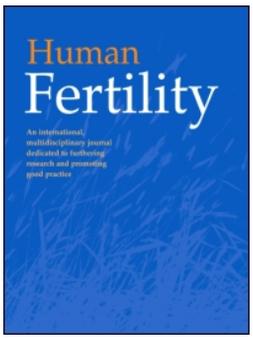

Taylor & Francis
Taylor & Francis Group

Human Fertility

an international, multidisciplinary journal dedicated to furthering research and promoting good practice

ISSN: 1464-7273 (Print) 1742-8149 (Online) Journal homepage: www.tandfonline.com/journals/ihuf20

Ctrl + Alt + Conceive: fertility awareness in the age of Artificial Intelligence, how do large language models compare?

Bola Grace, Jiarui Zhu , Hema Dudakia , Favour Ajao-Rotimi & Nora Colton

To cite this article: Bola Grace, Jiarui Zhu , Hema Dudakia , Favour Ajao-Rotimi & Nora Colton (2025) Ctrl + Alt + Conceive: fertility awareness in the age of Artificial Intelligence, how do large language models compare?, Human Fertility, 28:1, 2584673, DOI: 10.1080/14647273.2025.2584673

To link to this article: https://doi.org/10.1080/14647273.2025.2584673

HUMAN FERTILITY
2025, VOL. 28, NO. 1, 2584673
https://doi.org/10.1080/14647273.2025.2584673

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Ctrl + Alt + Conceive: fertility awareness in the age of Artificial Intelligence, how do large language models compare?

Bola Grace[a,b,c], Jiarui Zhu[b], Hema Dudakia[a], Favour Ajao-Rotimi[c] and Nora Colton[b]

[a]UCL Institute for Women's Health, Faculty of Population Health Sciences, University College London, London, UK; [b]UCL Global Business School for Health, Faculty of Population Health Sciences, University College London, London, UK; [c]Research Department, Cambridge Matrix, Cambridge, UK

ABSTRACT

Technology continues to change how we manage our health, and recent breakthroughs in Artificial Intelligence have increased the adoption of Large Language Models (LLMs) in healthcare. Since the launch of ChatGPT, LLMs have been increasingly used for health information; this study, therefore, aimed to qualitatively assess fertility information provided by LLMs. Content generated by four LLM platforms: ChatGPT, Gemini, Copilot, Perplexity, were analysed comparatively. Thirty-seven prompts were generated, covering five topics: menstrual cycle, conception, risk factors, assisted reproductive technologies and age-related fertility decline. Prompts were analysed for concordance, comprehensibility and conciseness. Safety warnings for all platforms were recorded. LLM platforms generally provided concordant answers for menstrual cycle, conception, and risk factors. However, content on assisted reproductive technologies was the least accurate. Perplexity provided the highest number of strongly-concordant and poorly-concordant responses. Comprehensibility was similar across platforms. ChatGPT was the most concise. Not all platforms provided warning or safety messages regarding potential inaccuracies. LLMs present an opportunity to expand access to fertility and reproductive health information not only for individuals and patients, but also for clinicians, researchers, educators, charities, reproductive health organisations and policymakers. Nevertheless, attention must be paid to the quality of information generated in order to ensure that professionals have accurate guidance, and that individuals can access quality information to help achieve their desired fertility and reproductive health intentions.

ARTICLE HISTORY
Received 22 November 2024
Accepted 21 October 2025

KEYWORDS
Fertility knowledge; fertility awareness; Artificial Intelligence; large language models; reproductive health education

Introduction

Generative Artificial Intelligence (GenAI) refers to AI technologies which are able to generate new content, such as text, images, or other media, based on the data they have been trained on (Jackson & Pinto, 2024). The rise of GenAI in healthcare is transforming the way that medical data is generated, processed and used, and the technology holds high potential for the advancement of health information (Raza et al., 2024). Large language models (LLMs), an increasingly powerful form of Gen AI, are capable of analysing and generating human-like text based on the data on which they have been trained (Reddy, 2024).

LLMs can analyse vast amounts of health literature, synthesise findings and provide summaries in a format that is easy for non-experts to understand (Clusmann et al., 2023; Raza et al., 2024). As these models, which are fed with an extremely large dataset, are increasingly used across population groups, they are becoming almost as good as humans in answering medical questions (Gilson et al., 2023). These AI technologies therefore offer a significant potential for healthcare information to be more

CONTACT Bola Grace bola.grace@ucl.ac.uk UCL Institute for Women's Health, Faculty of Population Health Sciences, University College London, London WC1E 6AU, UK.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/14647273.2025.2584673.

adaptable (Tam et al., 2024), understandable and could provide access to encyclopaedic information on healthcare (Lu et al., 2024), including fertility, enabling proactive and tailored patient information (Grace, 2025).

As individuals continue to delay childbearing for various justifiable reasons (Mertes et al., 2023), the average age of first-time parents continues to increase, especially in high-income countries. According to the UK Office for National Statistics (ONS, 2024), total fertility rate - the rate needed to maintain a population's size - in England and Wales has fallen to a record low of 1.44 per woman. These data highlight the largest decline since records began and there is a similar trend in other high-income countries (OECD, 2024). Due to these demographic shifts in recent years, several stakeholder groups have been driving initiatives to improve fertility education (Beilby & Hammarberg, 2024; Cheshire et al., 2024; Ekstrand Ragnar et al., 2025).

Several research studies have shown that fertility awareness is low across many population groups (Grace et al., 2023a; Hammarberg et al., 2017; Kudesia et al., 2017). Good knowledge and awareness of fertility and reproductive health has the potential to minimise the impact of infertility from modifiable factors, highlighting the importance of good fertility education across reproductive age (Martins et al., 2024). Additionally, fertility education can help improve reproductive autonomy, body literacy and the ability to make good health decisions beyond child-bearing intentions.

Improving fertility education requires a good understanding of the sources of reproductive health and fertility information. Online digital sources are often cited as popular sources of fertility information across reproductive age (Grace et al., 2023b). Reasons for this include access, perceived anonymity, and the speed of obtaining answers. Google is already widely used for seeking medical information (Schneider-Kamp & Kristensen, 2019).

In terms of other digital sources, since the public launch of Chat-GPT, LLMs are increasingly being used to seek health information (Kung et al., 2023; Lautrup et al., 2023; Lu et al., 2024; Sarraju et al., 2023), highlighting the usefulness of LLMS in improving health education. However, research (Farquhar et al., 2024) has highlighted risks of inaccuracies, including hallucination, generation false outputs, and unsubstantiated answers associated with AI-generated information. This study aimed to comparatively assess fertility and reproductive health information generated via LLM platforms via qualitative content analysis to explore their potential as useful tools in reproductive health education.

Materials and methods

Two questionnaires on fertility knowledge (Grace et al., 2023a; Kudesia et al., 2017), covering a range of topics on fertility and reproductive health, were used to generate 37 questions (See Supplementary material 1). The two fertility knowledge questionnaires, published within the last 7 years, were selected due to the breadth of topics covered which included: Menstrual cycle (5 questions), Conception (7 questions), Assisted reproductive technologies (ARTs) (8 questions), Age-related fertility decline (8 questions), and Risk factors associated with fertility (9 questions).

Four LLM platforms: ChatGPT 4.0 Free; Copilot Free; Gemini 1.0 Free (formerly known as Bard); and Perplexity Free, were selected for analysis based on popularity and access (Google, 2024; Microsoft, 2024; OpenAI, 2024; Perplexity AI, 2023). A comparative analysis of the responses from these LLM platforms was conducted. Only free versions of LLMs were used, as a wider population of individuals are more likely to access these free versions than paid versions. Safety features or warning messages provided by LLMs were recorded.

To convert the questions from the questionnaires to prompts, text summarisation was conducted, and multiple-choice questions were converted to closed questions. To minimise the impact of previous responses affecting subsequent ones, each prompt was inputted independently for each LLM platform. Screen captures of LLM responses were saved to provide a snapshot of information at the time of research and for independent review. Text for each prompt was then copied into a Microsoft Excel file for assessment and scoring. Four researchers (HD, JZ, FA, and BG) collected data provided by LLM platforms over a four-month period between June and September 2024.

To allow for a comprehensive qualitative assessment of results provided by LLMs, a scoring system was developed based on the criteria for evaluating LLMs by Lautrup et al. (2023) and Van Veen et al.

(2024). Lautrup et al. (2023) applied adaptation methods to eight LLMs, across four clinical tasks and provided a methodology for assessment via evaluation of results for completeness, correctness and conciseness. Additionally, Van Veen et al. (2024) qualitatively assessed the capability of ChatGPT on 123 prompts and developed a model for LLM evaluation, including whether they are concise, comprehensive and comprehensible. In summary, in the present study, each response was assessed on Likert scale of 1–5 across three key measures: concordance, comprehensibility and conciseness.

'Concordance' is scored based on correct responses against the standard questionnaire answers, where a score of 5 indicates high concordance and a score of 1 indicates low concordance. For 'Comprehensibility' 5 indicates clear and easy to understand responses; and 1 indicates difficulty in understanding responses. Similarly, for 'Conciseness' 5 indicates short content with relevant information, and 1 indicates a verbose response with less relevant information. For concordance, in addition to standard questionnaire answers, where applicable, the accuracy of answers which could be subject to change over time (e.g. ART costs), was assessed using recent information from evidence-based sources: the UK Human Fertilisation and Embryology Authority and the American Society for Reproductive Medicine.

Following initial assessment by the first reviewer, all scores were independently checked by an additional two reviewers, to ensure consistency, and any discrepancies were resolved by discussion between researchers. Comparative analysis of all five sources of information and data visualisation was carried out using Tableau Data Analysis Software, V. 2024.2. Ethical considerations are critical in research. Although this research falls outside the scope of studies requiring University College London Institutional Review Board Ethics Approval, as it did not include human participants, ethical considerations of the use of LLMs were taken into account by authors (Ong et al., 2024).

Results

The comparative analysis of the LLM platforms across measures of concordance, comprehensibility and conciseness for different fertility and reproductive health topics is summarised in this section.

How do LLM platforms compare with each other?

Across measures, LLM platforms performed better for comprehensibility than conciseness or concordance. ChatGPT and Copilot returned similar concordance for fertility and reproductive health information, which was expected, given both use the OpenAI GPT-4 LLM. Both achieved slightly higher concordance than Gemini and Perplexity, although the latter has the highest incidence of high-scoring concordance (5) and low-scoring concordance (1). Concordance depended on the questions asked. All platforms Scored 5 on 50% of responses, but also provided inaccurate responses to different extent, dependent on topic. In terms of conciseness, Perplexity tended to be more verbose while ChatGPT was more concise. Comprehensibility was generally high for all platforms; all scored very similarly, as summarised in Figure 1.

How do LLM platforms compare across different fertility and reproductive health topics?

Analysis based on fertility and reproductive health topics show that all platforms returned generally concordant information on menstrual cycle and conception (Figure 2). LLMs also returned concordant information on risk factors. Information on assisted reproductive technologies was low across all LLM platforms. All platforms were highly comprehensible and generally concise.

To highlight high concordance for age-related fertility decline, when prompted for: 'around what age does the female fertility start to decline?' to which the questionnaire response is 'around 30–35 years', the LLM platforms provided the following responses:

- ChatGPT: 'Female fertility starts to decline around the age of 30, with a more significant decrease after age 35'.
- Gemini: 'Female fertility starts to decline gradually in the early 30s, and more rapidly after 35'.

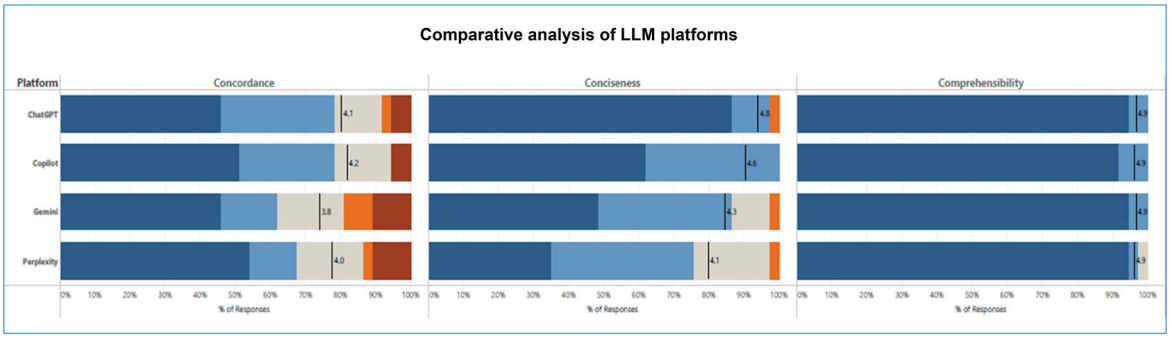

Figure 1. Comparative analysis of LLM platforms.
The colour gradient denotes high versus low scores. The stronger the shade of blue, the higher the score, and the stronger the shade of red, the lower the score with grey indicating the neutral mid-point (3). The black lines denote the average score for each measure.

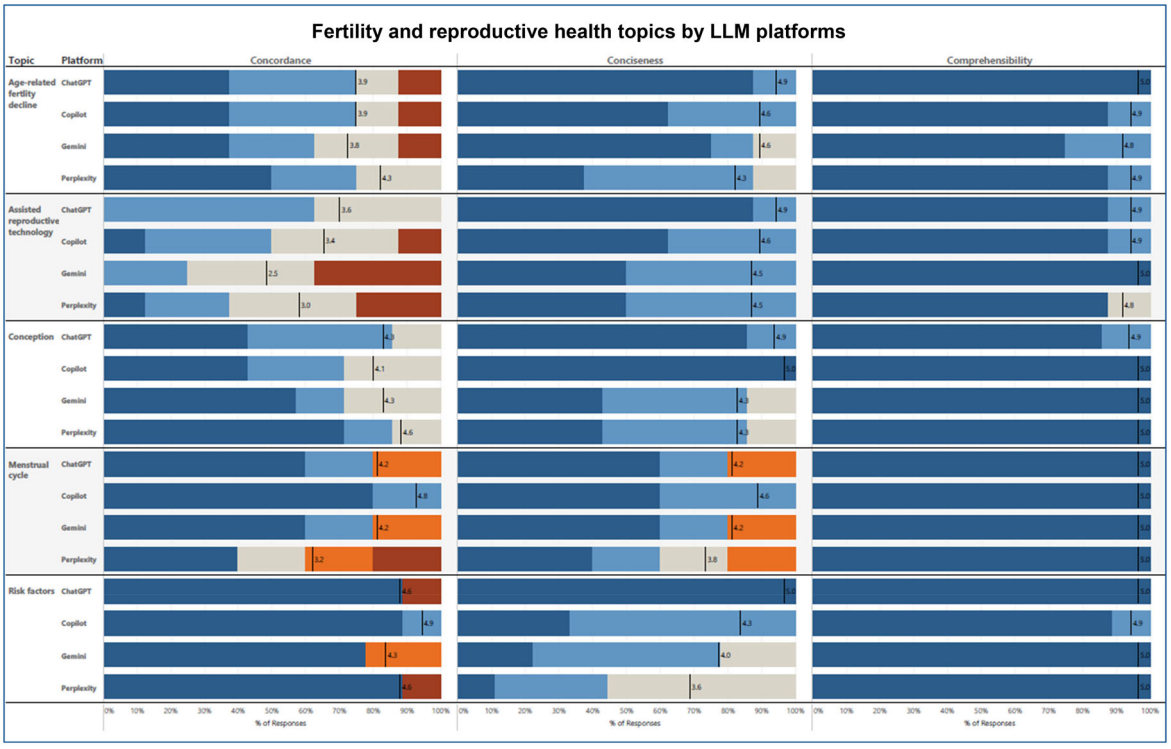

Figure 2. Fertility and reproductive health topics assessment by LLM platforms.

- CoPilot: 'Female fertility generally begins to decline at around age 32 and then drops off more dramatically after age 37'.
- Perplexity: 'Female fertility typically starts to decline in the late 20s, with a more significant decrease after age 35'.

In another example of high concordance, when prompted for: 'Over the course of 1 month, what is the percent chance that a 30-year-old woman who is trying to get pregnant will get pregnant?' to which the questionnaire response is '20%', the following responses were generated:

- ChatGPT: 'The percent chance that a 30-year-old woman who is trying to get pregnant will conceive in one month is about 20%'.
- Gemini: 'Around 20% chance'.

- CoPilot: 'A healthy, fertile 30-year-old woman has a 20% chance of getting pregnant each month'.
- Perplexity: 'A healthy 30-year-old woman has about a 20% chance of getting pregnant each month'.

These responses scored high for concordance, conciseness and comprehensibility.

To highlight varied concordance for ARTs, when prompted for: 'When using frozen eggs from women less than 37 years old, what is the live birth rate per thawed egg?' to which the questionnaire response is '≤10%', the LLM platforms generated the following responses:

- ChatGPT: 'When using frozen eggs from women less than 37 years old, the live birth rate per thawed egg is approximately 2% to 12%'.
- Gemini: 'The live birth rate per thawed egg using frozen eggs from women under 37 is around 60-70%'.
- CoPilot: 'The live birth rate per thawed egg from women less than 37 years old varies, but one study suggests it could be around 70%'.
- Perplexity: 'When using frozen eggs from women less than 37 years old, the live birth rate per thawed egg is approximately 70% when at least 20 mature eggs are thawed'.

While ChatGPT had high concordance, the others had low concordance, especially when safety features are considered. At the time of the study, no caveats, warnings or safety features were included for Co Pilot and Gemini, but ChatGPT and Gemina provided safety features.

In another example of varied concordance, when prompted for: 'Can having less than 9 periods in a year be normal for some women and not require any further evaluation?' to which the questionnaire response is 'No', the following responses were generated:

- ChatGPT: 'No, having less than 9 periods a year is generally not considered normal and typically requires further evaluation to determine the underlying cause'.
- Gemini: 'No. Having less than 9 periods in a year (fewer than one every 42 days) is typically not normal and may warrant evaluation by a healthcare professional'.
- CoPilot: 'While menstrual cycles can vary, having less than 9 periods in a year is generally not considered normal and may warrant further medical evaluation'.
- Perplexity: 'Yes, having less than nine periods in a year can be normal for some women, particularly those with oligomenorrhea, and may not require further evaluation unless accompanied by other concerning symptoms'.

While ChatGPT, Gemini and CoPilot had high concordance, Perplexity did not. All platforms were concise and generally comprehensible.

Safety features

LLM safety was considered in terms of warning messages provided by LLM platforms. All safety warning messages were recorded as summarised in Table 1. At the time of the study, where provided, full text of safety features is outlined in Supplementary material 2.

Discussion

This study comparatively analysed fertility and reproductive health information obtained via LLM platforms, with insights into AI-based outputs. It provided qualitative assessment of the information across three key measures - concordance, comprehensibility, and conciseness.

Concordance, comprehensibility and conciseness across platforms

Overall, LLM platforms generally returned correct answers to fertility and reproductive health questions around menstrual cycle, conception, and risk factors; however, content on assisted reproductive

Table 1. Summary of safety features.

Platform	Safety feature summary
ChatGPT	Upon creating an account, ChatGPT advises against sharing sensitive information and warns users of potential inaccuracies. A warning is also displayed under the prompt input that warns of potential inaccuracies
Gemini	Before inputting a prompt, a message is displayed warning that saved chats may be reviewed for improvement purposes. Similar to ChatGPT a warning is displayed below the prompt input that warns of potential inaccuracies. In addition, responses can be cross referenced with Google search and any discrepancies are highlighted
Copilot	No warnings or safety features displayed
Perplexity	No warnings or safety features displayed however, upon creating an account a message is displayed highlighting that Perplexity is accurate and includes citations
Google	No warnings or safety features displayed

technologies was the least accurate. These findings are similar to other studies, where ChatGPT tended to provide accurate information on health topics (Shieh et al., 2024; Torun et al., 2024; Walker et al., 2023) including fertility and reproductive health (Beilby & Hammarberg, 2024; Chervenak et al., 2023). LLMs use a generative approach by synthesising a new set of information from large datasets, rather than surfacing individual, human-generated articles (Clusmann et al., 2023).

From a technological perspective, it has been reported that LLMs are still in their infancy. However, the pace of their evolution has been fast (Chiarello et al., 2024). LLM platforms have rapidly developed over the past couple of years and are showing a high potential to transform access to medical information (Raza et al., 2024). The quality of machine-generated content obtained via LLM platforms is improving rapidly. Within a short time, LLM platforms are proving to be more comprehensive sources of reliable health information (Waldock et al., 2024) than search engine outputs.

ChatGPT by OpenAI, Copilot by Microsoft and Gemini by Google are the leading LLM platforms. ChatGPT is by far the most popular and well-known GenAI product (Fletcher & Nielsen, 2024). Comparative analysis of the LLM platforms evaluated in this study showed that ChatGPT 4 and Copilot have similar concordance for fertility and reproductive health information. This is unsurprising as both platforms are based on the same Generative Pre-Trained Transformers (GPT-4), whereas Google's Gemini and Perplexity AI's LLM platforms are based on proprietary LLMs. In this study, ChatGPT and Copilot returned slightly more concordant responses than Gemini and Perplexity, even though Perplexity provided the highest number of strongly-concordant and poorly-concordant responses versus other LLM platforms.

Another important measure of health information sources is how easy and understandable the information presented is. We found that, for most prompts, the information provided by LLM platforms was almost indistinguishable from human-generated content, highlighting high comprehensibility. This is corroborated by previous articles which highlight that LLM platforms are able to generate human-like healthcare information (Reddy, 2024). Nevertheless, readability remains crucial from a patient perspective, and it remains important to ensure that responses are tailored to the literacy levels of intended users (Tepe et al., 2024). In terms of conciseness, our findings show that machine-generated text is often more efficient at conveying a message, with less use of redundant words or phrases, than human generated text.

Fertility and reproductive health topics, safety and ethical considerations

In terms of fertility and reproductive health topics, information generated on general reproductive biology such as menstrual cycle and conception was largely accurate. This was similar to information on risk factors associated with fertility. However, generated text about important aspects of ARTs was inaccurate, presenting potential risk and safety issues for fertility patients where LLMs are used as sources of information for patients and healthcare professionals alike. For example, when asked about birth rate per thawed egg when using frozen eggs, some platforms in this study provided inaccurate responses.

One explanation for the difference in concordance on these topics is the large body of source material on basic reproductive biology compared to the much smaller body of source material on ARTs and their fast-changing nature due to the advancement in ARTs through innovation. Another possible explanation is LLM hallucination. This refers to GenAI systems generating inaccurate, fabricated, non-existent responses with a high level of confidence (Farquhar et al., 2024). This adds to the challenges associated with online sources of fertility information.

LLMs can only be as reliable as their training datasets. When models are fed incorrect and/or biased information, their responses will contain the same inaccuracies. Interestingly, Perplexity, which provided a safety message on accuracy based on citations, provided the highest number of strongly concordant as well as the highest number of poorly concordant responses. For some prompts, information generated from the scientific literature referenced by the platform needed to be more representative of the broader population i.e. Perplexity in this study appeared to overweigh data for rare conditions versus those for a normal healthy population. This has implications for training datasets, even when based on published papers, as scientific literature is not foolproof to inaccuracies. Furthermore, the information can change over time.

The rise of social media influencers who are able to rapidly generate fertility content online using LLMs combined with the potential for plausible, yet unverified, inaccurate and/or fabricated information, as well access to a large audience could create a perfect storm for fertility misinformation. Reproductive health organisations need to pay attention to this risk. This becomes more complicated as evidence shows that misinformation spreads up to six times faster online than accurate information (Grace et al., 2025; Menz et al., 2024). Evidence shows that when not carefully monitored, LLM platforms have the potential to drive an infodemic threat in public health (De Angelis et al., 2023).

Security and safety concerns also present risks to the application of LLMs in clinical care (Andrew, 2024). ChatGPT and Gemini provided users with warning of potential inaccuracies; the other platforms evaluated did not provide such caveats. In this study we found evidence of LLM platforms' inaccurate response to success rates of egg-freezing. No safety features were included for two platforms (Co-pilot and Perplexity) at the time of study. It is important for users to understand the limitations of information generated by LLMs. In their article on GenAI safety principles, Obika et al. (2024) highlight the importance of safety considerations. They provide a framework for medical summarisation using generative AI with considerations for intended purpose, identification, evaluation, mitigation of risks, as well as monitoring.

In terms of broader implications, an important consideration is the impact on health equity in fertility. Despite significant potential, use of LLMs in reproductive health, is still relatively low from a global perspective, with use concentrated in high income countries (Fletcher & Nielsen, 2024). To promote health equity, it is important that LLMs are trained on data which are representative of diverse population groups and that they are thoroughly tested for bias. Ethical considerations should always be taken into account when using LLMs, to ensure that the interests of patients are prioritised (Li et al., 2023).

Study strengths and limitations

In terms of study strengths, this study provides a qualitative content analysis of four LLM platforms on fertility and reproductive health information, allowing for direct comparison of AI-generated outputs, based on the same test instrument. The number of LLM platforms reviewed, number of reviewers, and prompts based on fertility questionnaires are key study strengths. A key limitation is the subjectivity inherent with qualitative assessments, even though the authors provided some mitigation by including multiple reviewers.

Another limitation is the generalisability of the study findings. The prompts are based on established questionnaires which have been reviewed by researchers; it is likely that non-experts, in real-world usage, are likely to use simpler terms when prompting LLM platforms and this will generate different responses depending on the sophistication and phrasing of each prompt. Comprehensibility is more than likely to vary among the general public, depending on factors such as educational status or levels of reading comprehension. Additionally, LLMs continue to be updated, and improvements to current versions of the platforms can be made within a relatively short period of time. However, this study

provides a snapshot of information which could be useful for several stakeholders, especially experts, clinicians, academics, educators, researchers, policy makers and reproductive health groups who seek to generate content which can help improve fertility and reproductive health education.

Conclusions

GenAI is transforming how we access information, and the use of LLMs for healthcare information will continue to develop rapidly. As the concerted effort by various groups to improve fertility awareness continues to increase, LLMs present an opportunity to widen the sources of reliable fertility and reproductive health information for individuals, healthcare professionals, educators, researchers, charities, policymakers, reproductive health organisations, and other stakeholders. This has a huge potential as a tool for improving reproductive health education for individuals and notably, for increasing clinician awareness. However, attention must be paid to the quality of information generated, in order to ensure that individuals obtain accurate, reliable and understandable information to help them achieve their desired reproductive health and family building intentions.

Acknowledgements

The authors are grateful to CM for their support with Data Visualisation. All authors contributed to the review and approval of the final manuscript.

Authors' contributions

CRediT: **Bola Grace**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing; **Jiarui Zhu**: Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing; **Hema Dudakia**: Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing; **Nora Colton**: Writing – review & editing.

Disclosure statement

No conflict of interest was reported by the authors.

Funding

No specific funding was used to generate this article.

ORCID

Bola Grace http://orcid.org/0000-0001-5943-1700
Nora Colton http://orcid.org/0000-0001-8679-0098

Data availability statement

Data are underlying this review can be shared on reasonable request made to the corresponding author.

References

Andrew, A. (2024). Potential applications and implications of large language models in primary care. *Family Medicine and Community Health*, *12*(Suppl. 1), e002602. https://doi.org/10.1136/fmch-2023-002602

Beilby, K., & Hammarberg, K. (2024). ChatGPT: A reliable fertility decision-making tool? *Human Reproduction (Oxford, England)*, *39*(3), 443–447. https://doi.org/10.1093/humrep/dead272

Chervenak, J., Lieman, H., Blanco-Breindel, M., & Jindal, S. (2023). The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility*, *120*(3 Pt 2), 575–583. / https://doi.org/10.1016/j.fertnstert.2023.05.151

Cheshire, J., Chu, J., Boivin, J., Dugdale, G., Harper, J., & Balen, A. (2024). The Fertility Education Initiative: Responding to the need for enhanced fertility and reproductive health awareness amongst young people in the United Kingdom. *Human Fertility (Cambridge, England)*, *27*(1), 2417940. https://doi.org/10.1080/14647273.2024.2417940

Chiarello, F., Giordano, V., Spada, I., Barandoni, S., & Fantoni, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, *133*, 103002. https://doi.org/10.1016/j.technovation.2024.103002

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 141. https://doi.org/10.1038/s43856-023-00370-1

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120. https://doi.org/10.3389/fpubh.2023.1166120

Ekstrand Ragnar, M., Hammarberg, K., Carvalho, A., Delbaere, I., Fincham, A., Harper, J., Serdarogullari, M., Simopoulou, M., Antoniadou Stylianou, C., Sylvest, R., & Grace, B. (2025). Defending access to reproductive health information. *Human Reproduction Open*, *2025*(2), hoaf016. https://doi.org/10.1093/hropen/hoaf016

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Fletcher, R., & Nielsen, R. K. (2024). *What does the public in six countries think of generative AI in news?* Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-4zb8-cg87

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*, e45312. https://doi.org/10.2196/45312

Google. (2024). *Gemini 1.0*. Large language model. https://gemini.google.com/

Grace, B. (2025). Accelerating women's health research and innovation. *The Lancet Obstetrics, Gynaecology, & Women's Health*, *1*(1), e11–e12. https://doi.org/10.1016/j.lanogw.2025.100003

Grace, B., Shawe, J., & Stephenson, J. (2023a). Exploring fertility knowledge amongst healthcare professional and lay population groups in the UK: A mixed methods study. *Human Fertility (Cambridge, England)*, *26*(2), 302–311. https://doi.org/10.1080/14647273.2022.2153349

Grace, B., Shawe, J., & Stephenson, J. (2023b). A mixed methods study investigating sources of fertility and reproductive health information in the UK. *Sexual & Reproductive Healthcare*, *36*, 100826. https://doi.org/10.1016/j.srhc.2023.100826

Grace, B., Wise, L. A., Nieroda, M., Egbunike, J., & Usman, N. O. (2025). Digital health technologies to transform women's health innovation and inclusive research. *British Medical Journal*, *391*, e085682. https://doi.org/10.1136/bmj-2025-085682

Hammarberg, K., Collins, V., Holden, C., Young, K., & McLachlan, R. (2017). Men's knowledge, attitudes and behaviours relating to fertility. *Human Reproduction Update*, *23*(4), 458–480. https://doi.org/10.1093/humupd/dmx005

Jackson, J. M., & Pinto, M. D. (2024). Human near the loop: Implications for Artificial Intelligence in healthcare. *Clinical Nursing Research*, *33*(2-3), 135–137. https://doi.org/10.1177/10547738241227699

Kudesia, R., Chernyak, E., & McAvey, B. (2017). Low fertility awareness in United States reproductive-aged women and medical trainees: Creation and validation of the Fertility & Infertility Treatment Knowledge Score (FIT-KS). *Fertility and Sterility*, *108*(4), 711–717. https://doi.org/10.1016/j.fertnstert.2017.07.1158

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lautrup, A. D., Hyrup, T., Schneider-Kamp, A., Dahl, M., Lindholt, J. S., & Schneider-Kamp, P. (2023). Heart-to-heart with ChatGPT: The impact of patients consulting AI for cardiovascular health advice. *Open Heart*, *10*(2), e002455. https://doi.org/10.1136/openhrt-2023-002455

Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, *5*(6), e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3

Lu, Z., Peng, Y., Cohen, T., Ghassemi, M., Weng, C., & Tian, S. (2024). Large language models in biomedicine and health: Current research landscape and future directions. *Journal of the American Medical Informatics Association: JAMIA*, *31*(9), 1801–1811. https://doi.org/10.1093/jamia/ocae202

Martins, M. V., Koert, E., Sylvest, R., Maeda, E., Moura-Ramos, M., Hammarberg, K., & Harper, J. (2024). Fertility education: Recommendations for developing and implementing tools to improve fertility literacy. *Human Reproduction (Oxford, England)*, *39*(2), 293–302. https://doi.org/10.1093/humrep/dead253

Menz, B. D., Kuderer, N. M., Bacchi, S., Modi, N. D., Chin-Yee, B., Hu, T., Rickard, C., Haseloff, M., Vitry, A., McKinnon, R. A., Kichenadasse, G., Rowland, A., Sorich, M. J., & Hopkins, A. M. (2024). Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: Repeated cross sectional analysis. *BMJ (Clinical Research ed.)*, *384*, e078538. https://doi.org/10.1136/bmj-2023-078538

Mertes, H., Harper, J., Boivin, J., Ekstrand Ragnar, M., Grace, B., Moura-Ramos, M., Rautakallio-Hokkanen, S., Simopoulou, M., & Hammarberg, K. (2023). Stimulating fertility awareness: The importance of getting the language right. *Human Reproduction Open*, *2023*(2), hoad009. https://doi.org/10.1093/hropen/hoad009

Microsoft. (2024). *Copilot Free* [Large language model]. https://copilot.microsoft.com/

Obika, D., Kelly, C., Ding, N., Farrance, C., Krause, J., Mittal, P., Cheung, D., Cole-Lewis, H., Elish, M., Karthikesalingam, A., Webster, D., Patel, B., & Howell, M. (2024). Safety principles for medical summarization using generative AI. *Nature Medicine*, *30*(12), 3417–3419. https://doi.org/10.1038/s41591-024-03313-y

OECD. (2024). *Society at a Glance 2024*. https://doi.org/10.1787/918d8db3-en

Ong, J. C. L., Chang, S. Y.-H., William, W., Butte, A. J., Shah, N. H., Chew, L. S. T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., & Ting, D. S. W. (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, *6*(6), e428–e432. https://doi.org/10.1016/S2589-7500(24)00061-X

ONS. (2024). Births in England and Wales: 2023. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2023

OpenAI. (2024). *ChatGPT-4* [Large multimodal model]. https://chat.openai.com/chat

Perplexity AI. (2023). *Perplexity*. Large language model. https://www.perplexity.ai

Raza, M. M., Venkatesh, K. P., & Kvedar, J. C. (2024). Generative AI and large language models in health care: Pathways to implementation. *NPJ Digital Medicine*, *7*(1), 62. https://doi.org/10.1038/s41746-023-00988-4

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science*, *19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

Sarraju, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F., & Laffin, L. (2023). Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*, *329*(10), 842–844. https://doi.org/10.1001/jama.2023.1044

Schneider-Kamp, A., & Kristensen, D. B. (2019). Redistribution of Medical Responsibility in the Network of the Hyperconnected Self. In: Otrel-Cass, K. (Ed.), *Hyperconnectivity and digital reality* (pp 83–102). Springer. https://doi.org/10.1007/978-3-030-24143-8_6

Shieh, A., Tran, B., He, G., Kumar, M., Freed, J. A., & Majety, P. (2024). Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Scientific Reports*, *14*(1), 9330. https://doi.org/10.1038/s41598-024-58760-x

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1), 258. https://doi.org/10.1038/s41746-024-01258-7

Tepe, M., Emekli, E., Tepe, M., & Emekli, E. (2024). Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: A study on readability and accuracy. *Cureus*, *16*(5), e59960. https://doi.org/10.7759/CUREUS.59960

Torun, C. A.-O., Sarmis, A. A.-O., & Oguz, A. A.-O. (2024). Is ChatGPT an accurate and reliable source of information for patients with vaccine and statin hesitancy? *Medeniyet Medical Journal*, *39*(1), 1–7. https://doi.org/10.4274/MMJ.galenos.2024.03154

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. S. (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, *30*(4), 1134–1142. https://doi.org/10.1038/s41591-024-02855-5

Waldock, W. J., Zhang, J., Guni, A., Nabeel, A., Darzi, A., & Ashrafian, H. (2024). The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *26*, e56532. https://doi.org/10.2196/56532

Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, B. P., Raptis, D. A., & Staubli, S. M. (2023). Reliability of medical information provided by ChatGPT: Assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*, *25*, e47479. https://doi.org/10.2196/47479