Golestan University of Medical Sciences
Journal of Clinical and Basic Research
Online ISSN: 2538-3736

The performance of GPT-3.5 and GPT-4 on genetic tests at PhD-level: GPT-4 as a promising tool for genomic medicine and education

Teymoor Khosravi [1], Arian Rahimzadeh [1], Farzaneh Motallebi [1], Fatemeh Vaghefi [1]
Zainab Mohammad Al Sudani [1], Morteza Oladnabi [2,3,4*]

1. Student Research Committee, Golestan University of Medical Sciences, Gorgan, Iran
2. Gorgan Congenital Malformations Research Center, Golestan University of Medical Sciences, Gorgan, Iran
3. Department of Medical Genetics, School of Advanced Technologies in Medicine, Golestan University of Medical Sciences, Gorgan, Iran
4. Ischemic Disorders Research Center, Golestan University of Medical Sciences, Gorgan, Iran

* Correspondence: Morteza Oladnabi. Gorgan Congenital Malformations Research Center, Golestan University of Medical Sciences, Gorgan, Iran. Tel: +9891044958390; Email: oladnabidozin@yahoo.com

Abstract

Background: Natural Language Processing (NLP) has empowered AI models to understand and generate human language, with transformer-based architectures like GPT-3 and GPT-4 marking significant advancements. GPT-4, equipped with a larger parameter count and multimodal capabilities, offers enhanced accuracy and contextual understanding over its predecessor, GPT-3.5. However, challenges such as factual inaccuracies remain. This study aims to evaluate GPT-4's performance on genetics-related tasks, assessing its strengths and limitations compared to GPT-3.5.

Methods: We assessed GPT-4's performance across five key genetic tasks: (1) understanding basic genetic concepts, (2) interpreting family pedigrees, (3) analyzing genetic mutations, (4) solving population genetics problems, and (5) answering medical genetics Ph.D. entrance exam questions. Both open-ended and multiple-choice questions (MCQs) were used, some of which required forced justification to evaluate reasoning. GPT-4's multimodal capabilities were also tested using pedigree images for inheritance pattern analysis.

Results: GPT-4 demonstrated perfect accuracy in Task 1 (Basic genetic concepts) and Task 3 (Genetic mutation interpretation), correctly answering all 10 and 16 questions, respectively. In Task 2 (Pedigree analysis), GPT-4 answered 24 out of 71 questions correctly, with 47 incorrect responses. For Task 4 (Population genetics problems), GPT-4 provided 30 correct answers out of 34. In Task 5, which assessed performance on a Ph.D. entrance exam, GPT-4 correctly answered 58 out of 80 questions. Performance was notably higher for MCQs than for open-ended questions.

Conclusion: GPT-4 substantially improves over GPT-3.5, particularly in understanding genetic concepts and interpreting genetic mutations. Despite these advances, its performance in more complex tasks, such as pedigree analysis, reveals areas that require further refinement. These findings highlight GPT-4's potential in advancing genetic education and research. Future studies should further explore GPT-4's capabilities and address its limitations in tasks that demand higher reasoning and factual accuracy.

Article History

Received: 28 September 2024
Received in revised form: 14 November 2024
Accepted: 23 December 2024
Published online: 30 December 2024
DOI: 10.29252/JCBR.8.4.22

Keywords

Natural language processing
Generative artificial intelligence
Genetics

Article Type: Original Article

OPEN ACCESS
© The author(s)

Highlights

What is current knowledge?

Currently, the field of genetics leverages Natural Language Processing (NLP) techniques and artificial intelligence (AI) models, such as GPT-3 and GPT-4, to process and interpret genetic data. Transformer-based architectures have been at the forefront of these advancements, showcasing remarkable capabilities in tasks such as translation, summarization, and question-answering. While GPT-4 surpasses its predecessor, GPT-3.5, by offering multimodal input compatibility-processing both text and images-challenges persist, including factual inaccuracies and the generation of unreliable content. These models, despite their enhancements, still exhibit limitations in performing more complex tasks, such as accurately interpreting genetic family pedigrees.

What is new here?

This study introduces a novel evaluation of GPT-4's performance in genetic tasks at a PhD-level, comparing it to GPT-3.5. What is new in this research is the assessment of GPT-4's capabilities across five distinct genetic tasks, including understanding basic genetic concepts, interpreting family pedigrees, analyzing genetic mutations, solving population genetics problems, and answering questions from a medical genetics Ph.D. entrance exam. The study also explores GPT-4's multimodal capabilities, such as analyzing pedigree images to identify inheritance patterns. The results reveal significant improvements in GPT-4's performance, particularly in understanding basic genetic concepts and interpreting genetic mutations, although there remains considerable room for improvement in more complex tasks like pedigree analysis.

Introduction

Natural Language Processing (NLP) in artificial intelligence (AI) is used to understand and generate human language (Figure 1). Common NLP techniques include Named Entity Recognition (NER), Sentiment Analysis, Text Summarization, and Machine Translation. These techniques enable AI to process large amounts of language data, leading to more human-like interactions (1). In recent years, the transformer architecture has emerged as the preeminent model in NLP, outshining its predecessors, the recurrent neural networks. This advancement has been facilitated by the integration of attention mechanisms and the optimization of parallel processing (2). Leveraging increased computational resources has enabled the development of models with expansive parameter counts, capable of achieving performances comparable to human levels. Notably, unsupervised pre-training on extensive internet corpora, exemplified by Bidirectional Encoder Representations from Transformers (BERT)-based models, has markedly enhanced model quality (3). The Generative Pre-trained Transformer (GPT), predicated on the transformer architecture, demonstrates proficiency in tasks such as translation, summarization, and question-answering. Successive iterations, specifically GPT-2 and GPT-3, have improved upon transferability and have shown remarkable performance improvements proportional to their augmented parameter sizes (4). InstructGPT refines responses further by employing Reinforcement Learning from Human Feedback (RLHF) during fine-tuning (5). Comparative studies suggest a user preference for InstructGPT over GPT-3, even as standard NLP benchmark datasets exhibit suboptimal performance against user expectations. It is posited that ChatGPT (GPT-3.5), likely by integrating variegated user feedback loops, surpasses GPT-3 in conversational tasks (5). Thus, the collection of human-sourced data for model fine-tuning is deemed critical for optimizing performance and aligning with user expectations (6).

Looking to the future, the evolution of GPT-3.5 into its next iteration, GPT-4, is projected to manifest as a larger-scale model capable of interpreting multimodal inputs including both text and images (7). GPT-4 is poised to set a new benchmark in the realm of NLP-driven chatbots, attributed to its enhanced capabilities. As the most advanced tool in the NLP toolkit to date, extensive empirical studies into GPT-4's proficiency across an array of tasks-ranging from text generation to summarization to translation-are underway (8). It is anticipated that these developments will translate into substantial improvements in the model's overall functionality and its efficacy in executing complex NLP tasks (8).

[DOI: 10.29252/JCBR.8.4.22]

What makes GPT-4 superior to GPT-3.5

The GPT-4 and GPT-3.5 models have garnered considerable recognition for their remarkable capabilities in the field of artificial intelligence. This evaluative comparison seeks to scrutinize the salient characteristics of both GPT-4 and GPT-3.5 in order to discern the enhancements introduced with GPT-4 (9). An examination will be conducted focusing on elements such as the architecture, scale of the model, integration of multimodal inputs, the extent of the context window, the length of text output, methodologies underpinning training, computational speed, precision, rates of factual inaccuracies, and constraints on prompts (9,10). Through this comparative analysis, we intend to illuminate the strides made in the domain of natural language processing and explicate the manner in which GPT-4 furthers the frontier, presenting augmented capabilities for forthcoming innovation within the discipline (11).

Architecture: Both GPT-3.5 and GPT-4 are predicated on the sophisticated Transformer architecture, which employs self-attention mechanisms to discern intricate patterns and relationships within input sequences. Although GPT-4's specific architectural intricacies remain undisclosed, the architecture of GPT-3.5 is publicly recognized, encompassing 175 billion parameters (8).

Model size: GPT-4 represents a significant leap over its predecessor, GPT-3.5, with an estimated 1.76 trillion parameters, a stark contrast to the 175 billion parameters of GPT-3.5. This dramatic upsurge in model size is indicative of enhanced capabilities in processing complex language tasks with greater proficiency (9).

Multimodal inputs: GPT-4 heralds the introduction of multimodal input compatibility, capable of assimilating and interpreting both textual and visual data. This multivalent functionality is a marked advancement from GPT-3.5, which is constrained to processing solely textual input (7).

Context window length: The context window for GPT-4 is considerably broader than that of GPT-3.5, accepting between 8192 and 32768 tokens, substantially more than GPT-3.5's 2048 token capacity. Such an expanded contextual purview grants GPT-4 an elevated degree of precision and contextual relevance in generating responses (12).

Text output length: GPT-4 exhibits the capacity to render text outputs of up to 24,000 words, eclipsing the 3,000-word maximum of GPT-3.5. This capability enables GPT-4 to provide responses that are more elaborate and comprehensive (13).

Training process: The training regimen for GPT-4 fuses RLHF with a Rule-Based Reward Model (RBRM) approach. RBRMs-zero-shot classifiers-supply additional reward signals amid the fine-tuning phase of RLHF. This symbiotic approach seeks to bolster safety and dependability by curbing artifacts like content hallucinations, a step beyond the training methodologies applied to GPT-3.5 (9).

Speed: Speed reflects the model's response time and hinges on factors such as model and input sizes, hardware, and optimization techniques. Typically, smaller models and shorter inputs facilitate quicker response times. GPT-3.5 outpaces GPT-4 due to its comparatively smaller parameter count and context window; however, the speed differential remains negligible for many users (9).

Accuracy: Accuracy evaluates the correctness and pertinence of the model's responses. It is influenced by model size, training data, the nature of the task, and chosen metrics. With more parameters and a more extensive training dataset, GPT-4 outperforms GPT-3.5 in accuracy. It registers fewer factual errors and is more contextually apposite, yielding more dependable and efficacious responses (14).

Factual error rates: This metric assesses the model's propensity to generate mistakes or factual inconsistencies. Fewer factual errors are typically reported in larger models with comprehensive datasets. GPT-4 demonstrates a lower rate of factual inaccuracies compared to GPT-3.5 owing to its vast parameter pool and enlarged dataset, which enhances its factual consistency and verification mechanisms (15).

Prompt restrictions: Prompt restrictions quantify limitations on user requests per unit time, contingent upon model size, computational resources, pricing strategy, and policy framework. Due to their resource-intensive nature, larger models like GPT-4 impose prompt limitations, capping the number of permissible user inquiries per hour. In contrast, GPT-3.5 allows for unlimited user requests within the same time frame (16).

Table 1 provides a comprehensive summary comparing the two versions. Notwithstanding its innovations, GPT-4 shares certain constraints with GPT-3.5, as extensively delineated in OpenAI's "GPT-4 System Card." A cardinal issue is that of 'hallucination', wherein the AI fabricates nonsensical or inaccurately informed content. While GPT-4 evidences a reduction in such occurrences as compared to GPT-3.5, the challenge persists and necessitates ongoing rectification (17). The potential generation of harmful content, including hate speech or incendiary material, is also a significant concern. To this end, two iterations exist: GPT-4 Early and GPT-4 Launch. The latter has integrated safety protocols designed to foster safer outputs, yet safeguards remain less effective under conditions of minimal safety intervention (9). Furthermore, GPT-4 makes strides in contending with disinformation and influence operations. It surpasses GPT-3.5 in curbing the creation of disinformation; still, efforts must continue to thwart the misuse of GPT-4 in fabricating deceptive or manipulative narratives (18). The imperative to enhance the model's resilience against exploitation for the creation of disinformation is vital. The purpose behind OpenAI's exposition is to clarify and deepen understanding regarding GPT-4's competencies, safety-related issues, and the strategies employed to alleviate associated risks (18).

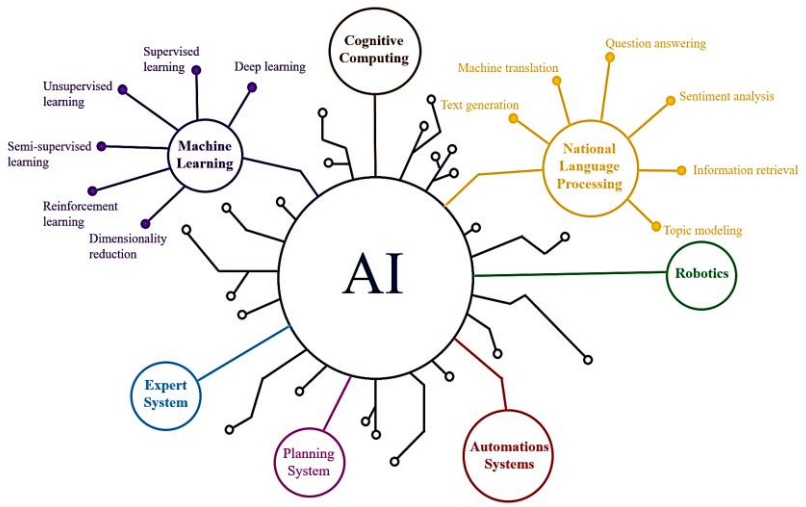

Figure 1. Categorization of artificial intelligence (AI) domains. OpenAI ChatGPT is a natural language processing (NLP) system that is trained on extensive data.

Table 1. Comparison of GPT-3.5 and GPT-4 across various parameters

Aspect	GPT-3.5	GPT-4
Architecture	Transformer	Transformer (Details not disclosed)
Model size	175 billion parameters	Approximately 1.76 trillion parameters
Multimodal inputs	Text only	Text and images
Context window length	2,048 tokens	8,192 to 32,768 tokens
Text output length	Maximum of 3,000 words	Maximum of 24,000 words
Training process	RLHF only	RLHF with RBRMs
Speed	Faster	Slower
Accuracy	Lower	Higher
Factual errors	More	Less
Prompt restrictions	None	Hourly limit

Reinforcement Learning from Human Feedback (RLHF); Reinforcement Learning from Human Feedback with a Rule-Based Reward Model (RBRM)

[DOI: 10.29252/JCBR.8.4.22]

The core of this study pivots on appraising the language comprehension prowess of GPT-4, with a particular focus on its role as a chatbot versed in genetics. The objective is to discern and elucidate the limitations inherent to GPT-4 in addressing genetics-related questions through five distinct tasks. Further, this inquiry aspires to benchmark GPT-4's performance vis-à-vis that of GPT-3.5 in response generation. Herewith, the investigation seeks to unearth insights into GPT-4's strengths and frailties as a linguistic model within the genetic arena and to elucidate how it sets itself apart from its precursor model.

Methods

The purpose of this research is to evaluate the understanding and responsiveness of ChatGPT 4 in the context of genetic questions. In comparison to its predecessor, ChatGPT 3.5, we aim to determine whether ChatGPT 4 has improved its ability to comprehend and respond to genetic queries. Previously, we demonstrated that GPT-3.5 provides correct answers to approximately 70% of the genetics-related questions (19).

One notable advancement in GPT-4 is its enhanced capability to comprehend images. This improvement enables GPT-4 to effectively incorporate familial pedigree images as prompts in Task 2, enhancing its understanding and interpretation of genetic information. In this evaluation, we conducted an analysis of GPT-4's performance by implementing the same tasks used in previous research. These tasks include:

1. Assessing ChatGPT's understanding of basic genetic concepts.
2. Evaluating ChatGPT's interpretation of family pedigrees and identification of inheritance patterns.
3. Analyzing the reliability of ChatGPT's interpretation of genetic mutations.
4. Testing ChatGPT's ability to solve genetic population problems.
5. Assessing ChatGPT's performance in passing a medical genetics Ph.D. entrance exam.

Tasks 1, 2, and 3 comprise open-ended questions, while tasks 4 and 5 consist of multiple-choice questions (MCQs). For the MCQs, we employed forced justification (FJ). This involved engineering prompts that explicitly requested the model to furnish reasoning and persuasive arguments in support of the correct answers. Conversely, there were also MCQ prompts that did not require explicit reasoning. These prompts were designed to assess the model's ability to arrive at the correct answer without explicitly justifying its choice. The workflow of our study encompassed these steps and ensured a comprehensive training regimen for the ChatGPT 4 model (Figure 2).

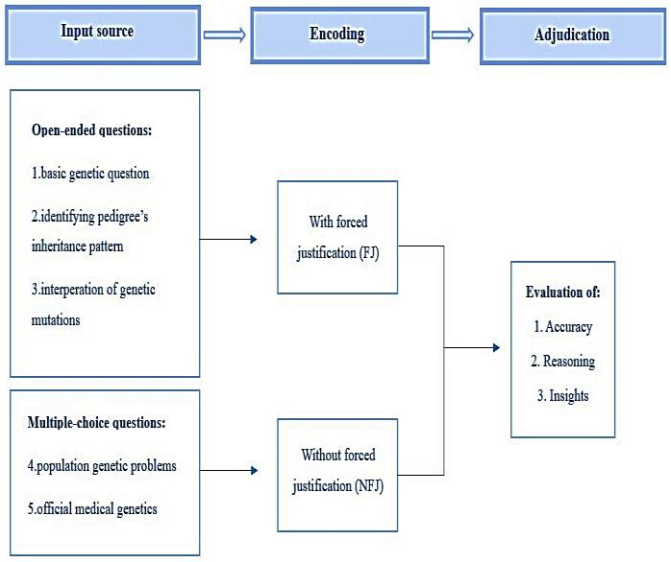

Figure 2. Study workflow illustrating input source, encoding, and adjudication.

Results

GPT-4 performed well on Task 1, which involved basic genetic concepts, by providing complete and comprehensive answers to all ten questions, with no incorrect answers (Supplementary Material 1). For Task 2, which focused on analyzing family pedigrees and identifying inheritance patterns, GPT-4 was able to provide correct answers to 24 out of 71 questions. However, it gave incorrect answers to 47 questions (Supplementary Material 2). In Task 3, which involved notations of genetic mutations, GPT-4 successfully answered all 16 questions correctly, demonstrating a strong understanding of the topic (Supplementary Material 3). For Task 4, which consisted of MCQs about genetic population problems, the chatbot provided correct answers to 30 out of 34 questions. It gave incorrect answers to four questions (Supplementary Material 4). In Task 5, the assessment format in question involved a series of MCQs, which required answering a medical genetics Ph.D. entrance exam, the AI answered 58 out of 80 questions correctly. However, it provided incorrect answers to 19 questions (Supplementary Material 5). Figure 3 illustrates the results obtained, part a show the comparison between MCQs and open-ended questions revealing that GPT-4 displayed a higher accuracy rate in answering MCQs compared to open-ended questions. Part b provides a comprehensive overview of the performance across the five tasks through pie charts. Tasks 1 and 3 achieved a perfect success rate, indicating that GPT-4 accurately understood and responded to basic genetic concepts (Task 1) and the interpretation of genetic mutations (Task 3). However, Tasks 2, 4, and 5 exhibited varying levels of incorrect responses, indicating areas where GPT-4 encountered challenges or made errors.

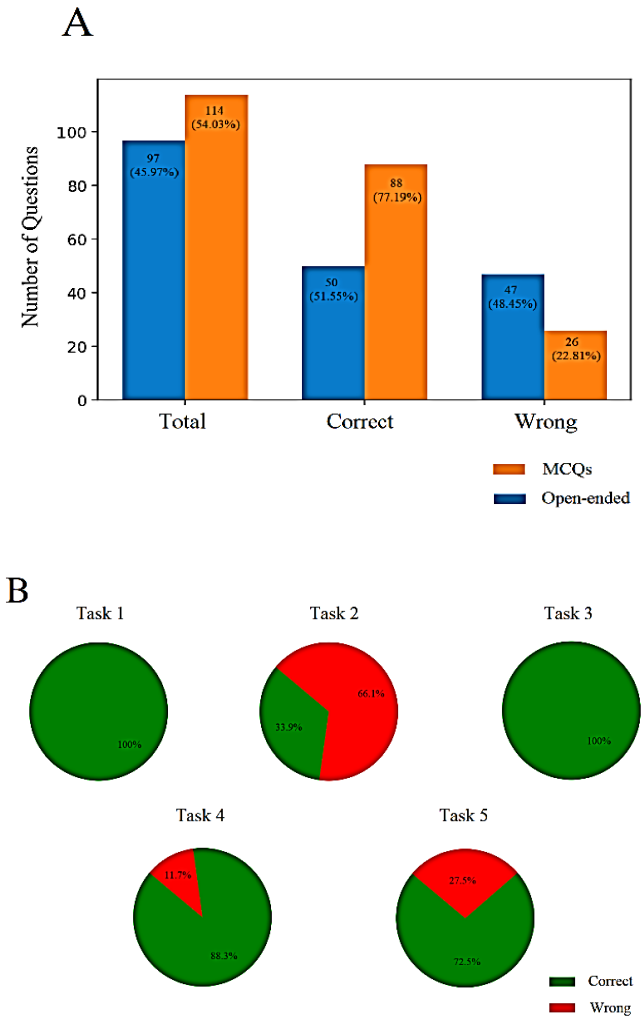

Figure 3. Evaluation of GPT-4's performance in addressing genetics-related questions. The performance of the chatbot in multiple-choice questions (MCQs) and open-ended questions is depicted in (a), along with specific performance metrics for each task in the study shown in (b).

Discussion

GPT-4, with its advanced intelligence capabilities, showcases remarkable proficiency in processing longer prompts and engaging in extended conversations more effectively. It has also exhibited greater factual accuracy compared to GPT-3.5. However, GPT-3.5 outperforms GPT-4 in response generation speed and lacks the hourly prompt restrictions imposed by GPT-4. GPT-4's larger model size allows it to handle complex tasks and generate more accurate responses. The results of our study suggest that GPT-4 has improved performance compared to GPT-3.5 in various genetics-related tasks. As graphically summarized in Figure 4, GPT-4 outperformed GPT-3.5 across all tasks. GPT-4 achieved perfect or near-perfect accuracy in understanding basic genetic concepts (Task 1), interpretation of genetic mutations (Task 3), and solving genetic population problems (Task 4). It also demonstrated better performance in interpreting inheritance patterns of family pedigrees (Task 2) compared to GPT-3.5. Additionally, GPT-4 achieved higher accuracy in the medical genetics Ph.D. entrance exam (Task 5) compared to GPT-3.5. It is also worth noting that GPT-4 answered only 24 out of 71 questions correctly in Task 2, approximately a 33.8% success rate, indicating significant room for improvement in this area. Several studies have shed light on the comparative performance of GPT-3.5 and GPT-4 in different scientific topics. Notably, a study conducted a practicing ophthalmology written examination, where the performance of GPT-3.5, GPT-4, and human users was assessed. Interestingly, GPT-4 and humans surpassed the passing threshold, while GPT-3.5 fell short (20). Furthermore, an evaluation focused on the Turkish Medical Specialization Exam (TUS) questions, comparing the responsiveness of GPT-4

[DOI: 10.29252/JCBR.8.4.22]

and GPT-3.5. The results revealed that GPT-4 exhibited a higher overall success rate compared to GPT-3.5 (14).
Another investigation sought to examine the performance of GPT-4 and GPT-3.5 across three different examinations. The findings demonstrated that GPT-4 consistently outperformed GPT-3.5, achieving a mean.

GPT-4 achieved an accuracy of 80.7% in all versions of the MFE (Medical Field Examinations), while GPT-3.5 achieved a mean accuracy of 56.6% in only two out of three versions (21). The results consistently indicated that GPT-4 outperformed GPT-3.5 in terms of accuracy, particularly in general, clinical, and clinical sentence questions. Additionally, GPT-4 performed better in handling difficult questions and specific disease-related inquiries. With regard to education, the applications of this study are significant. The findings can be applied to enhance genetic education, develop assessment tools, and create virtual teaching assistants. GPT-4's improved performance in understanding basic genetic concepts, interpreting genetic mutations, and answering genetics-related questions accurately can be leveraged to develop interactive and intelligent educational tools. Furthermore, the results of this study can guide the development of assessment systems and practice exams for genetics-related subjects. Additionally, virtual teaching assistants powered by GPT-4 can provide personalized explanations, answer student questions, and engage in interactive discussions, thereby enhancing the learning experience (22). It is important to emphasize that although GPT-4 shows promising performance, ethical considerations should be taken into account when integrating AI models into educational settings. Human supervision and critical evaluation are essential to ensure accurate and responsible use of these technologies. Future directions for this study could involve fine-tuning the models specifically for genetics-related tasks and expanding the range of tasks to cover more complex genetic scenarios (23). Additionally, comparing the performance of GPT-4 and GPT-3.5 with other state-of-the-art language models or specialized genetic models would provide a comprehensive understanding of their relative strengths and weaknesses.

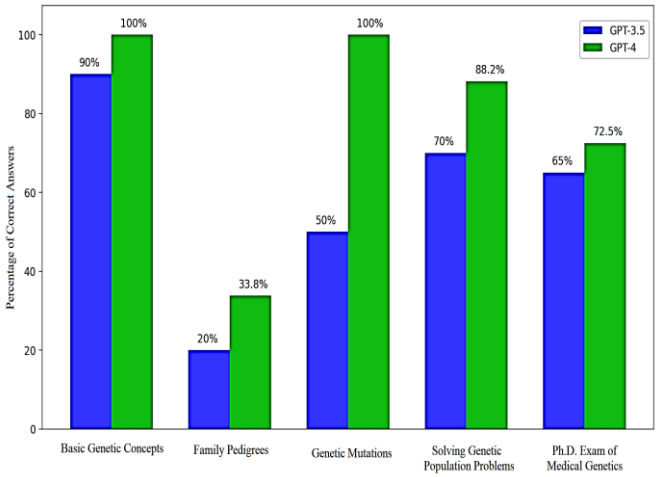

Figure 4. Comparative assessment of GPT-4 and GPT-3.5 models in answering questions related to genetics.

Conclusion

This study demonstrates the superior performance of GPT-4 compared to GPT-3.5 in various genetics-related tasks. These findings, along with other research in different scientific domains, highlight the potential of GPT-4 in advancing scientific understanding and education. Future research can further explore the capabilities of GPT-4 and refine its applications in genetics and beyond.

Acknowledgement

Not applicable.

Funding sources

The present study received financial support from Golestan University of Medical Sciences with grant number 113845.

Ethical statement

Our research was carried out in adherence to the guidelines established by the Ethics Committee of Golestan University of Medical Sciences (Ethics Code: IR.GOUMS.REC.1402.458).

Conflicts of interest

The authors declare that there is no conflict of interest.

Author contributions

Teymoor Khosravi: Contributed to the conceptualization, methodology, investigation, formal analysis, data curation, and writing of the original draft. Arian Rahimzadeh: Participated in the conceptualization, methodology, investigation, formal analysis, and writing of the original draft. Farzaneh Motallebi: Contributed to visualization, methodology, formal analysis, and writing of the original draft. Fatemeh Vaghefi: Focused on visualization, methodology, formal analysis, and writing of the original draft. Zainab M. Al Sudani: Focused on methodology and formal analysis. Morteza Oladnabi: Supervised the project, validated the findings, and contributed to the writing, review, and editing of the manuscript. All authors read and approved the final manuscript.

References

1. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: State of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713-44. [View at Publisher] [DOI] [PMID] [Google Scholar]
2. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Information Fusion. 2023;99:101861. [View at Publisher] [DOI] [Google Scholar]
3. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology. 2023;1(2):100017. [View at Publisher] [DOI] [Google Scholar]
4. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: Past, present and future. AI Open. 2021;2:225-50. [View at Publisher] [DOI] [Google Scholar]
5. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35. 2022:27730-44. [View at Publisher] [Google Scholar]
6. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:220607682. 2022. [View at Publisher] [Google Scholar]
7. Rahaman MS, Ahsan MT, Anjum N, Terano HJR, Rahman MM. From ChatGPT-3 to GPT-4: a significant advancement in ai-driven NLP tools. Journal of Engineering and Emerging Technologies. 2023;2(1):1-11. [View at Publisher] [DOI] [Google Scholar]
8. Chang EY, editor. Examining GPT-4: Capabilities, Implications and Future Directions. The 10th International Conference on Computational Science and Computational Intelligence; 2023. [View at Publisher] [Google Scholar]
9. Koubaa A. GPT-4 vs. GPT-3.5: A concise showdown. 2023. [PPR] [View at Publisher] [DOI] [Google Scholar]
10. Ghosn Y, El Sardouk O, Jabbour Y, Jrad M, Hussein Kamareddine M, Abbas N, et al. ChatGPT 4 Versus ChatGPT 3.5 on The Final FRCR Part A Sample Questions. Assessing Performance and Accuracy of Explanations. medRxiv. 2023. [PPR] [View at Publisher] [DOI] [Google Scholar]
11. Egli A. ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? Clinical Infectious Diseases. 2023;77(9):1322-8. [DOI] [PMID]
12. Espejel JL, Ettifouri EH, Alassan MSY, Chouham EM, Dahhane W. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. Natural Language Processing Journal. 2023;5:100032. [View at Publisher] [DOI] [Google Scholar]
13. Kozachek D, editor. Investigating the Perception of the Future in GPT-3, -3.5 and GPT-4. Proceedings of the 15th Conference on Creativity and Cognition; 2023. [View at Publisher] [DOI] [Google Scholar]
14. Kilic ME. AI in Medical Education: A Comparative Analysis of GPT-4 and GPT-3.5 on Turkish Medical Specialization Exam Performance. medRxiv. 2023. [PPR] [View at Publisher] [DOI] [Google Scholar]
15. Wang W, Shi J, Tu Z, Yuan Y, Huang J-t, Jiao W, et al. The Earth is Flat? Unveiling Factual Errors in Large Language Models. arXiv preprint. 2024. [View at Publisher] [Google Scholar]
16. Adesso G. Towards the ultimate brain: Exploring scientific discovery with ChatGPT AI. AI Magazine. 2023;44(3):328-42. [View at Publisher] [DOI] [Google Scholar]
17. Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints. 2023. [View at Publisher] [DOI] [Google Scholar]
18. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;3:121-154. [View at Publisher] [DOI] [Google Scholar]
19. Khosravi T, Al Sudani ZM, Oladnabi M. To what extent does ChatGPT understand genetics? Innovations in Education and Teaching International. 2024;61(6):1320-1329. [View at Publisher] [DOI] [Google Scholar]
20. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB, 3rd. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JB JS Open Access. 2023;8(3):e23.00056. [View at Publisher] [DOI] [PMID] [Google Scholar]

[DOI: 10.29252/JCBR.8.4.22]

21. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. medRxiv. 2023. [PPR] [View at Publisher] [DOI] [Google Scholar]
22. Cebesoy UB, Oztekin C. Genetics literacy: Insights from science teachers' knowledge, attitude, and teaching perceptions. International Journal of Science and Mathematics Education. 2018;16:1247-68. [View at Publisher] [DOI] [Google Scholar]
23. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1-23. [View at Publisher] [DOI] [Google Scholar]

How to Cite:

Khosravi T, Rahimzadeh A, Motallebi F, Vaghefi F, Al Sudani ZM, Oladnabi M. The performance of GPT-3.5 and GPT-4 on genetic tests at PhD-level: GPT-4 as a promising tool for genomic medicine and education. *JCBR.* 2024;8(4):22-6.

[DOI: 10.29252/JCBR.8.4.22]