared statistic = 9.85), however the performance of each subject within any domain was not significantly different. Values in brackets denote the percentage proportion (%).

difference was observed for TTR. In Part One, correct responses had a slightly higher median TTR (0.66 [IQR 0.63–0.68]) compared with incorrect responses (0.65 [IQR 0.62–0.67]), with a negligible effect size of − 0.01 ($p < 0.001$). Similarly, for Part Two, correct responses had a median TTR of 0.62 (IQR 0.57–0.67), which was marginally higher than the 0.59 (IQR 0.54–0.65) of incorrect responses, with an effect size of − 0.03 ($p < 0.001$). These findings suggest that the linguistic complexity, as measured by unique token count and TTR, has a statistically significant association with the accuracy of responses. However, the effect sizes indicate that the actual difference in linguistic complexity between correct and incorrect responses is not substantial enough to meaningfully influence ChatGPT's performance.

Confidence and uncertainty in ChatGPT responses

In the evaluation of ChatGPT's self-assessment of confidence, it was observed that for 192 questions, representing 10.5% of the total, ChatGPT allocated an identical probability score to each answer option. This suggests a lack of discriminatory power, which may indicate that ChatGPT was either equally uncertain about all options or failed to parse the question correctly. These instances are explored in the discussion to provide insights into the model's limitations and areas for improvement.

Of the remaining probabilities, 1072 were associated with correct answers accurately identified by ChatGPT. Conversely, 567 probabilities pertained to answers incorrectly identified as correct, another 567 were allocated to correct answers erroneously identified as incorrect, and 3100

probabilities corresponded to answers correctly identified as incorrect (Fig. 3). The high value of 3100 probabilities in this latter category is explained by the multiple incorrect answers available per falsely answered question. Specifically, each question in the MRCOG exams can have between 5 to 18 options, which leads to a higher count of probability scores in the categories involving multiple incorrect options.

The median confidence level for both correctly identified correct answers and incorrectly identified correct answers was 70.0% (Interquartile Range [IQR]: 60–90, $p < 0.001$). For correct answers misclassified as incorrect, the median confidence was 10.0% (IQR: 0–10), whereas for incorrect answers rightly identified as such, the median confidence was 5.0% (IQR: 0–10, $p < 0.001$). Despite statistical significance, the practical difference in confidence levels between these groups was minimal.

The median entropy for ChatGPT's correct responses (where ChatGPT's answer matched the correct exam answer) was 1.46 (IQR 0.44–1.77) and similarly, the median entropy for its incorrect responses (where ChatGPT's answer did not match the correct exam answer) was 1.46 (IQR: 0.67–1.77, $p < 0.001$, Fig. 4). The identical median values suggest that ChatGPT's distribution of probabilities does not discernibly distinguish between its correct and incorrect responses.

Table 3 | ChatGPT performance accuracy in the MRCOG part two

Knowledge area	Correct	Incorrect	Total
Antenatal care	48 (39.0%)	75 (61.0%)	123
Clinical skills	9 (52.9%)	8 (47.1%)	17
Core surgical skills	28 (40.0%)	42 (60.0%)	70
Early pregnancy care	24 (52.2%)	22 (47.8%)	46
Gynaecological oncology	29 (50.9%)	28 (49.1%)	57
Gynaecological problems	107 (51.2%)	102 (48.8%)	209
Management of delivery	17 (56.7%)	13 (43.3%)	30
Management of labour	16 (35.6%)	29 (64.4%)	45
Maternal medicine	95 (55.6%)	76 (44.4%)	171
Postoperative care	13 (56.5%)	10 (43.5%)	23
Postpartum problems	18 (56.2%)	14 (43.8%)	32
Sexual & reproductive health	17 (51.5%)	16 (48.5%)	33
Subfertility	18 (50.0%)	18 (50.0%)	36
Teaching & research	25 (58.1%)	18 (41.9%)	43
Urogynaecology & pelvic floor problems	34 (63.0%)	20 (37.0%)	54
Total	498	491	989

Part Two comprises single best answer (SBA) and extended matching questions (EMQ) from 15 knowledge areas (subjects). Accuracy across the knowledge areas did no vary significantly ($p = 0.10$, $\chi^2 = 21.05$). Values in brackets denote the percentage proportion (%).

Discussion

This study presents a novel and in-depth evaluation of the potential for LLMs as tools in women's health, specifically O&G. Leveraging a substantial dataset of questions from the Royal College of Obstetricians and Gynaecologists' MRCOG Part One and Part Two examinations, we have detailed a comprehensive analysis of ChatGPT's capabilities in understanding and applying medical knowledge and reasoning to an internationally-recognised standard of excellence. ChatGPT exhibited a notable level of proficiency in the MRCOG Part One examination, displaying an ability to evaluate medical content based on the current MRCOG syllabus[17]. The syllabus covers basic and applied science knowledge necessary for qualified medical professionals before they begin specialty training in O&G. In contrast, the Part Two examination, which tests candidates with several years of training in O&G on the application of their knowledge (i.e. clinical reasoning) to representative clinical scenarios, ChatGPT's performance was poorer. While ChatGPT outperformed random chance, its responses were, on average, as frequently incorrect as they were correct. The significant dif-

Table 4 | Comparing ChatGPT's performance accuracy between SBAs and EMQs

Question Type	Correct	Incorrect	Total
Single best answer (SBA)	318 (54.0%)	271 (46.0%)	589
Extended matching questions (EMQ)	180 (45.0%)	220 (55.0%)	400
Total	498	491	989

ChatGPT performed better in single best answer (SBA) questions than extended matching questions (EMQ), $p = 0.01$, $\chi^2 = 7.35$. Values in brackets denote the percentage proportion (%).

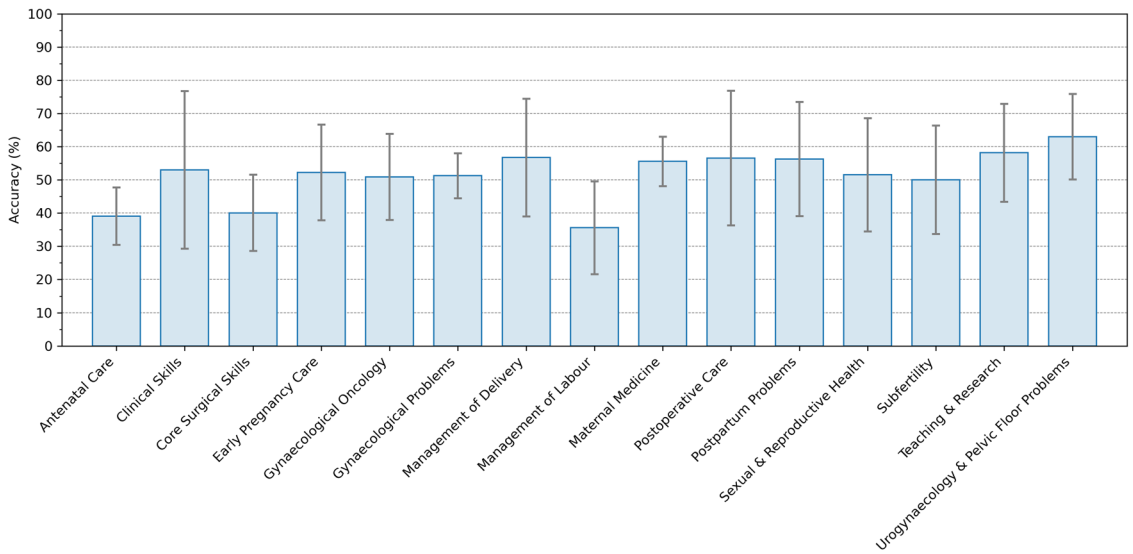

Fig. 2 | Performance of ChatGPT on the MRCOG part two examination. The distribution of accuracies across various knowledge areas showed no significant variation ($p = 0.10$, $\chi^2 = 21.05$). Urogynaecology & Pelvic Floor Problems emerged as the area with the highest accuracy at 63.0% (95% Confidence Interval [CI]: 50.1–75.8), contrasting with Management of Labour which had the lowest at 35.6% (95% CI: 21.6–49.5).

Table 5 | Evaluation of linguistic complexity as a factor in ChatGPT performance description

Linguistic complexity metric	MRCOG Part	Response category	Median (IQR)	Effect size	*p*-value
Unique Token count	One	Correct responses	122 (114–134)	−2	0.05
		Incorrect responses	120 (112–131)		
	Two	Correct responses	157 (135–182)	+1	0.60
		Incorrect responses	158 (134–188)		
Type-token ratio (TTR)	One	Correct responses	0.66 (0.63–0.68)	−0.01	< 0.001
		Incorrect responses	0.65 (0.62–0.67)		
	Two	Correct responses	0.62 (0.57–0.67)	−0.03	< 0.001
		Incorrect responses	0.59 (0.54–0.65)		

Values for unique token count have been rounded to whole numbers. Values for TTR have been rounded to 2 significant figures. Effect size describes the average difference between correct and incorrect responses for the linguistic complexity metrics. There was a significant difference observed for the unique token count metric for the MRCOG Part One whereby incorrect responses had on average two unique tokens less than correct responses. A significant different was observed for the type-token ratio, whereby incorrect responses had an average TTR of −0.01 less than correct responses for part one ($p < 0.001$) and −0.03 for part two. These results suggest that, while there was a significant difference, the effect size (actual difference in these values) was not of meaningful importance in ChatGPT's performance accuracy.

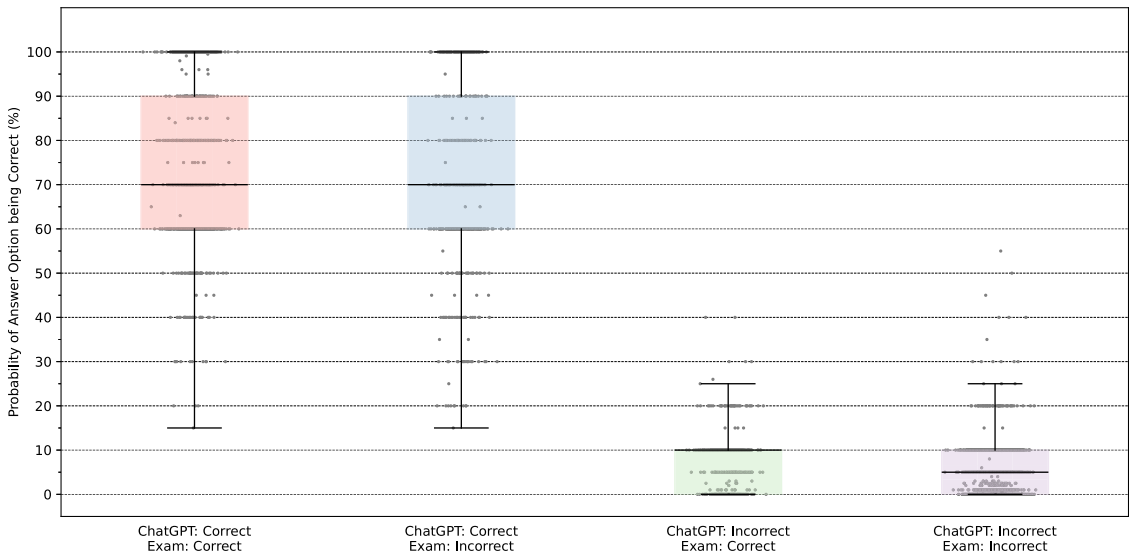

Fig. 3 | Box-and-Whisker Plot Depicting the Distribution of Confidence Scores Attributed by ChatGPT to Different Categories of Responses. The y-axis represents the probability scores (expressed as a percentage) that ChatGPT assigned to its answers, indicative of self-assessed confidence. The categories on the x-axis represent four scenarios: ChatGPT correctly identifying a correct answer (red), ChatGPT incorrectly identifying a correct answer (blue), ChatGPT incorrectly identifying an incorrect answer as correct (green), and ChatGPT correctly identifying an incorrect answer (purple). The central line in each box denotes the median confidence score, while the bounds of the boxes represent the interquartile range (IQR). The median confidence level for correctly identified correct answers and incorrectly identified correct answers was both at 70.0%, while for correct answers misclassified as incorrect, the median was significantly lower at 10.0%. Incorrect answers accurately identified as such had a median confidence of 5.0%. Despite the presence of statistical significance, the minimal practical variance in confidence scores suggests a limitation in ChatGPT's ability to self-evaluate the certainty of its responses accurately.

ference in performance between the MRCOG Part One and Part Two exams highlights several important factors. This discrepancy could be attributed to several factors:

1. Nature of the Questions: Part Two questions require more complex clinical reasoning and decision-making, areas where ChatGPT's limitations in understanding nuanced medical contexts and integrating experiential knowledge become apparent.
2. Linguistic Complexity: Although our analysis showed that linguistic complexity (unique token count and type-token ratio) had a statistically significant but minor impact on performance, the nature of clinical reasoning questions in Part Two may inherently require a deeper level of comprehension and synthesis that goes beyond vocabulary breadth and diversity.
3. Contextual Understanding: ChatGPT, while proficient in processing text, lacks the ability to fully grasp the context and subtleties of clinical scenarios. This limitation affects its performance in questions that demand a holistic understanding of patient care and decision-making processes.
4. Training Data Limitations: The model's training data may not encompass the specific and detailed clinical scenarios represented in Part Two, limiting its ability to accurately predict and reason through these questions.

These factors collectively underscore the current limitations of ChatGPT in medical applications, particularly in complex clinical decision-making tasks. Understanding these limitations is crucial for developing and refining AI models to enhance their reliability and safety in clinical settings.

Our study considered the setup and evaluation of the MRCOG examinations. The dataset consisted of questions sourced from a variety of resources, ensuring they were beyond the training data available to ChatGPT. This approach aimed to mitigate any prior exposure and potential memorisation by the model. By employing a dual-review

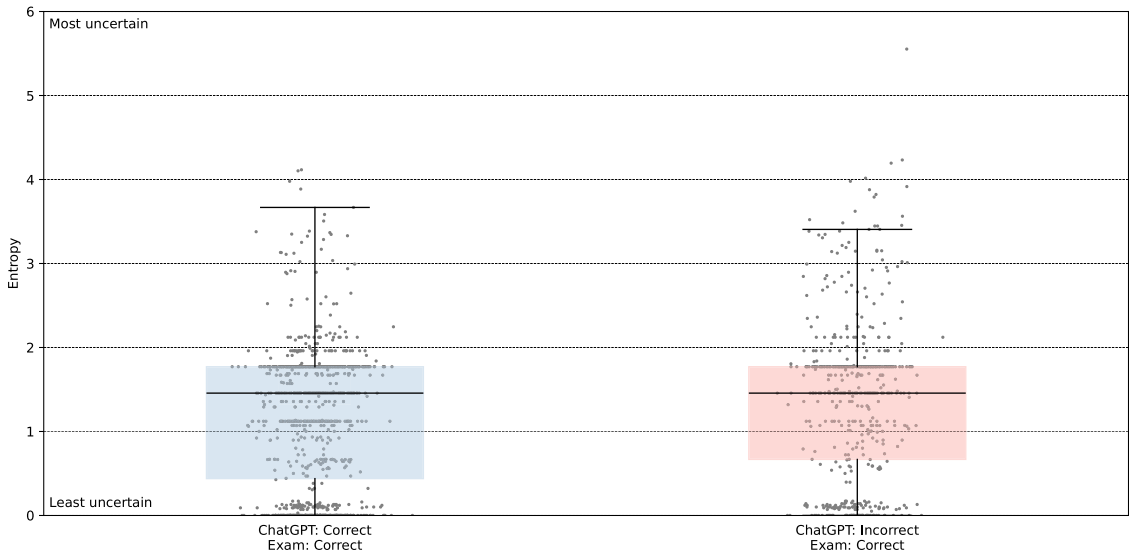

Fig. 4 | Comparative entropy distribution for correct and incorrect responses by ChatGPT. This box-and-whisker plot displays the entropy values for ChatGPT responses, stratified by the model's accuracy and the actual correctness of the exam answers. The y-axis represents entropy, a measure of uncertainty, with higher values indicating greater uncertainty. The blue box represents responses where ChatGPT's answers were correct for questions with correct exam answers, showing a median entropy of 1.46 (IQR: 0.44–1.77). The red box denotes responses where ChatGPT's answers were incorrect for questions with correct exam answers, with an identical median entropy of 1.46 (IQR: 0.67–1.77, $p < 0.001$). The consistent median entropy across both categories indicates that ChatGPT's confidence does not significantly vary between its correct and incorrect responses, despite the statistical significance, calling into question the model's self-assessment accuracy.

process for validation, we ensured both technical and clinical accuracy, which strengthens the reliability of our findings. Extended Matching Questions (EMQs) posed a specific challenge for ChatGPT due to their format, which often includes a larger number of possible answer choices compared to Single Best Answer (SBA) questions. EMQs require more complex reasoning and the ability to integrate multiple pieces of information, highlighting the limitations in the model's clinical reasoning capabilities.

To evaluate the repeatability of our tests, we standardised the interaction with ChatGPT using consistent parameters and prompt structures. This included controlling for variables such as the temperature setting of the model to ensure deterministic outputs, which is critical for replicating results in subsequent studies. While our study did not involve repeated runs of the same set of questions, we believe this best represents how this technology is currently being utilised by clinicians and patients. The detailed setup and standardised procedures provide a framework that can be easily replicated for further research. Future studies could expand on this work by comparing results across multiple iterations and with different versions of ChatGPT, as well as other large language models. This would provide a deeper understanding of the repeatability and robustness of the findings, contributing to the broader field of AI applications in clinical settings.

This study demonstrates several significant limitations not only in ChatGPT's domain knowledge but also the understanding and application of complex clinical knowledge and reasoning. Given this discrepancy in performance between Part One and Part Two examinations, it would be premature to suggest ChatGPT possesses a comparable or useful level of understanding within O&G.

This conclusion is reinforced when considering ChatGPT's overall self-reported confidence and certainty in its answers. It displayed a high degree of confidence in incorrect responses, performing poorly when presented with the correct answer as an option, as evidenced in SBA and EMQ formats. Although statistical significance was observed, the practical implications of this finding remain equivocal, necessitating further investigation to ascertain whether ChatGPT possesses an inherent ability to gauge the veracity of its generated answers with any degree of reliability. This indicates that ChatGPT does not have a reliable mechanism for self-evaluating its confidence or certainty, as evidenced by similar scores for both correct and incorrect responses. This misalignment between confidence and correctness raises concerns regarding the reliability of ChatGPT in clinical decision-making or patient information-giving scenarios.

It was observed that for 192 questions, ChatGPT assigned identical probability scores to each answer option. This lack of discriminatory power suggests either a uniform uncertainty or a failure to properly interpret the question. These instances highlight a significant limitation in ChatGPT's ability to differentiate between multiple answer choices, which is critical in a clinical context where accurate decision-making is paramount. By understanding these instances, we can identify specific areas where the model's performance can be improved, thereby enhancing its reliability and safety in clinical applications.

ChatGPT's capabilities extend beyond medical knowledge to include a wide range of applications such as language translation, content creation, and customer service. However, this study specifically focuses on benchmarking ChatGPT's performance in the domain of obstetrics and gynaecology by using the MRCOG examinations as a standard. These exams are recognised internationally and cover essential knowledge and skills required in O&G, which is a critical component of women's health.

Women's health encompasses a broad range of issues, including reproductive health, maternal health and conditions that disproportionately affect women. While the MRCOG exams primarily focus on O&G, they contribute significantly to the broader field of women's health by ensuring that practitioners are well-versed in the medical and clinical aspects of caring for women during pregnancy, childbirth, and reproductive health.

Given the scope of our study, we have chosen to limit our evaluation to ChatGPT's medical knowledge capacities, specifically within O&G, to provide a clear and focused analysis. Expanding the discussion to include all potential applications of ChatGPT would dilute the relevance and applicability of our findings within the clinical and patient care contexts we aim to address.

There is growing concern globally surrounding AI safety; our findings support this[23]. While LLMs such as ChatGPT undoubtedly possess substantial potential in several domains it has demonstrated significant limitations in medicine and healthcare[24–27]. Impressive performance in one task does not necessarily translate to equivocal performance in others. Users of this technology, both medical practitioners and patients alike, need be aware. As these AI models continue to develop, we hope to see an improvement in women's health. Women's health is a field with a significant

diagnostic and treatment gap[5,6,8,28]. Caution must be taken that, through these technologies, it does not widen. Safety in the context of women's health must be a priority. Work is currently underway to develop and evaluate LLMs trained instead on region-specific clinical best practice guidelines. We are also developing a platform for safely testing LLMs based on local and international clinical consensus. Through this work, we hope to see the development of reliable, robust and safe AI models that can be of utility.

There are several important strengths to this study. We evaluated ChatGPT with data unlikely to have been used in its training. This enabled a more direct and robust interrogation of its aptitude in clinical knowledge and reasoning without the associated bias of testing the AI model on previously learned questions and answers. In essence, we have avoided testing a system on an examination it has already memorised, forcing it instead to use its current domain-specific knowledge and reasoning. We have also evaluated different levels of expected clinical aptitude by examining ChatGPT on Parts One and Two of the MRCOG. Our evaluation encompassed not only the accuracy of responses but the model's linguistic processing capabilities and its self-assessment of confidence and certainty. We have demonstrated that the poor performance of ChatGPT is not attributable to linguistic complexity. Likewise, we have shown that ChatGPT is equally as confident when it is wrong as when it is correct. Currently, ChatGPT will answer most questions, with relative disregard for safety or accuracy beyond a generic disclaimer. This study was limited in that it did not compare ChatGPT's performance directly against the performance of candidates undertaking the same examinations – these data are not provided by the RCOG. We posit, however, that LLMs with the potential demonstrated by ChatGPT need to demonstrate at least near-perfect performance. Especially if they are to be made as publicly available as ChatGPT.

In light of our findings, we suggest that for LLMs to be viable in medical practice, they must first unequivocally demonstrate domain competence in both knowledge and reasoning. Such competence entails not only matching (or surpassing) human experts in clinical knowledge and reasoning tasks, which in itself is insufficient to capture the complexities of clinical medicine, but also possessing an acute awareness of the AI's own boundaries of knowledge and the associated risks when these boundaries are approached or breached.

In conclusion, while ChatGPT's performance is impressive from the perspective of the progression of large language models (LLMs), it is not satisfactory for clinical practice. The model demonstrated commendable accuracy in basic medical knowledge, but its limitations in clinical reasoning and decision-making tasks, coupled with a high degree of confidence in incorrect answers, highlight the need for significant refinement. Therefore, despite its potential, ChatGPT in its current form is not ready for use in clinical settings or for providing medical information in women's health.

Methods

Data acquisition and processing

We extracted single best answer (SBA) and extended matching questions (EMQ) questions for the MRCOG Part One and Part Two examinations from online sources regarded as unavailable to LLMs trained on publicly available data. This was done to reduce the possibility of evaluating ChatGPT on examination questions it had already observed and memorised. Sources included the Royal College of Obstetrics and Gynaecology[29] and publishers making their content only available to users with the appropriate license. Only questions and data published after 2012 were used to ensure that the dataset included only questions from the new examination format introduced by the RCOG in 2012. Data extraction was permitted under the exception of Section 29 of the UK Copyright, Designs and Patents Act 1988 which allows researchers to make copies of copyright works for non-commercial research. Prior to inclusion in the study database, each question and corresponding answers underwent validation. This involved a dual-review system where a data scientist ensured the technical accuracy of the conversion to the study format and the clinical team confirmed the medical accuracy and relevance in line with current guidelines. Questions that relied on information from previous questions were updated to include that information, while those that were duplicated or included images for interpretation were omitted.

Questions were converted into a format for simplified interpretation by ChatGPT. We chose the JavaScript Open Notation (JSON) format for its flexibility and widespread use in data interchange[30]. JSON's hierarchical structure allows for the representation of complex question and answer formats, facilitating the efficient parsing of data by ChatGPT. To ensure tabular data retained its context and was interpretable, we developed a conversion protocol that preserved the relational structure of tables, converting them into nested JSON objects that ChatGPT could systematically evaluate. The background information for the examination (including the type and nature of question being asked, e.g. SBA or EMQ) was incorporated into the instruction given to ChatGPT. The knowledge area and domain of understanding assigned by the publisher for each question was recorded with the question for sub-analysis (e.g. anatomy, biophysics, urogynaecology). Where the subject was not provided by the source, these were assigned by the clinical team.

Interfacing with the OpenAI application programming interface (API)[31] was accomplished using a Python script. We ensured that each query to the API was structured to mirror the interactive nature of the ChatGPT interface, including the provision of context where necessary and the structured format of the JSON-encoded data. We systematically designed the prompts following guidelines and best practices outlined[32,33], ensuring clarity and consistency in the queries provided to ChatGPT. Each prompt underwent multiple iterations to optimise the language and structure, enhancing the model's ability to generate accurate and reliable responses (Supplementary Text 3). Parameters such as temperature, which controls the randomness of the response, were set to zero to favour deterministic outputs, providing consistency across multiple requests. The complete prompt was then provided to ChatGPT and the responses recorded. ChatGPT was presented with each prompt individually to avoid contamination of responses, ensuring that each response was generated based on the input provided without influence from neighbouring questions. ChatGPT was not subsequently informed of the correct answer. The response was then compared against the correct answer for each question.

Linguistic complexity analysis

We then investigated the role of linguistic complexity in model performance. Each question was tokenised and metrics including unique token count and type-token ratio (TTR) were computed[34]. The unique token count represented the total number of distinct words used (the breadth of vocabulary) while the TTR provided a measure of lexical diversity (the diversity of that vocabulary relative to the total number of words used). These were selected as metrics of linguistic complexity because they offer insights into the variety and richness of the language used within the questions. These metrics are indicative of the complexity ChatGPT must navigate to understand and respond to a question, hypothesizing that a higher linguistic complexity might affect ChatGPT's performance.

Self-assessed confidence and uncertainty

Finally, we aimed to determine the extent to which ChatGPT could self-assess the confidence and uncertainty of its responses. We conducted a series of experiments wherein ChatGPT was instructed to assign a probabilistic confidence score, ranging from 0–1 (0–100%), to each answer option within a question. The decision to assess confidence using a probability score is grounded in the probabilistic nature of ChatGPT's language model. These confidence scores were then utilised as an indicator of the model's self-perceived accuracy when the correct answer was identified. The responses deemed incorrect by ChatGPT were bifurcated into two categories: those incorrectly classified as erroneous and those accurately classified as such. A higher confidence score was interpreted as indicative of greater certainty in the response.

Entropy was calculated for the distribution of confidence scores to quantitatively measure the model's uncertainty. Entropy was calculated using the Shannon entropy formula[35], a fundamental concept in information theory that measures the unpredictability or randomness of information

content. In this context, it quantifies the degree of uncertainty in ChatGPT's predictions. Entropy values inversely correlate with uncertainty; thus, lower entropy signifies greater confidence in the responses, and higher entropy indicates greater uncertainty. This analysis provided a statistical layer to the confidence scores, enriching our understanding of the model's performance. A statistically significant difference in confidence or uncertainty levels across different categories would imply an intrinsic capability of ChatGPT to discern the boundaries of its knowledge within specific domains.

Statistical analysis

We adhere to STROBE guidelines where applicable (Supplementary Table 3). Categorical variables are expressed as frequencies and percentages. Continuous variables not normally distributed are described using medians and interquartile ranges (IQR). The accuracy metric was defined as the ratio of correct predictions to total predictions made by ChatGPT, where a correct prediction is denoted as a congruence between ChatGPT's prediction and the true value (e.g., both ChatGPT prediction and correct answer are 'A'). Accuracy and probability values are reported as percentages. Differences between categorical variables were assessed with the Chi-square test. Continuous variables were evaluated using the Mann-Whitney U test, considering a p-value of less than 0.05 as statistically significant. The analysis utilised the GPT-4 model ("gpt-4", accessed 9th November, 2023) to assess responses to queries. All statistical computations were conducted with Python (version 3.9.17), employing libraries including Pandas (version 1.5.3), NumPy (version 1.23.5), Matplotlib (version 3.7.1), OpenAI (version 0.28.1), and TikToken (version 0.5.1).

Data availability

The data underpinning the findings of this study are proprietary and, as such, cannot be openly disclosed. This restriction is due to the data being subject to confidentiality agreements and intellectual property rights held by external partners. The proprietary nature of the data prevents us from providing access to or sharing the raw data sets or the specific data analysis methods that would enable replication of the results. We ensure that all necessary permissions were obtained for the use of this data in the research, and all analyses were conducted in accordance with relevant guidelines and regulations. To maintain the integrity of the research and the confidentiality of the data, we have provided comprehensive descriptions of the methods and analyses in the manuscript, enabling understanding of the methodologies applied and the conclusions drawn.

Code availability

The code used in this study cannot be made publicly available due to intellectual property constraints and licensing agreements with third-party software providers. Additionally, the code includes proprietary algorithms and methods developed specifically for this research, which are protected under the terms of use by the developers. However, we are committed to transparency and reproducibility; thus, we are open to providing detailed descriptions of the methodologies and computational approaches upon request to the corresponding author for verification and collaborative purposes by qualified researchers.

Received: 14 January 2024; Accepted: 24 June 2024;
Published online: 15 July 2024

References

1. OpenAI. ChatGPT. 2023.
2. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
3. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
4. Gronowski, A. M. & Yarbrough, M. L. The Women's Health Diagnostic Gap. *Endocrinology* **159**, 776–778 (2018).
5. Clancy, C. M. & Massion, C. T. American Women's Health Care: A Patchwork Quilt With Gaps. *JAMA* **268**, 1918–1920 (1992).
6. Owens, G. Gender differences in health care expenditures, resource utilization, and quality of care. *J. Managed Care Pharm.* **14**, 2–6 (2008).
7. Shih, T. et al. The Rising Burden of Preeclampsia in the United States Impacts Both Maternal and Child Health. *Am. J. Perinatol.* **33**, 329–338 (2016).
8. Amin, A., Remme, M., Allotey, P. & Askew, I. Gender equality by 2045: reimagining a healthier future for women and girls. *J. Publishing Group*. **373**, n1621 (2021).
9. Kocoń, J. et al. ChatGPT: Jack of all trades, master of none. *Information Fusion*. **99**, 101861 (2023).
10. Li, S. W. et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am. J. Obstetrics Gynecology*. **229**, 172.e1 (2023).
11. Biswas, S. S. Role of chat gpt in public health. *Ann. Biomed. Eng.* **51**, 868–869 (2023).
12. Cascella, M., Montomoli, J., Bellini, V. & Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J. Med. Syst.* **47**, 33 (2023).
13. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol. Sci*. **3**, 100324 (2023).
14. Hu, J.-M., Liu, F.-C., Chu, C.-M. & Chang, Y.-T. Health care trainees' and professionals' perceptions of ChatGPT in improving medical knowledge training: rapid survey study. *J. Med. Internet Res.* **25**, e49385 (2023).
15. Wang, C. et al. Ethical considerations of using ChatGPT in health care. *J. Med. Internet Res.* **25**, e48009 (2023).
16. Temsah, M.-H. et al. Chatgpt and the future of digital health: a study on healthcare workers' perceptions and expectations. *Healthcare* **2023**, 1812 (2023). MDPI.
17. Campbell, A. J. The new MRCOG curriculum. *Obstet., Gynaecol. Reprod. Med.* **30**, 156–158 (2020).
18. Council G. M. Apply for registration. 2023. https://www.gmc-uk.org/registration-and-licensing/join-the-register/registration-applications/application-registration.
19. Royal College of Obstetricians and Gynaecologists U. Training and support in O&G. 2023. https://www.rcog.org.uk/careers-and-training/training/.
20. Studies DoGM. Joint Master of Medicine (Obstetrics & Gynaecology)/ MRCOG Part 3 Clinical Assessment Examination (Hong Kong). 2023. https://medicine.nus.edu.sg/dgms/master-of-medicine/obstetrics-gynaecology/ (accessed 30/11/2023 2023).
21. Royal College of Obstetricians and Gynaecologists U. MRCOG Part 1 exam centres. 2023.
22. Royal College of Obstetricians and Gynaecologists U. MRCOG Part 1 Exam. 2023. https://www.rcog.org.uk/careers-and-training/exams/mrcog-our-specialty-training-exam/mrcog-part-1/.
23. Amodei D., et al. Concrete problems in AI safety. *arXiv preprint arXiv:160606565* 2016.
24. Oztermeli, A. D. & Oztermeli, A. ChatGPT performance in the medical specialty exam: An observational study. *Medicine* **102**, e34673 (2023).
25. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* **2**, e0000198 (2023).
26. Joly-Chevrier, M., Nguyen, A. X.-L., Lesko-Krleza, M. & Lefrançois, P. Performance of ChatGPT on a practice dermatology board certification examination. *J. Cutan. Med. Surg.* **27**, 407–409 (2023).
27. Giannos, P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *J Neurology Open* **5**, e000451 (2023).
28. Winchester N. Women's health outcomes: Is there a gender gap? In Focus. 01/07/2021 ed: House of Lords Library; 2021.
29. Royal College of Obstetricians and Gynaecologists U. RCOG eLearning. 2023. https://elearning.rcog.org.uk/ (accessed 23/11/2023 2023).

30. Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M. & Vrgoč, D. Foundations of JSON schema. *Proc. 25th Int. Conf. World Wide Web* **2016**, 263–273 (2016).
31. OpenAI. OpenAI API. 2023.
32. White J., et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:230211382* 2023.
33. OpenAI. Prompt Engineering. 2023. https://platform.openai.com/docs/guides/prompt-engineering (accessed 14/10/2023 2023).
34. Herdan, G. Type-token mathematics: A textbook of mathematical linguistics. The Hague: Mouton & Co (1960).
35. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

Acknowledgements

This work was supported by the Medical Research Council (MRC) under the UK Research and Innovation (UKRI) grant MR/X029689/1, and by The Alan Turing Institute's Enrichment Scheme.

Author contributions

M.B. was responsible for the data curation, and formal analysis of the study. MB also took the lead in writing the original draft of the manuscript. I.D. and E.M. contributed to the data curation. W.C. and G.D.J. were involved in the conceptualisation, formal analysis, and development of the methodology. G.D.J. additionally provided resources and software, supervised the project, and contributed to writing the original draft. Both W.C. and M.V. were engaged in reviewing and editing 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/s44294-024-00028-w.

Correspondence and requests for materials should be addressed to Gabriel Davis Jones.