

Investigating the use of a large language model in general practice

WESLEY CHORNEY^{1, A–G}, SING HUI LING^{2, B, D, E}
ORCID ID: 0009-0008-2598-6837

¹ School of Medicine, University College Cork, Cork, Ireland

² School of Medicine, University of Limerick, Limerick, Ireland

A – Study Design, B – Data Collection, C – Statistical Analysis, D – Data Interpretation, E – Manuscript Preparation, F – Literature Search, G – Funds Collection

Summary Background. Large language models have demonstrated strong performance on many tasks. In particular, they have been shown to pass many medical knowledge tests. However, the majority of these studies do not make use of fine tuning.

Objectives. Evaluate the suitability of fine-tuned, large language models in the context of general practice by evaluating performance on the Applied Knowledge Test (AKT) of the Royal College of General Practitioners.

Material and methods. We evaluate the performance of ChatGPT 3.5 in three distinct cases using publicly available practice questions from the Royal College of General Practitioners. In the baseline case, questions are simply input and the answer recorded. In the second case, prompt engineering is used before the questions. Finally, the model is fine-tuned using a subset of the questions and evaluated on the remaining ones.

Results. The fine-tuned model outperforms both the baseline ($p = 0.005$) and prompt engineering cases ($p = 0.010$). Furthermore, the model achieves a passing mark on the AKT, with a mean score of 72.03%.

Conclusions. With further development, fine-tuned, large language models could potentially be used by general practitioners to facilitate areas of their practice. Care must be used to ensure that the models conform to stringent standards to avoid misinforming patients or misguiding care.

Key words: artificial intelligence, general practice, medicine.

Chorney W, Ling SH. Investigating the use of a large language model in general practice. *Fam Med Prim Care Rev* 2025; 27(1): 19–24, doi: <https://doi.org/10.5114/fmpcr.2025.146965>.

Background

Artificial intelligence (AI) makes use of complex models that are able to learn representations of input data in order to process it and produce a desired output. Typically, inputs to models are uniform – they exist in a certain well-defined domain with various restrictions for membership. For example, a model able to recognize carcinoid tumors in the lung might expect input to have a 256×256 resolution and be in black and white. Many AI models have been developed for use in health care, with applications ranging from scheduling [1], histology and molecular profiling [2], to radiology [3]. However, owing to the expectation of structured and well-regulated data, many of these models may overestimate performance [4].

Recent approaches to deep learning allow for models to work with highly unstructured data [5], which widens the scope of use of AI models. In particular, many modern natural language processing models can take user-generated text and output a coherent response (these are typically referred to as large language models, or LLMs). Perhaps the most well-known amongst these models are generative pretrained transformers (GPTs) [6], which have been heavily popularized in the form of models such as ChatGPT. Studies have shown that ChatGPT can pass Step 1 of the United States Medical Licensing Examination (USMLE) [7]. Further studies have shown that ChatGPT performs relatively evenly across all clinical subjects tested on the USMLE exam [8], but how well ChatGPT performs in more advanced contexts is an open question.

Given that general practitioners (GPs) are in short supply [9], a model capable of supporting clinicians could be an asset.

For example, given satisfactory testing, a model could help incoming triage patients, identify those patients who may need a medication review, or provide patients with some information to streamline their consultation. A requisite first step for such a model is possession of specialized and accurate clinical knowledge. Given the digitization of medicine [10], it is important to evaluate how robust large language models are in order to determine in what situations they can be applied. General practice is an ideal scenario to examine this, as it is a varied case mix in terms of presentation, short appointment times, and largely undifferentiated patients.

In the United Kingdom, physicians must complete the Applied Knowledge Test (AKT) of the Royal Society of General Practitioners in order to become a fully qualified general practitioner. The AKT is a more interesting test than the USMLE for two reasons. First, the AKT is designed to test higher order reasoning instead of rote recall [11], whereas Step 1 of the USMLE has been criticized for testing rote recall of clinical facts [12]. Second, as the qualifying exam meant to evaluate whether practitioners have knowledge suitable for independent practice, it is a more specialized and difficult exam.

LLMs have been investigated widely in the medical context. For instance, it has been shown that ChatGPT can pass the USMLE Step 1 examination, the Italian Residency Admission National Exam (IRANE), the German State Examination in Medicine (GSEM) M1 and M2, and the Turkish Medical Specialty Examination (MSE) without the need for prompt engineering or fine tuning [7, 13–15]. Prompt engineering allowed ChatGPT to pass the China National Medical Licensing Examination (CNMLE), Basic Life Support (BLS), and Advanced Cardiovascular Life Support (ACLS) [16, 17]. ChatGPT 3.5 was unable to pass the Japanese



National Medical Licensing Examination (JNMLE), although the authors did not make use of prompt engineering or fine tuning [18]. On the Ophthalmology Knowledge Assessment Program (OKAP) examination and the Neurosurgery Oral Board Preparation Question Bank (NOBPQB), ChatGPT 3.5 answered 57% and 62.4% of questions correctly when tested (both of which are scaled tests without passing scores) [19, 20]. The authors in [11] showed that ChatGPT could not perform at a sufficient level to pass the AKT; however, it is worth noting that the authors used neither prompt engineering nor fine tuning.

Year	Prompt engineering	Fine tuning	Test	Passing mark	Reference
2022	x	x	NOBPQB	~	[20]
2023	x	x	USMLE Step 1	✓	[7]
2023	x	x	AKT	x	[11]
2023	✓	x	CNMLE	✓	[16]
2023	x	x	JNMLE	x	[18]
2023	x	x	IRANE	✓	[13]
2023	x	x	GSEM M1	✓	[14]
2023	x	x	GSEM M2	✓	[14]
2023	✓	x	BLS	✓	[17]
2023	✓	x	ACLS	✓	[17]
2023	x	x	OKAP	~	[19]
2023	x	x	MSE	✓	[15]
2024	✓	✓	AKT	✓	Proposed study

By prompt engineering, we refer specifically to giving context to the model instead of simply giving a question as input. Exams without passing thresholds are marked with ~.

Table 1 provides a concise overview of the above discussion. A probable reason for ChatGPT failing to pass certain examinations is that in the majority of studies, neither prompt engineering nor fine tuning was used. Proper prompts are a knowledge transfer method that provide proper context to ChatGPT in order to enhance its ability to provide a relevant response [21] and improves the performance of ChatGPT in a variety of domains, including data extraction [22] and in the medical domain [23]. Fine tuning allows ChatGPT to be trained further for specific tasks, thereby increasing its performance. For instance, fine tuning has been shown to markedly improve the performance of ChatGPT in domains such as science [24] and education [25, 26].

Objectives

Before models such as ChatGPT can be investigated further in a specific clinical context (such as general practice), it is necessary to have an accurate measure of their strengths and weaknesses so that further developments can be made. In this study, we investigate the ability of ChatGPT 3.5 to provide answers to questions taken from the AKT. In particular, we compare the performance of ChatGPT 3.5 without any modifications versus fine-tuned versions of ChatGPT 3.5. In addition, we show how well-engineered prompts affect the ability of the model to answer these questions. In summary, we

- demonstrate that prompt engineering does not appear to offer any improvement over the baseline performance of ChatGPT,
- show that fine-tuning can markedly improve the performance of ChatGPT 3.5 when answering questions in the Applied Knowledge Test, and

- remark on the clinical significance of the model's results and potential applications.

Material and methods

ChatGPT

ChatGPT 3.5 is a large language model with over 175 billion parameters. It has been trained on an extremely large corpus of text in many languages, totaling over 45 terabytes of data [27]. Owing to its extensive corpus and in-depth training, ChatGPT performs very well across a number of question answering tasks [28].

While ChatGPT 3.5 performs very well in a number of different domains without any additional modification, there are domains in which it struggles. For instance, ChatGPT is known to perform poorly in solving mathematical problems and higher-order reasoning [29], as well as often providing factually inaccurate or biased information [30]. Some techniques exist in order to address these shortcomings. First, prompt engineering can be used in order to provide context to ChatGPT such that it outputs a better response. Well-crafted prompts function as knowledge transfer methods, such that ChatGPT can provide a more relevant response [21]. In addition, a good prompt makes the user's intent manifest in such a manner that ChatGPT is able to provide an answer that aligns with the user's intent [31]. Second, fine tuning can be used to provide additional training samples in order to improve the performance of ChatGPT in a certain domain. Additional training samples allow ChatGPT to learn more suitable responses in a focused domain and, in certain cases, can even improve generalization [32].

We compare the performance of ChatGPT 3.5 in answering questions on the AKT in three different contexts: first, with no additional modifications (*baseline*); second, by including prompt engineering (*PE-ChatGPT*); and third, with fine tuning (*FT-ChatGPT*). In every case, ten distinct, identical trials are carried out in order to generate confidence intervals. Each trial is conducted in a new session in order to avoid previous input to ChatGPT biasing future answers. In the baseline case, questions from the AKT are input without any modification to ChatGPT, and the answer is recorded. In the PE-ChatGPT trials, the following prompt prefaces each question: *Act as a general practitioner from the United Kingdom taking questions from a patient or student and select the most medically accurate answer of the multiple choices following the given question.* In the FT-ChatGPT case, a randomly selected dataset of 41 questions is selected from the 82 AKT questions (those questions involving images or tables are removed) and used to fine-tune the model, after which the remaining 41 questions are used in a manner identical to the trials described above. Furthermore, training hyperparameters are chosen via grid hyperparameter. While this improves model results on the specific dataset, in adapting the model to future datasets, hyperparameter tuning must be repeated to determine optimal hyperparameters for the new dataset. Furthermore, good results on the AKT dataset do not guarantee good results on other datasets – the extent to which finetuning improves model performance on other datasets could be explored in future work.

AKT practice questions

In this study, we use practice questions for the AKT made available online through the Royal College of General Practitioners [33, 34]. These questions are designed to be representative of questions on the AKT, which is a computerized test wherein candidates must answer 200 questions in 190 minutes at a workstation. Importantly, these questions are designed to test higher order reasoning rather than factual recall, so they represent a more difficult task for ChatGPT – as evidenced by its inability to achieve a passing grade when unaided by prompts or

fine tuning [11]. Figure 1 provides an example of the questions that ChatGPT is faced with.

2. Weakness

A 35-year-old woman has noticed increasing weakness and numbness in her legs and arms over the past few days. Today she is feeling short of breath on exertion.

Six weeks ago she had diarrhoea. Stool culture confirmed infection with *Campylobacter jejuni*. She was treated with ciprofloxacin and recovered fully.

Which is the SINGLE MOST likely diagnosis? Select ONE option only.

- A. Drug-induced neuropathy
- B. Guillain-Barré syndrome
- C. Multiple sclerosis
- D. Myasthenia gravis
- E. Transverse myelitis

Answer B. (Guillain-Barré syndrome)

Guillain-Barré syndrome is an acute ascending polyneuropathy which usually occurs following an infection. It can eventually affect the respiratory muscles

Content Guide reference: Neurological

Curriculum: 3.18 Care of people with neurological problems

Figure 1. A sample AKT practice question from [35]

Results

Figure 2 and Table 2 show the overall results. Note that for Figure 2, the y-axis begins at 0.5 for better visibility.

Table 2. Results for each method, reported as accuracy plus or minus standard deviation

Method	Accuracy
Standard	67.68% $\pm$ 2.09%
Prompt engineering	67.80% $\pm$ 2.80%
Fine tuning	72.03% $\pm$ 3.60%

The means are nearly identical between the standard and prompt engineering cases, although the standard deviation is noticeably higher in the prompt engineering case. It is possible that a more detailed prompt could give better context to the model, but we note in particular that the relevant information to frame the task was included in the prompt, including the role, location, and specific task that ChatGPT was asked to perform. The general nature of the questions may be such that a prompt has little effect on the results. That the questions are medically oriented is immediately clear from the text, and because they cover a relatively wide variety of topics, a prompt may not be able to convey the breadth of knowledge.

Most notable is that the performance of the model improves significantly with fine tuning. Figure 3 displays the 95% confidence intervals concerning the means for each method

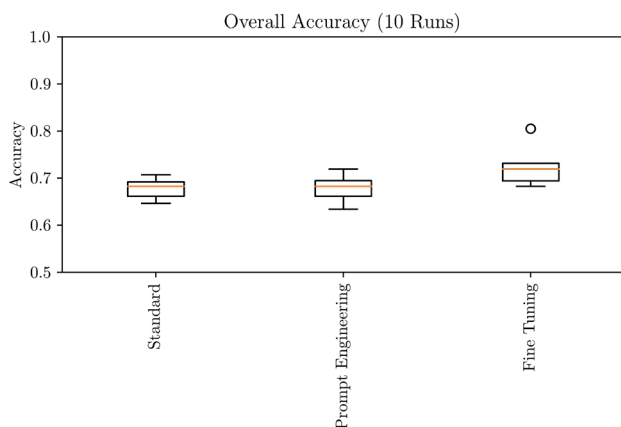


Figure 2. Boxplot of accuracies for the standard, prompt engineering, and fine-tuning cases over ten runs

(again, note the y-axis scaling, which was chosen for visibility). Although there are minimal differences between the standard and prompt engineering cases, the fine-tuning case performs significantly better than the previous two cases. We performed two two-tailed *t*-Tests assuming unequal variances with significance level $\alpha = 0.025$ (we initially selected $\alpha = 0.05$ but applied Bonferroni correction owing to the multiple comparisons) to the fine-tuning versus standard case and the fine-tuning versus prompt engineering case. At the given significance level, fine tuning significantly outperformed both the standard case ($p = 0.005$) and the prompt engineering case ($p = 0.010$). In particular, the mean passing mark for the AKT in 2023 was 69.75% [35, 36] – meaning that the fine-tuned model would be expected to achieve a passing mark.

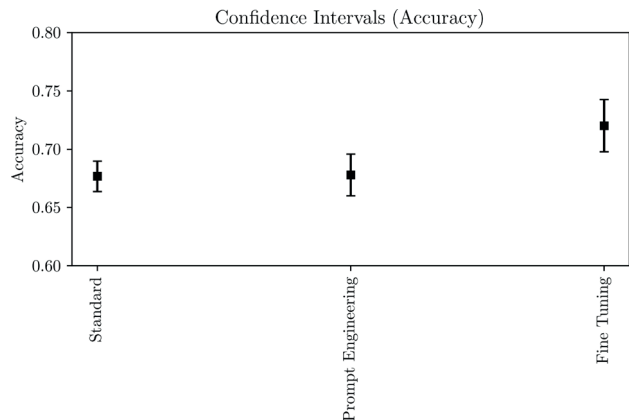


Figure 3. 95% confidence intervals over the ten runs for each method

Discussion

The effect of fine tuning on the performance of ChatGPT on the dataset was significant, although prompt engineering made little difference on the results. Consulting Table 1, we note that ChatGPT 3.5 is able to pass a significant number of examinations without the need for prompt engineering and can achieve a passing mark on additional examinations once prompt engineering is used. The AKT is notable in that ChatGPT was unable to pass the examination with prompt engineering (as we showed). However, none of the previous studies investigated applied a fine-tuned model to medical board exams, and the current study demonstrates that proper fine tuning could lead to notable improvements in performance. In particular, this could be of note for those exams in which ChatGPT did not receive a passing mark. Generally, while prompt engineering ideally contextualizes questions input into ChatGPT, it is limited in that it cannot provide actual samples through which backpropagation can occur. In the context of answering a diverse array of medical questions, the advantage of fine tuning over prompt engineering is clear. The current study demonstrates that the performance is significantly increased by fine tuning, but more investigations could better quantify the precise impact of fine tuning. Therefore, the authors hope that the current study stimulates more research on the performance of fine-tuned models on medical examinations.

A passing score on the AKT is also encouraging from a clinical perspective. We reiterate once again that the value of models such as ChatGPT in the general practice setting is, insofar as the authors are concerned, to accentuate the work of a general practitioner. There are two possible applications: the first is in answering simple patient questions, such that patients have a better idea of what they might expect upon consultation with their general practitioner, which could streamline consultations. However, it is important to note that in automated, unsuper-

vised cases such as this, a model would be limited to answering only those questions in which it is very confident of the answers. Therefore, it would also be necessary to quantify the confidence of each response (for instance, ensuring that the main points in responses to the same question do not differ).

The second application would require less additional work and would involve a general practitioner directly interfacing with a model. The current study demonstrated that a fine-tuned model produced a satisfactory understanding of general practice, given appropriate input. Patients may not be adept at providing appropriate input, but a general practitioner would be able to translate findings into medical jargon, which could then be input into a model. Therefore, the model could be used to produce a rapid second opinion in order to verify a diagnosis or provide differential diagnoses.

Finally, we note that the positive performance of fine-tuned models is interesting from an educational perspective as well. ChatGPT has been noted for its potential to create interactive, engaging learning environments [37], which could complement a medical curriculum. However, it is paramount that these models present accurate information in a rational manner [38]. Otherwise, students could be misled, and false information be perpetuated. Careful fine tuning helps to ensure that the model is presenting accurate information on a specific topic, which would then improve student learning. Students using a fine-tuned model should still use it to accentuate a medical curriculum rather than replace it entirely, lest connections be made inappropriately between topics. Overall, we believe that performant large language models are an exciting addition to the medical field and have the potential to integrate well into many different domains.

While the results are encouraging, there are a number of limitations to point out. The first is that two datasets were used in training the model, designed specifically for the AKT – while this is a test designed to assess knowledge of general practice, it is limited to knowledge only and does not cover the depth and breadth of clinical practice. Furthermore, these tests are specific to the United Kingdom, and therefore the results may not translate to clinical practice in other countries. Additionally, the question format in the AKT is essentially uniform. A short case is presented, and the correct answer (or answers) must be determined from a list of multiple-choice questions. While this is a convenient method for assessing knowledge, even the ability of a large language model to map a clinical preamble to a diagnosis or other piece of relevant information, interfacing with patients, as discussed above, would require providing answers, information, and relevant follow-up questions based on information supplied by patients. Evaluating how successful a model is in tasks such as this would be a source of interesting and important future work.

We also note that ChatGPT and other large language models often make use of reinforcement learning with human feedback (RLHF) [39], which allows the model to continuously improve and adapt its output to the needs of the user. With the goal of integrating a large language model into clinical practice, it would be of great interest to use RLHF. In so doing, the model could improve how it serves the needs of its users. However, the specific manner in which RLHF is used is important. For instance, ChatGPT implements RLHF by asking users which type of answer they prefer. In the medical context, trained physicians should choose which response by the large language model is best in order to ensure that the provided information is medically accurate.

While the present study investigated whether an existing large language model could be fine-tuned in order to demonstrate aptitude on a medical test, more resources could be devoted to creating a large language model specific to the medical domain. This would entail carefully selecting information relevant to the purpose of the model, choices regarding the model architecture, and optimizing training [40]. In addition, RLHF could be used with a population of medical specialists to further optimize model output. Future work could then investigate how a fine-tuned, general large language model compares with domain-specific large language models. Furthermore, such models could be combined with other specialized or general large language models in order to create an ensembled model. Ensembled models tend to achieve higher performance due to leveraging relative strengths of different models. Recent work has investigated creating ensembles of large language models [41], and this could also be investigated in the medical context. The proposed method of fine-tuning ChatGPT could be used as one component of a larger ensembled model capable of providing higher-quality medical information.

Conclusions

The explosive research in large language models is exciting, as is their relatively high performance on a broad number of tests. Performant large language models have the potential to positively impact many different domains, especially the medical domain, where there can be many shortages. In many countries, wait times to see general practitioners can be long, and consultations could be streamlined with access to a performant model.

While most studies benchmarking large language models in medicine do not make use of prompt engineering or fine tuning, we investigated both in the context of general practice by using sample AKT questions provided by the Royal College of General Practitioners. Notably, we demonstrated that on the given dataset, prompt engineering does not appear to improve model performance over a baseline. However, we did observe a statistically significant increase in model performance after applying fine tuning. Previous studies seemed to indicate that using prompt engineering in conjunction with ChatGPT (or even baseline performance) alone was sufficient to obtain passing marks in many examinations; however, our results show that this is not the case for general practice.

We envision various avenues for future work. First, our fine-tuning process focused specifically on providing information in the same format that the model would be tested on. It would be possible to create a more extensive fine-tuning process by including information from reputable online, medically oriented sources. Such a methodology could create a better knowledge base for the model when answering questions. Specifically, because the dataset in question is most pertinent to general practitioners in the United Kingdom, the model could be fine-tuned with information from the relevant authorities, such as the Royal College of General Practitioners or the National Health Service. Second, we believe that additional studies investigating the performance of fine-tuned models on other medical board exams versus prompt engineering and baseline performance could be interesting. More studies could help quantify how impactful fine tuning is on model performance for specific tests of knowledge. Finally, the creation of a large language model with a curated training set (instead of relying solely on fine tuning) could provide even stronger results, though the creation of such a dataset could be nontrivial.

Source of funding: This work was funded from the authors' own resources.

Conflicts of interest: The authors declare no conflicts of interest.

References

1. Knight D, Aakre CA, Anstine CV, et al. Artificial intelligence for patient scheduling in the real-world health care setting: A metanarrative review. *Health Policy and Technology* 2023; 12(8): 100824, doi: 10.1016/j.hlpt.2023.100824.
2. Stenzinger A, Alber M, Allgäuer M, et al. Artificial intelligence and pathology: From principles to practice and future applications in histomorphology and molecular profiling. *Semin Cancer Biol* 2022; 84: 129–143, doi: 10.1016/j.semcancer.2021.02.011.
3. Syed AB, Zoga AC. Artificial intelligence in radiology: current technology and future directions. *Semin Musculoskelet Radiol* 2018; 22(5): 540–545.
4. Chorney W, Wang H. Towards federated transfer learning in electrocardiogram signal analysis. *Comput Biol Med* 2024; 170: 107984, doi: 10.1016/j.combiomed.2024.107984.
5. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019; 25(1): 24–29.
6. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 2020; 33: 1877–1901.
7. Gilson A, Safranek CW, Huang T, et al. 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 Med Educ* 2023; 9: e45312, doi: 10.2196/45312.
8. Garabet R, Mackey BP, Cross J, et al. ChatGPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. *Med Sci Educ* 2023; 34(1): 145–152, doi: 10.1007/s40670-023-01956-z.
9. Hoffmann K, Schober J, Schäfer, et al. Tackling the reasons for GP shortage: The workload of GPs in rural and urban areas in Austria. A cross-sectional study within the framework of QUALICOPC: Kathryn Hoffmann. *The European Journal of Public Health* 2013; 23(Suppl. 1), doi: 10.1093/eurpub/ckt126.310.
10. Harmon DM, Noseworthy PA, Yao X. The digitization and decentralization of clinical trials. *Mayo Clin Proc* 2023; 98(10): 1568–1578, doi: 10.1016/j.mayocp.2022.10.001.
11. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a Large Language Model (ChatGPT) in General Practice with the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Med Educ* 2023; 9: e46599, doi: 10.2196/46599.
12. Zalzal HG, Cheng J, Shah RK. Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education. *OTO Open* 2023; 7(4): e94, doi: 10.1002/oto2.94.
13. Bonetti MA, Giorgino R, Gallo Afflitto G, et al. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? *Ann Biomed Eng* 2024; 52(4): 745–749, doi: 10.1007/s10439-023-03318-7.
14. Jung LB, Gudera JA, Wiegand TL, et al. ChatGPT Passes German State Examination in Medicine with Picture Questions Omitted. *Dtsch Arztebl Int* 2023; 120(21): 373–374, doi: 10.3238/arztebl.m2023.0113.
15. Ozttermeli AD, Ozttermeli A. ChatGPT performance in the medical specialty exam: An observational study. *Medicine (Baltimore)* 2023; 102(32): e34673, doi: 10.1097/MD.00000000000034673.
16. Wang H, Wu W, Dou Z, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. *Int J Med Inform* 2023; 177: 105173, doi: 10.1016/j.ijmedinf.2023.105173.
17. Zhu L, Mou W, Yang T, et al. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. *Resuscitation* 2023; 188: 109783, doi: 10.1016/j.resuscitation.2023.109783.
18. Yanagita Y, Yokokawa D, Uchida S, et al. Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study. *JMIR Form Res* 2023; 7: e48023, doi: 10.2196/48023.
19. Teebagay S, Colwell L, Wood E, et al. Improved Performance of ChatGPT-4 on the OKAP Examination: A Comparative Study with ChatGPT-3.5. *J Acad Ophthalmol* 2023; 15(2): e184–e187, doi: 10.1055/s-0043-1774399.
20. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* 2023; 93(5): 1090–1098, doi: 10.1227/neu.0000000000002551.
21. White J, Fu Q, Hays S, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv 2023; arXiv:2302.11382 [cs.CL], doi: 10.48550/arXiv.2302.11382.
22. Polak MP, Morgan D. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. arXiv 2023; arXiv:2303.05352 [cs.CL], doi: 10.48550/arXiv.2303.05352.
23. Shin E, Ramanathan M. Evaluation of prompt engineering strategies for pharmacokinetic data analysis with the ChatGPT large language model. *J Pharmacokinet Pharmacodyn* 2024; 51(2): 101–108, doi: 10.1007/s10928-023-09892-6.
24. Zhang W, Wang Q, Kong X, et al. Fine-tuning ChatGPT Achieves State-of-the-Art Performance for Chemical Text Mining. *ChemRxiv* 2023, doi: 10.26434/chemrxiv-2023-k7ct5.
25. Latif E, Zhai X. Fine-tuning ChatGPT for Automatic Scoring. arXiv 2023; arXiv:2310.10072 [cs.CL], doi: 10.48550/arXiv.2310.10072.
26. Fan Y, Jiang F, Li P, et al. GrammarGPT: Exploring Open-Source LLMs for Native Chinese Grammatical Error Correction with Supervised Fine-Tuning. arXiv 2023; arXiv:2307.13923 [cs.CL], doi: 10.48550/arXiv.2307.13923.
27. Wu T, He S, Liu J, et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA. Journal of Automatica Sinica* 2023; 10(5): 1122–1136.
28. Tan Y, Min D, Li Y, et al. Evaluation of ChatGPT as a question answering system for answering complex questions. arXiv 2023; arXiv:2303.07992 [cs.CL], doi: 10.48550/arXiv.2303.07992.
29. Zhou J, Ke P, Qiu X, et al. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng* 2023; 25(1–3): 1–6.
30. Abuyaman O. Strengths and Weaknesses of ChatGPT Models for Scientific Writing About Medical Vitamin B12: Mixed Methods Study. *JMIR Form Res* 2023; 7: e49459, doi: 10.2196/49459.
31. Ekin S. Prompt Engineering for ChatGPT: A Quick Guide To Techniques, Tips, and Best Practices. *Authorea Preprints* 2023, doi: 10.36227/techrxiv.22683919.v1.
32. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. arXiv 2022; arXiv:2210.11416 [cs.CL], doi: 10.48550/arXiv.2210.11416.
33. Royal College of General Practitioners. MRCGP: Applied Knowledge Test (AKT). Available from URL: <https://www.rcgp.org.uk/mrcgp-exams/applied-knowledge-test>.
34. Royal College of General Practitioners. Akt example questions. 2019. Available from URL: <https://gp-training.hee.nhs.uk/cornwall/wp-content/uploads/sites/86/2021/04/RCGP-Sample-questions-2019-with-answers.pdf>.
35. Royal College of General Practitioners. MRCGP Applied Knowledge Test (AKT) Feedback Report AKT 48, April 2023. Available from URL: <https://www.rcgp.org.uk/getmedia/e7ff7308-cd7a-4038-9bdf-fe179736f2e3/April-2023-AKT-feedback-report.pdf>.

36. Royal College of General Practitioners. MRCGP Applied Knowledge Test (AKT) Feedback Report AKT 49, October 2023. Available from URL: <https://www.rcgp.org.uk/getmedia/36507354-f398-4b65-a79c-c0d45edfcac3/October-2023-AKT-Feedback-Report.pdf>.
37. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ* 2024; 17(5): 926–931, doi: 10.1002/ase.2270.
38. Shin-Yee Wong R, Ming LC, Raja Ali RA. The Intersection of ChatGPT, Clinical Medicine, and Medical Education. *JMIR Med Educ* 2023; 9: e47274, doi: 10.2196/47274.
39. Kaufmann T, Weng P, Bengs V, et al. A survey of reinforcement learning from human feedback. arXiv 2023; arXiv:2312.14925 [cs.CL], doi: 10.48550/arXiv.2312.14925.
40. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA* 2023; 330(9): 866–869.
41. Jiang D, Ren X, Lin BY. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. arXiv 2023; arXiv:2306.02561 [cs.CL], doi: 10.48550/arXiv.2306.02561.

Tables: 2

Figures: 3

References: 41

Received: 20.08.2024

Reviewed: 05.11.2024

Accepted: 15.12.2024

Address for correspondence:

Wesley Chorney, PhD

School of Medicine

University College Cork, Ireland

E-mail: wes.chorney@gmail.com