Open Access

Journal of Artificial Intelligence & Robotics

Article type: Research Article

Enhancing the accuracy of GPT models in kidney stone diagnosis: A comparison and optimization of GPT 3.5 and GPT 4.0

Haoyang Zeng*

Department of Machine Learning, Shantou University, China.

Corresponding Author: Haoyang Zeng

Department of Machine Learning, Shantou University, China.

Email: 21hyzeng1@stu.edu.cn

Received: Jun 06, 2025

Accepted: Jul 18, 2025

Published Online: Jul 25, 2025

Website: www.joaiar.org

Volume 2 [2025] Issue 2

Abstract

Background: As artificial intelligence advances, Large Language Models (LLMs) have shown tremendous potential in medical diagnosis and treatment, yet existing research has not extensively explored their application in kidney stones. This study evaluated the effectiveness of the most commonly used GPT 3.5 and GPT 4.0 models in addressing clinical questions related to kidney stones and attempted to enhance the performance of existing models through fine-tuning techniques.

Methods: Eighty clinical questions related to renal calculi were proposed by urology experts, and GPT 3.5 and GPT 4.0 were utilized to provide binary (yes/no) answers. Subsequently, the response accuracy of both models was analysed. Finally, we employed variants of these questions and employed fine-tuning techniques to optimize the GPT 3.5 Turbo model, evaluating its training outcomes.

Results: The results revealed that GPT 3.5 and GPT 4.0 achieved accuracy rates of 56.67% and 90.83%, respectively, with GPT 4.0 significantly outperforming GPT 3.5 ($p<0.05$). However, both models exhibited instability in their responses. Finally, through fine-tuning, the accuracy of the GPT-3.5 Turbo model stabilized at 96.25%, surpassing that of the tested model. Additionally, the fine-tuned models demonstrated stability in their responses.

Conclusion: This study demonstrated the advantages and limitations of GPT 3.5 and GPT 4.0 in addressing issues related to kidney stones, highlighting the potential of enhancing AI framework performance through iterative training.

Keywords: LLMs; Kidney stones; GPT 3.5; Fine-tuned model; AI.

Introduction

With the rapid development of Artificial Intelligence (AI) technology and the continuous advancement of medical informatics, Large Language Models (LLMs) have attracted the attention of researchers and have increasingly dominated the field of AI research. LLMs are deep learning models composed of billions to trillions of parameters and are employed in AI systems for processing natural language tasks [1]. These models excel in natural language processing tasks such as text generation, text classification, and question-answering systems. Some studies suggest that LLMs can be applied in the medical field, including image evaluation, diagnostic assistance, and risk prediction [2-4]. However, they still have some limitations; in terms of accuracy, consistency, and transparency, the responses from LLMs are not always satisfactory. Clearly, there is significant room for improvement in the application of existing LLMs in certain specialized medical domains.

Kidney stones, a prevalent urological condition, can develop at multiple sites within the urinary system, such as the renal pelvis, calyces, and ureter [5]. In recent years, with the increase in living standards, the incidence of kidney stones has increased. These stones can lead to a spectrum of symptoms and complications, including pain, haematuria, and hydronephrosis, severely impacting the quality of life of patients. Given the variability in patient physiology, stone composition, and the diverse options for treatment and factors contributing to recurrence, managing individual differences, making treatment decisions, and managing recurrences present significant challenges in clinical practice. Consequently, implementing a specialized clinical decision support system underpinned by AI technology and based on relevant clinical information about kidney stones could greatly enhance the diagnostic capabilities and decision-making proficiency of urological clinicians.

In this study, clinicians, drawing on their experience in treating kidney stones and guided by the *Medical Management of Kidney Stones: AUA Guidelines* [6], formulated 80 clinical questions to evaluate GPT 3.5 and GPT 4.0. Both versions of the GPT were configured to respond with “yes” or “no” answers, with each question repeated three times. Senior urologists assessed the correct answers based on clinical experience and guidelines. Additionally, we applied data augmentation to the original questions to generate variants for building the training set. This involved fine-tuning GPT 3.5 Turbo to optimize the model, ensuring that its responses more closely align with clinical realities.

Materials and methods

Question formulation

The urologist, leveraging both clinical experience and guidelines, composed 80 clinical questions related to kidney stones, covering epidemiology, diagnosis, treatment, prognosis, and follow-up, and provided the correct answers. Another senior urologist then reviewed these kidney stone-related questions and answers for accuracy and relevance.

LLMs Responses

Versions GPT 3.5 and GPT 4.0 were selected as test subjects, with 80 clinical questions related to kidney stones inputted to obtain responses from the two models. The responses were formatted as binary “yes” or “no” to ensure conciseness. To guarantee the reproducibility of the GPT-generated answers, the “new Chat” feature was used for both models; each question was treated as a separate input, and this process was repeated three times.

The refinement of model construction

Fine-tuning involves employing techniques such as data augmentation [7], early stopping strategies, and freezing layers to make slight adjustments or optimizations to a model’s parameters, enhancing its performance on new tasks or datasets. Commonly applied in transfer learning, these techniques aim to boost model performance on specific tasks [8]. In this study, the GPT 3.5 Turbo model underwent fine-tuning for iterative optimization, refining its performance parameters to more accurately address clinical issues related to kidney stones. To achieve this goal, we utilized OpenAI’s Command-Line Interface (CLI) tool to systematically extract text data from a series of test questions, converting these data into the JavaScript Object Notation Lines (JSONL) format. This process was designed to ensure the computational compatibility of the data and facilitate efficient data manipulation. During the fine-tuning phase, we opted for programmatic fine-tuning to ensure flexibility in adjusting parameters or configurations. In subsequent training sessions, we employed data augmentation techniques, using the NLTK library’s WordNet semantic database to create variants of the original questions (five variants per question for the training set) with the original questions serving as the test set. We then continued training the fine-tuned model on incorrect answers generated by the model, using data augmentation to expand the training set samples of incorrectly answered questions before reuploading the data for further iterative optimization of the fine-tuned model until satisfactory results were achieved. Moreover, empirical hyperparameter tuning was applied, setting “n_epochs=10,” “batch_size=8,” and “learning_rate_multiplier=1” to enhance model performance.

Statistical analysis

The accuracy and error rates of GPT 3.5 and GPT 4.0 were separately tabulated, and the correct response frequency for each question by both models was calculated. The fine-tuned model was also included in the comparison of accuracy rates. The data were processed and visualized using GraphPad Prism 9.5. Comparisons between groups were conducted using the Wilcoxon test, with a p value less than 0.05 considered to indicate statistical significance.

The work has been reported in line with the Standards for Quality Improvement Reporting Excellence (SQUIRE) criteria [9].

Results

Based on the clinical experience of senior physicians and guidelines, we established 80 questions concerning kidney stones to test the performance of GPT 3.5 and GPT 4.0, and enhanced the response accuracy of LLMs through training iterations (Figure 1). As depicted in the graphs, we individually assessed the accuracy and error rates for both GPT 3.5 and GPT 4.0 across three responses. The accuracies for GPT 3.5 and GPT 4.0 were 56.67% and 90.83%, respectively, with error rates of 43.33% and 9.17%, respectively (Figure 2). There was a statistically significant difference in the accuracy between GPT 3.5 and GPT 4.0 ($p<0.05$). Furthermore, the accuracy of the fine-tuned model reached 96.25% (Figure 3), which was significantly greater than that of both GPT 3.5 and GPT 4.0 ($p<0.05$). Finally, observation of GPT responses across three tests revealed inconsistencies, indicating instability in responses to the same question across different test sequences. However, this instability was not observed in the fine-tuned model, demonstrating the improvements made through fine-tuning techniques.

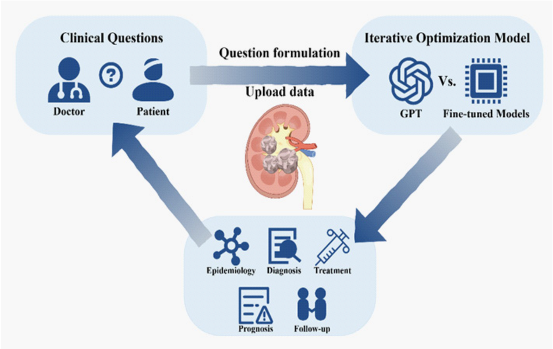

Figure 1: General schematic of this study. Questions are proposed through clinical experience in diagnosing and treating kidney stones. Evaluate different versions of answering ability by asking GPT, and improve the accuracy of LLMs answers through iterative optimization training. (Created with figdraw.com). LLMs Large Language Models.

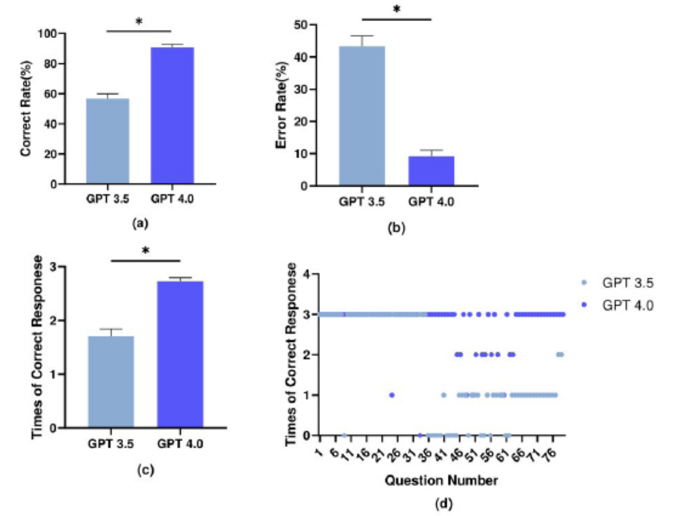

Figure 2: Ranking of statistical measures for assessing the performance of Large Language Models (LLMs) in response to inquiries. This bar graph displays the performance metrics for evaluating LLMs: (a) Average correctness rate of GPT across three test repetitions. (b) Average error rate of GPT across three test repetitions. (c) Average number of correct responses by GPT in three tests. (d) Average number of correct responses by GPT on each tested question.
"*" indicates a p-value<0.05.

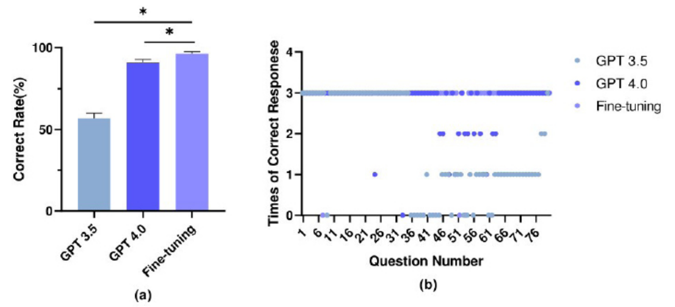

Figure 3: Ranking of statistical measures for assessing the performance of GPT and fine-tuned LLMs in response to inquiries. This bar graph displays the performance metrics for evaluating LLMs: (a) Average correctness rate of GPT across three test repetitions. (b) Average number of correct responses by GPT for each tested question.
"*" indicates a p-value<0.05.

Discussion

In this study, we assessed the accuracy of two versions of the GPT in responding to clinical questions related to kidney stones. By employing training sets constructed with variant questions and refining hyperparameters, the models were iteratively fine-tuned, enabling the fine-tuned model to better understand and respond to complex queries. However, due to limitations in training costs, we were unable to provide a model with more comprehensive clinical information about kidney stones. Therefore, to enhance the model's accuracy when addressing more complex clinical questions about kidney stones, we still need to improve the quality and quantity of the variants used.

Kidney stones are a common clinical condition, with urolithiasis occurring globally and constituting a significant health issue. The incidence of kidney stones is on the rise, and there is a high propensity for recurrence following treatment [10,11]. Patients often suffer from severe lumbar or abdominal pain, frequent urination, and hematuria, which severely impact their quality of life. Additionally, due to the high risk of new and recurring stones, the costs associated with the treatment and management of stone disease are considerable [5]. However, the diverse composition of kidney stones and significant individual variability pose challenges in diagnosis and treatment selection for clinicians [12]. In response, specialized LLMs for kidney stones could serve as valuable clinical decision support tools for diagnosing and treating kidney stones. This model not only provides clinicians with more accurate diagnostic and treatment recommendations but also offers objective, scientific advice during the selection of diagnostic or treatment plans. Furthermore, more personalized diagnostic and treatment strategies could be developed based on individual patient characteristics and conditions, thereby enhancing treatment outcomes and improving patients' quality of life.

LLMs such as the GPT and Codex are expansive neural network architectures based on deep learning technologies that are designed to process and generate natural language texts [13]. LLMs learn the structure and semantics of natural language and the interconnections among various domains of knowledge through extensive training on massive text datasets, exhibiting formidable capabilities in language comprehension and generation, as well as strong transfer learning performance. The GPT 3.5 turbo model used in this study also boasts advantages such as rapid response times, broad application scope, and robust adaptability. In recent years, LLMs have achieved impressive results in the medical field, particularly in providing decision support for medical education and clinical consultations [1,14,15]. However, LLMs are not without drawbacks, as the quantity and quality of datasets can significantly impact model performance. To our knowledge, no studies have yet applied LLMs specifically to kidney stones. Thus, our research aimed to develop an LLM tailored to resolving clinical issues related to kidney stones by inputting relevant clinical information and addressing some of the inherent limitations of LLMs.

Model fine-tuning refers to the technique of adapting a pretrained model for specific tasks or domains by training it further on data related to those specific tasks. The main techniques involved in this study include data augmentation, hyperparameter tuning, and setting binary responses.

Data augmentation, commonly employed in machine learning [16] and deep learning tasks, is a technique used to increase the size and diversity of training datasets. This method involves transforming and expanding the original data to generate new training samples, thereby increasing the diversity and richness of the data. In this study, data augmentation was applied to the original 80 kidney stone-related questions, producing a training set of 400 samples. This enhances the model's generalization ability, reduces overfitting, and effectively utilizes limited data. Additionally, all models in this study were set to use binary answers (yes/no) to simplify the decision-making process. Medical decisions often require rapid judgments and actions, and binary answers simplify the process by reducing complexity and uncertainty. Furthermore, setting binary answers minimizes the likelihood of errors, optimizes solutions, and reduces confusion and misjudgement in practical use. The use of binary answers also provides the most concise output, accelerating the decision-making process and optimizing the user experience. Compared to models that do not use binary answers, decision support models with binary answers facilitate clearer goals, simplifying the design and evaluation process. They also offer faster training speeds and lower training costs [17]. This study conducted preliminary hyperparameter adjustments based on domain knowledge and experience, adjusting the hyperparameters to "n_epochs=10," "batch_size=8," and "learning_rate_multiplier=1".

In this study, we tested the accuracy of GPT3.5 and GPT4.0 in resolving clinical kidney stone-related questions through a query-based approach. Additionally, by adjusting the hyperparameters, the newly generated fine-tuned model achieved an accuracy of 96.25% in responding to kidney stone-related questions, the highest among the tested models. However, we found that GPT3.5 had a significantly lower accuracy than GPT4.0, with an error rate approaching 50% for kidney stone-related questions. Moreover, GPT3.5 exhibited inconsistency in its responses, presenting contradictory answers to the same question. Although GPT4.0 reached an accuracy of 90.83%, it still showed instability. The reasons for this instability might include the following: First, the training data for GPT4.0 were updated as of April 2023, while GPT3.5's data were updated even earlier, causing a lag in the data that could affect the models' judgments. Second, as an LLM, the GPT has inherent limitations, such as poor causal reasoning capabilities, limited vocabulary and semantic understanding, and sample bias, which prevent it from providing accurate feedback in specialized clinical fields [18]. Additionally, the presence of multiple versions of clinical guidelines for kidney stones could lead to inconsistencies in responses. Therefore, by using a training and test set built from guideline-based and clinically experienced kidney stone questions and their variants, this study has somewhat improved the ability of the GPT to respond to clinical questions in the field of kidney stones.

To further enhance the clinical decision-making performance of LLMs in the context of kidney stones, we plan to iteratively optimize the existing models through the following steps in subsequent research. First, we intend to introduce more domain-specific knowledge by incorporating commonly used guidelines for the diagnosis and treatment of kidney stones and integrating the opinions and experiences of seasoned medical experts. This targeted optimization and adjustment of the model aimed to improve its reliability and practicality in real clinical applications [4]. Second, the hyperparameter tuning in this study was based on relevant experience. Future research will include methods such as cross-validation, random search, or Bayesian optimization [17,19], which will help in more accurately adjusting the hyperparameters to achieve better model responses. Additionally, establishing a real-time feedback mechanism will aid in enhancing model performance. By collecting data on the model's performance in actual clinical applications and continually iterating and optimizing based on these insights, the model can adapt to changes and demands in clinical practice, thereby maintaining continuous performance improvement. Finally, we plan to integrate multiple specialized models related to urological diseases, gradually building a more clinically relevant urological clinical decision support system. In addition, the "black box" nature of models [20], poor causal inference capabilities, and ethical issues still limit further performance enhancement [21]. Identifying and addressing these challenges will be crucial for further iterative optimization.

The significance of this study lies in demonstrating that through fine-tuning techniques, LLMs can be optimized to achieve better outcomes in resolving clinical questions related to kidney stones. Furthermore, the clinical adoption threshold for this fine-tuning technique is relatively low, allowing clinicians to generate training sets based on actual clinical practice and further adjust the model to fit the clinical environment. Additionally, characteristics such as low training costs, high accuracy, and good stability also facilitate the application of this model in actual clinical decision-making. Moreover, the model has shown potential in making clinical decisions in complex cases, contributing to personalized and precision medicine.

Conclusions

The significance of this study lies in demonstrating that through fine-tuning techniques, LLMs can be optimized to achieve better outcomes in resolving clinical questions related to kidney stones. Furthermore, the clinical adoption threshold for this fine-tuning technique is relatively low, allowing clinicians to generate training sets based on actual clinical practice and further adjust the model to fit the clinical environment. Additionally, characteristics such as low training costs, high accuracy, and good stability also facilitate the application of this model in actual clinical decision-making. Moreover, the model has shown potential in making clinical decisions in complex cases, contributing to personalized and precision medicine.

Declarations

Ethics approval and consent to participate: This research does not involve human or animal participation, thus ethical approval and participant consent were not necessary.

Consent for publication: All authors have reviewed and consented to the submission of this manuscript.

Availability of data and materials: All data generated and/or analyzed during this study are included in this published article.

Competing interests: The authors declare no competing financial, personal, or professional interests that could influence the conduct or interpretation of this research.

Funding: This study has received funding from the Shenzhen Medical Research Fund (A2301007, A2301008) and the Guangdong Basic and Applied Basic Research Foundation (2019A1515110038).

Authors contributions: Conceptualizing and designing the experiments: Huancheng Yang (HY), HL, HZ. Analysed the data: HZ, XC. Contributed reagents/materials/analysis: JH, Haoyuan Yuan (HY), ZY. Wrote the manuscript: HZ, XC. All authors have read and approved the final manuscript.

Acknowledgment: The author is grateful for the constructive suggestions provided by the editors and reviewers that contributed to the enhancement of this manuscript.

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023; 11: 129–36.

2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Loffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. Commun Med (Lond). 2023; 3: 141.

3. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst. 2024; 48: 22.

4. Javid M, Reddiboina M, Bhandari M. Emergence of artificial generative intelligence and its potential impact on urology. Can J

Urol. 2023; 30: 11588–11598.

5. Khan SR, Pearle MS, Robertson WG, Gambaro G, Canales BK, Doizi S, Traxer O, Tiselius HG. Kidney stones. Nat Rev Dis Primers. 2016; 2: 16008.

6. Pearle MS, Goldfarb DS, Assimos DG, Curhan G, Denu-Ciocca CJ, Matlaga BR, Monga M, Penniston KL, Preminger GM, Turk TM, White JR; American Urological Association. Medical management of kidney stones: AUA guideline. J Urol. 2014; 192: 316–24.

7. Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. J Neurosci Methods. 2020; 346: 108885.

8. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. Diagn Pathol. 2024; 19: 43.

9. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): Revised publication guidelines from a detailed consensus process. BMJ Qual Saf. 2016; 25: 986–92.

10. Siener R. Nutrition and kidney stone disease. Nutrients. 2021; 13.

11. Wang K, Ge J, Han W, Wang D, Zhao Y, Shen Y, Chen J, Chen D, Wu J, Shen N, Zhu S, Xue B, Xu X. Risk factors for kidney stone disease recurrence: A comprehensive meta-analysis. BMC Urol. 2022; 22: 62.

12. Crivelli JJ, Maalouf NM, Paiste HJ, Wood KD, Hughes AE, Oates GR, Assimos DG. Disparities in kidney stone disease: A scoping review. J Urol. 2021; 206: 517–25.

13. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020; 9: 14.

14. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, Bin Saleh K, Alowais SA, Alshaya OA, Rahman I, Al Yami MS, Albekairy AM. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm. 2023; 19: 1236–42.

15. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large language models in medical education: Opportunities, challenges, and future directions. JMIR Med Educ. 2023; 9: e48291.

16. Deo RC. Machine learning in medicine. Circulation. 2015; 132: 1920–30.

17. Dernoncourt F, Nemati S, Kassis EB, Ghassemi MM. Hyperparameter selection. In: Secondary analysis of electronic health records. Cham (CH); 2016. p. 419–27.

18. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, Naugler C. Assessing the research landscape and clinical utility of large language models: A scoping review. BMC Med Inform Decis Mak. 2024; 24: 72.

19. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A guide to cross-validation for artificial intelligence in medical imaging. Radiol Artif Intell. 2023; 5: e220232.

20. Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018; 10.

21. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: Artificial intelligence in healthcare. Cureus. 2023; 15: e43262.