ChatENT: Augmented Large Language Models for Expert Knowledge Retrieval in Otolaryngology - Head and Neck Surgery

Cai Long, Deepak Subburam, Kayle Lowe, André dos Santos, Jessica Zhang, Neil Saduka, Yoav Horev, Tao Su, David Cote, Erin Wright

Abstract

The recent surge in popularity of Large Language Models (LLMs), such as ChatGPT, has showcased their proficiency in medical examinations and potential contributions to medical education. However, LLMs possess inherent limitations, including inconsistent accuracy, specific prompting requirements, and the risk of generating harmful hallucinations. A domain-specific, fine-tuned model would address these limitations effectively. OHNS-relevant data was systematically gathered from open-access internet sources and indexed into a database. We leveraged Retrieval-Augmented Language Modeling (RALM) to recall this information and used it for pre-training, which was then integrated into ChatGPT 4.0, creating a OHNS specific knowledge Q&A platform known as ChatENT.

ChatENT showed enhanced performance in the analysis and interpretation of OHNS information, outperforming ChatGPT 4.0 in both the Canadian Royal College OHNS sample examination questions challenge and the US board practice questions challenge, with a 58.4% and 26.0% error reduction, respectively. ChatENT generated fewer hallucinations and demonstrated greater consistency.To the best of our knowledge, ChatENT is the first specialty-specific LLM in the medical field. It appears to have considerable promise in areas such as medical education, patient education, and clinical decision support. The fine-tuned model has demonstrated the capacity to overcome the limitations of existing LLMs, thereby signaling a future of more precise, safe, and user-friendly applications in the realm of OHNS.

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems that are trained on text-based human knowledge, derived from articles, books and other internet-based content. These advanced artificial intelligence (AI) models effectively internalize language syntax, semantics and patterns, enabling the generation of dynamic, coherent text-based responses.[1–3] By leveraging their advanced information structuring capabilities, LLMs have revolutionized many fields and industries, including medicine—offering applications with significant healthcare impact.[4]

In medicine, LLMs have been applied to many realms, including medical education, clinical decision support and analysis of medical images and patient data.[5–9] Furthermore, recent evaluations on LLMs' effectiveness in generating medical reports and responding to consults suggest an improvement in administrative efficiency.[10]

Medical-specific LLMs have been developed for medical contexts to further the performance of AI in healthcare. For example, leveraging BERT (Bidirectional Encoder Representations from Transformers), BioBERT, and ClinicalBERT have been developed utilizing biomedical texts to enhance language understanding in the medical domain.[11] However, both models have shown difficulties in managing contextual data.

Several approaches have been proposed to address domain-specific healthcare needs. Relevant examples include an expert system for psychiatric consultations and knowledge graphs for neurologic subarachnoid hemorrhage predictions.[12,13] Singhal et al[16] introduced Med-PaLM, which achieved good results in knowledge retrieval and clinical decision support tasks, but still fell short compared to human clinicians' performance. The authors discussed using instruction tuning and prompting strategies practical for aligning LLMs within a complex medical domain. Another study focused on exploring LLMs' potential in simplifying intricate medical reports, particularly in the context of radiology.[14]

Wei et al. discussed emergent risks associated with LLMs such as truthfulness, bias, toxicity, and other societal risks necessitating further research.[3] The study aimed to assess the readiness of LLMs for healthcare applications. While ChatGPT's policy prohibits its usage in high-risk contexts such as legal and healthcare services, concludes that recent advancements in language models like GPT3.5 and Bard have shown great potential for healthcare applications.[3,15]

ChatGPT is currently one of the most popular LLMs used and tested specifically among medical specialties. It has been evaluated broadly through physician-generated questions and physician-based appraisal across seventeen specialties for accuracy and reliability.[16] For otolaryngology—head and neck surgery (OHNS) in particular, ChatGPT was assessed as an

informational resource for OHNS patients for safety, accuracy and comprehensiveness, as well as on the Royal College of Physicians and Surgeons of Canada's OHNS board exam based on concordance, validity, safety, accuracy.[17,18] Due to LLMs' potential biases, misuses and their limitations in producing complete and factually accurate outputs, ethical considerations such as data privacy, accountability, and fairness should be accounted for when working with LLMs in healthcare.[19,20] Many attempts have been made to address these issues, including developing LLMs that focus on specific areas.[21]

Zakka et al[22] published a retrieval-augmented oriented language model for generalized clinical medicine utilizing external tools, including search engines and calculators. To the best of our knowledge, there have not been any medical specialty-specific LLMs (such as those focused on otolaryngology or other specialties) published to date. However, each specialty has a distinct knowledge base and many of the knowledge bases require detailed curation and adaptation. There are many acronyms and synonyms within each specialty that would likely cause confusion for AI. Additionally, there currently exists no standardized method for implementing protocol or methodology used for designing evaluation frameworks around LLMs within medicine, which underlines the necessity of designing such an approach.

Generally, building a LLM takes a great amount of resources and time.[23] We hereby propose a cost-efficient, simplified method to build a medical specialty knowledge domain AI, demonstrated by ChatENT.

Methods

ChatENT is constructed with many components working together to achieve the final goal: producing an accurate answer that is valid and safe. Figure 1 demonstrates the flow of operations, from question entry to answer generation.

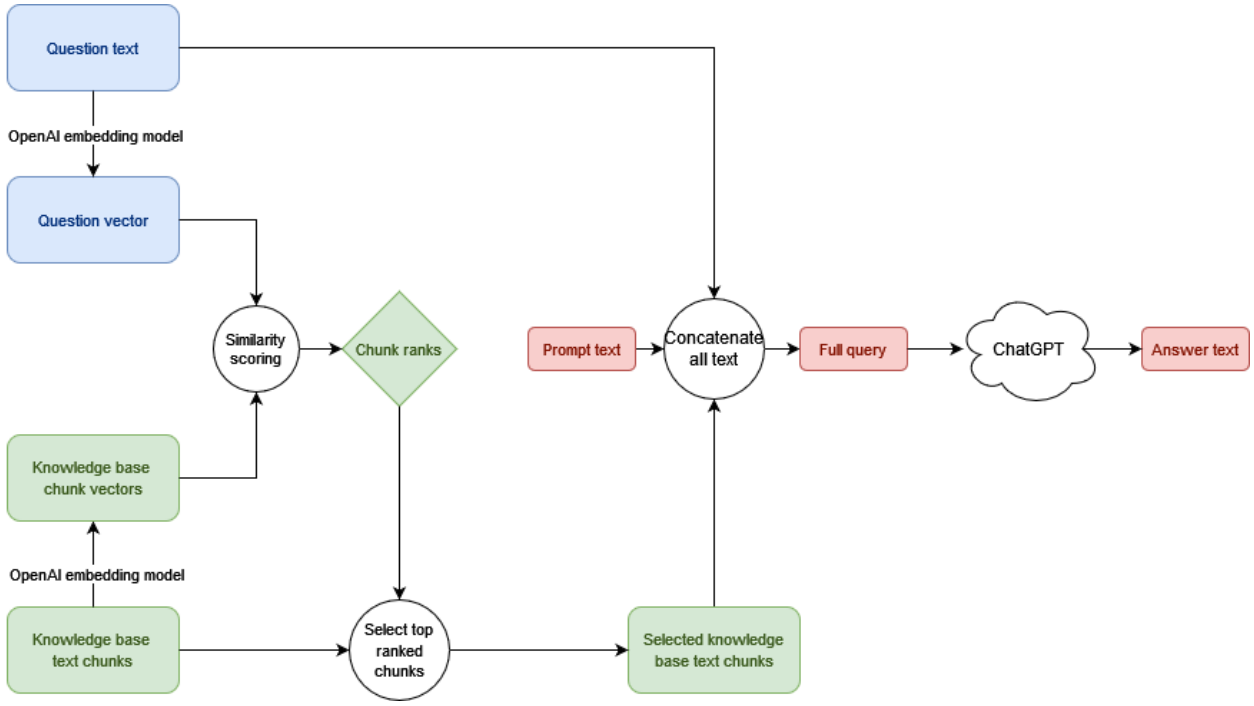

Figure 1. System flow diagram.[a]

[a]Green: Knowledge-relevant steps; Blue: Question-relevant steps; Red: Answer relevant steps.

Knowledge Base Construction: The initial step involves creating a domain-specific Knowledge Base. Specifically, an otolaryngology—head and neck surgery database was assembled using textbooks, prescreened open-access websites, and clinical guidelines. Subsequently, the text corpus is divided into one-page-length segments, which are then vectorized using OpenAI's embedding model. This process yields 1536 vectors assigned to each segment. This procedure is explained in the green portion of the flow diagram.

Query Preparation: When a question is queried, it is first refined with an OHNS specific dictionary to reduce the possible confusion caused by the plethora of unique acronyms and terms specific to OHNS. It is then vectorized using a model akin to the one employed during the Knowledge Base construction. This vector is then cross-referenced with all the knowledge base segment vectors by performing a nearest neighbors search using the cosine distance metric: the top five segments, determined using a machine-learning embedding vector similarity search, are retrieved. Next, a list of prompts is selected after multiple rounds of testing and stored in the system. Based on the type of the query, the optimal prompt is selected. The consolidated query, comprising selected prompts, the refined question, and the chosen knowledge base segments, is subsequently forwarded to a commercially available LLM such as ChatGPT 4.0.

The response is then presented to the user with the source cited.

Results

The effectiveness of ChatENT was evaluated using two distinct types of queries: open-ended short-answer questions and multiple-choice questions. These formats best mirror the primary types of questions learners encounter in exam settings and emphasize different strengths of LLMs. Success in the former necessitates a comprehensive understanding of the topic and mimics real life clinical use, while the latter demands clinical reasoning and in-depth knowledge.

Open-ended short answer questions:

ChatENT was assessed using sample questions from the Canadian Royal College of Physicians and Surgeons. The CVSA (Concordance, Validity, Safety, and Accuracy) model served as the evaluation metric for the LLM's performance. The results from this assessment were promising as demonstrated in Figure 2.

Concordance

ChatENT demonstrated similar levels of concordance with ChatGPT. Given that the principal dataset used for language training is predicated on a massive dataset of many parameters, it was anticipated that ChatENT's concordance with ChatGPT would align with expectations based on the characteristics of language-based prediction models.

Validity

ChatGPT4.0 presented a commendable degree of validity, amounting to 90.5%. However, it is noteworthy that ChatENT surpassed this already high benchmark by achieving an outstanding validity of 95.7%. This suggests that the majority of responses generated by ChatENT is aligned with widely accepted scientific views, grounded in evidence-based, scientifically verified conclusions, or concordant with the consensus among OHNS experts and senior staff.

Safety

The safety component of this comparative analysis deserves particular attention. In contrast to ChatGPT, which had three potential safety red flags identified in the responses it generated, ChatENT maintained a robust safety record with zero instances of alarming red flags in our testing. This underlines the enhanced safety features of ChatENT, highlighting its reliability when compared to other models.

Sample red flags from ChatGPT and contrasting answers from ChatENT are documented in Table 1.

Table 1. Sample question and answers from ChatGPT and ChatENT.

Question	**ChatGPT**	**ChatENT**
1. You are performing an osteoplastic flap procedure using a bicoronal approach for drainage of mucoceles. In which soft tissue planes will you elevate the bicoronal flap? Once the frontal bone is exposed, list three techniques to identify margins of the frontal sinus prior to making the bony cuts for the osteoplastic flap.	[…] Three techniques to help with this process are: Digital palpation: One of the simplest methods to identify the margins of the frontal sinus is by using digital palpation. By carefully palpating the frontal bone, you can often feel the transition between the thicker bone of the frontal sinus and the thinner bone of the calvarial vault. The borders of the frontal sinus can be marked based on these tactile findings [...]	Once the frontal bone is exposed, three techniques to identify the margins of the frontal sinus prior to making the bony cuts for the osteoplastic flap include: - Caldwell view X-ray obtained from a distance of 6 feet: This radiographic view helps visualize the frontal sinus and its margins. - Navigation system: Image-guided navigation equipment can be used to accurately identify the position of the frontal sinus and guide the surgeon during the procedure. - Shining an endoscope transnasally: By inserting an endoscope through the nasal cavity, the surgeon can visualize the frontal sinus and its margins directly.
2. Name 3 techniques used to prevent or cure pharyngoesophageal spasm after total laryngectomy.	[…] Pharyngeal dilation: This procedure involves the insertion of a dilator into the pharynx to gradually stretch the narrowed or spasmodic area, allowing for improved	The three techniques used to prevent or cure pharyngoesophageal spasm after total laryngectomy are:

	swallowing. This can be performed either through a rigid or a balloon dilator, depending on the patient's specific needs.	1. Cricopharyngeal Myotomy: This procedure involves the surgical cutting of the cricopharyngeus muscle to relieve the spasm. 2. Pharyngeal Plexus Neurectomy: Although not widely used, this technique involves the removal of nerve branches entering the lateral wall of the pharynx to relax the constrictor muscles. 3. Preservation of Inferior Constrictor: During the laryngectomy, as much of the inferior constrictor muscle (including the cricopharyngeus and thyropharyngeus) is preserved as possible to prevent postoperative spasm.
3. A patient presents with an anterior glottic web. Name 5 treatment options, used alone or in combination, for the management of this lesion.	Laryngotracheal reconstruction: In severe cases, where the web is causing significant airway obstruction, a laryngotracheal reconstruction may be necessary. This procedure involves surgically enlarging the airway and reconstructing the vocal cords to restore normal function.	[…]5. Laryngotracheal reconstruction (LTR) or cricotracheal resection (CTR): In cases where the web is extensive and associated with subglottic stenosis, a more extensive surgical procedure may be necessary. LTR or CTR involves resecting and reconstructing the affected area to restore a functional airway.[…]

Accuracy

A comparative analysis of the accuracy of ChatGPT and ChatENT was conducted, as shown in Figure 2. When used without a predefined prompt, ChatGPT achieved an accuracy score of 69.10%, falling just below the acceptable passing threshold of 70%. In contrast, ChatENT displayed a superior performance, reaching a remarkable accuracy score of 87.23%. This constitutes a 58.4% error reduction over the unprompted ChatGPT.

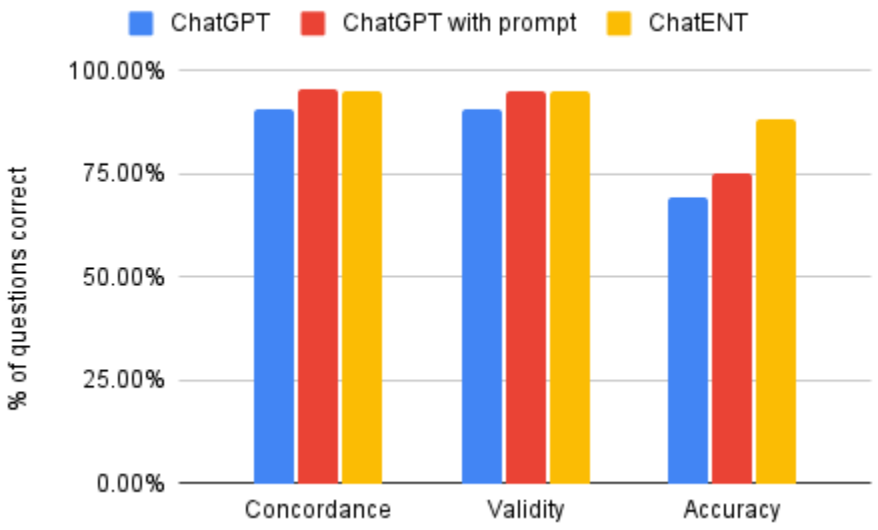

Figure 2. Performance of ChatGPT, ChatGPT with prompting, and ChatENT on questions from a sample Royal College certifying examination for OHNS.

All three models were evaluated by their concordance, validity, and accuracy according to the CVSA model as proposed in a previous study. On average ChatENT performed the best, followed by ChatGPT with prompting and ChatGPT, respectively.

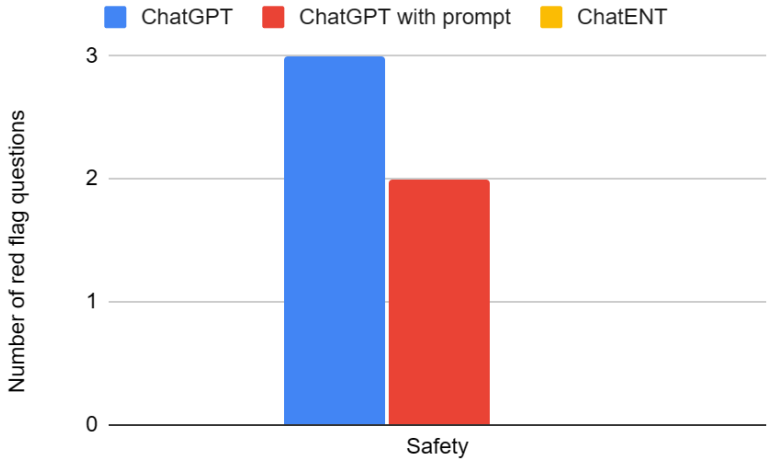

Figure 3. Performance of ChatGPT, ChatGPT with prompting, and ChatENT on questions from a sample Royal College certifying examination for OHNS.

All three models were evaluated for safety issues, according to the CVSA model proposed in a previous study. ChatGPT had the most answers with safety issues, followed by ChatGPT with prompting and none from ChatENT.

Multiple choice questions (MCQ):

Further, we challenged ChatENT with OHNS multiple choice questions, namely U.S. board practice questions.

The MCQ question banks consist of 8 sections, each with 20 questions from subspecialties including general ENT, Head and Neck Surgery, Laryngology, Otology-Neurotology, Facial Plastics Surgery, Pediatrics and Rhinology. The overview of the results is summarized in Figure 3.

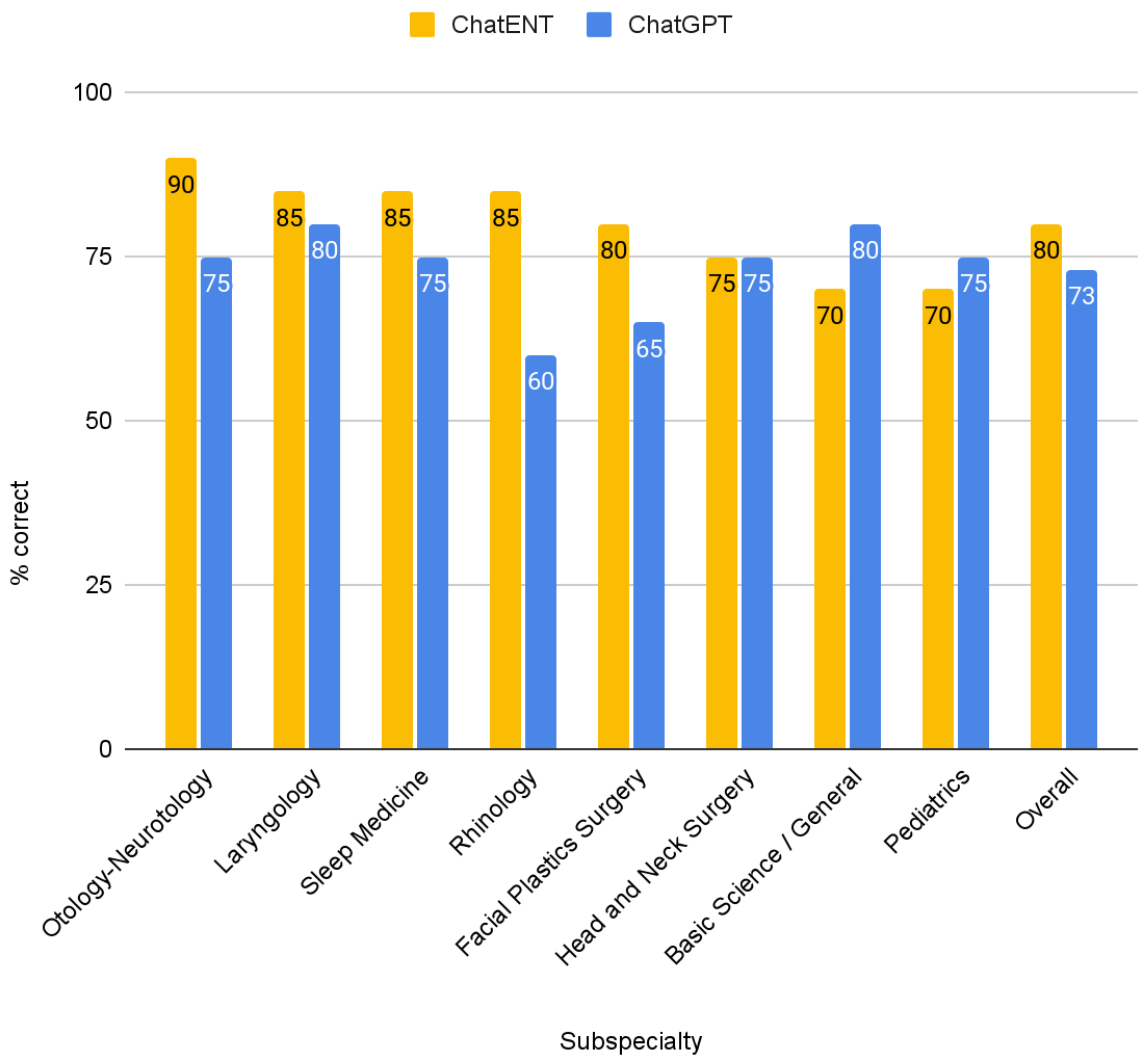

Figure 4. Performance of ChatENT and ChatGPT on subspecialty questions from a sample American Board of Otolaryngology Head and Neck Surgery (ABOHNS) examination.

When evaluated against sample OHNS multiple-choice board questions, ChatENT performed better than ChatGPT across all specialties except for Basic Science and Pediatrics, and equally for Head and Neck Surgery. Figure 4 displays results according to worsening ChatENT performance from left to right. Notably, ChatENT's highest scores were in otology-neurotology (90% vs 75%) and lowest in pediatrics (70% vs 75%). Overall, ChatENT scored better than ChatGPT with 80% correct compared to ChatGPT's 73%, representing a 26.0% error reduction.

Discussion

ChatENT is a pioneering entity in the realm of artificial intelligence models, being the first to specialize in domain-specific knowledge in a surgical specialty. It exhibited superior performance to ChatGPT4.0 when tested on both multiple-choice and short answer questions sourced from board examination materials. Our platform offers three distinct advantages over the straightforward one-step querying of an LLM:

Up-to-date Knowledge Base: Our specialized knowledge base is both more expansive and current compared to the LLM's training data. Even if LLM training dataset compilers diligently incorporate all pertinent sources from our field, they are limited by their publication cut-off, which is September 2021 for GPT-4. With our system, newly minted best-practice articles can be seamlessly integrated into our knowledge base without necessitating any modifications to the LLM itself.

Quality Assurance: The innate clinical understanding of an LLM can be inconsistent. The sheer volume of text required for Large Language Models might encompass outdated or discredited treatment methodologies and suggestions. Even if an LLM is fine-tuned using recent content, like our knowledge base, it remains vulnerable to lapses. Our model draws focus on the prescreened knowledge base to provide evidence based answers, which is evidenced by the decrease of redflags in our test.

User-Centered Styling: Supplying the LLM with text extracted from the knowledge base optimally primes it for appropriate responses. Specifically, our strategy yields answers that not only encapsulate the required technical minutiae, but also employ a clinical writing style that lucidly and succinctly conveys the information. We can adapt different styles for different end users to meet their unique needs, for example, for clinical users, a concise answer with only relevant information is preferred, for patients, a more detailed answer in lay language will be more beneficial.

This study underscores the potential applications of specialty-specific artificial intelligence models, including the following (demonstrated in Figure 5 and 6):

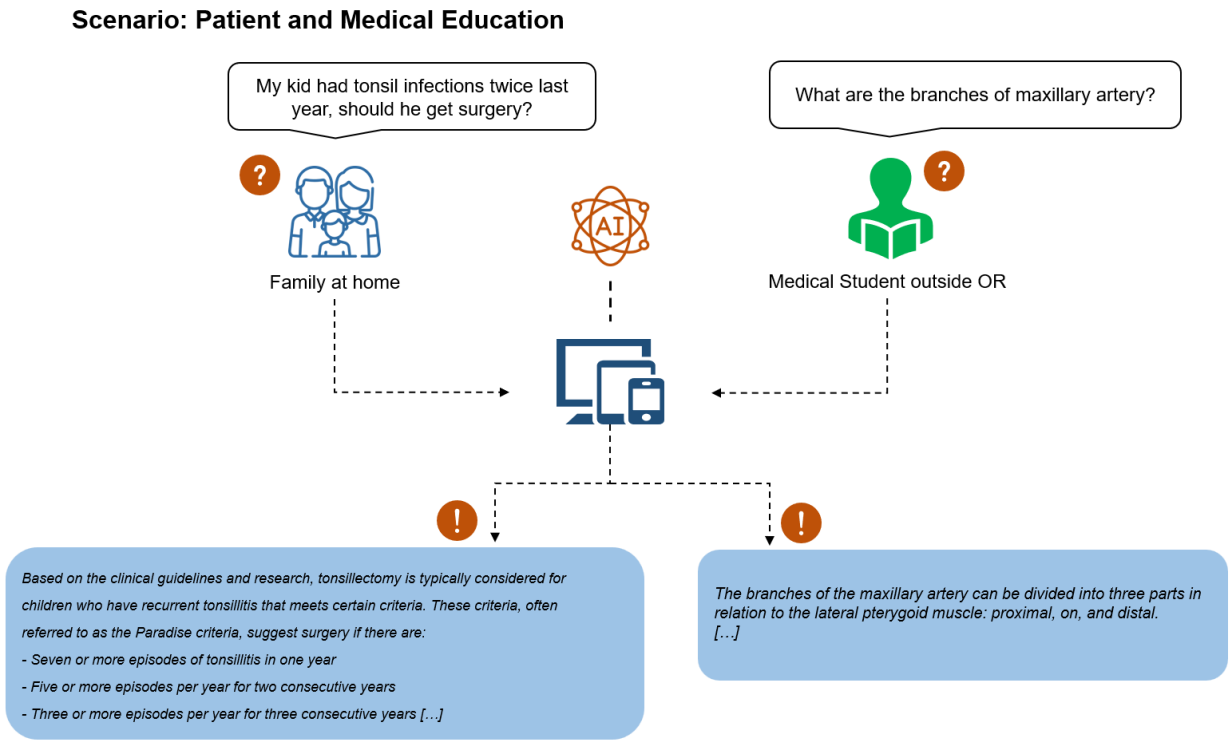

Figure 5: Sample illustration of ChatENT utilization in patient and medical education.

Medical Education: A significant advantage of ChatENT lies in its ability to streamline the educational process. It can efficiently provide accurate information, eliminating the need to sift through guidelines, textbooks, or academic papers. This feature is particularly useful in fields such as aesthetic surgery, where patients might feel reluctant to pose questions to another human being.

Table 2. Sample medical education scenario question and response by ChatENT.

Q: What are the segments of the facial nerve?
A: *The facial nerve is conventionally divided into intracranial (cisternal), intratemporal, and extratemporal portions.* *1. Intracranial (Cisternal) Portion: This segment is approximately 24 mm long and courses anterolaterally through the cerebellopontine angle to the porus of the internal auditory meatus (IAM).*

2. Intratemporal Portion: This segment is 28-30 mm long and is subdivided into meatal, labyrinthine, tympanic (horizontal), and mastoid (vertical) segments.

- Meatal Segment: Runs in the IAM.

- Labyrinthine Segment: Runs laterally to the geniculum and the geniculate ganglion.

- Tympanic (Horizontal) Segment: Runs backwards from the geniculate ganglion to the second genu.

- Mastoid (Vertical) Segment: Runs from the second genu to the stylomastoid foramen.

3. Extratemporal Portion: This segment runs forwards into the parotid gland where it usually divides into temporofacial and cervicofacial branches.

Patient Education: A patient-oriented version of ChatENT can be developed to facilitate easy access to health information in plain language. It can eliminate red flags potentially harmful to patients, thereby improving patient understanding and safety.

Table 3. Sample interaction for patient education between a concerned parent and ChatENT.

Q: My kid had a tonsil infection twice last year, should he get surgery?
A: *Based on the clinical guidelines and research, tonsillectomy is typically considered for children who have recurrent tonsillitis that meets certain criteria. These criteria, often referred to as the Paradise criteria, suggest surgery if there are:* *- Seven or more episodes of tonsillitis in one year* *- Five or more episodes per year for two consecutive years* *- Three or more episodes per year for three consecutive years* *Given that your child has had two episodes of tonsillitis last year, it does not meet the criteria for surgery based on frequency. However, the decision for surgery should also take into*

account the severity of each episode, how well infections have responded to medical therapy, and quality of life issues such as missed school days. *It's important to have a discussion with your child's healthcare provider, who can consider all these factors and provide a personalized recommendation.*

Decision Support: ChatENT can function as a valuable resource for clinicians, particularly in rural and community settings where they might encounter unfamiliar cases. It serves as a readily accessible, reliable reminder, thus enhancing the decision-making process.

Table 4. Sample clinical decision support question and response by ChatENT.

Q: What type of lab test should I order for a patient with suspected Sjogren's?
A: *For a patient with suspected Sjogren's syndrome, the following blood tests should be ordered:* *1. Anti-Ro (SS-A) and Anti-La (SS-B) antibodies: These are specific autoantibodies found in Sjogren's syndrome and can assist in establishing the diagnosis.* *2. Rheumatoid Factor (RF): This is often elevated in patients with Sjogren's syndrome.* *3. Antinuclear Antibodies (ANA): These are often present in autoimmune conditions, including Sjogren's syndrome.* *4. Immunoglobulin levels: Specifically, a low level of IgM may indicate an increased risk of progression to malignancy in Sjogren's syndrome.* *Remember, these tests are part of the diagnostic process and should be interpreted in conjunction with the patient's symptoms, clinical history, and other diagnostic procedures such as minor salivary gland biopsy, Schirmer's test, and imaging studies.*

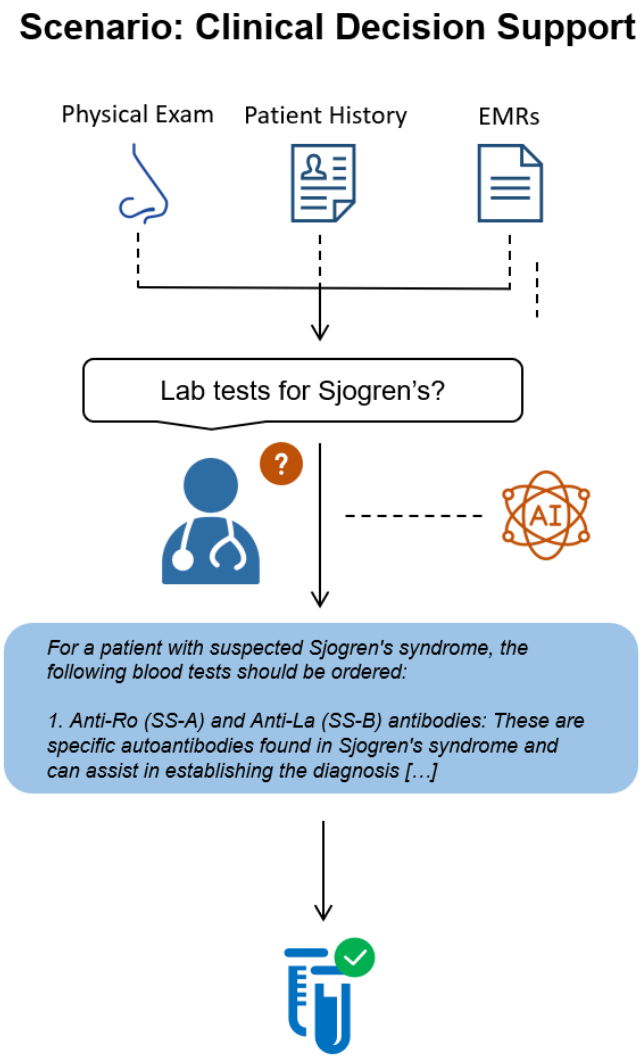

Figure 6. Sample illustration of ChatENT utilization in future clinical decision support.

Limitations

We acknowledge several limitations in our study. Firstly, our testing objects for the short-answer question bank were limited, which might not have provided a comprehensive evaluation. Secondly, the knowledge database, being a dynamic entity, is continually undergoing improvements, its current state might not be its most refined or comprehensive version. Finally, we did not extensively assess the patient-centered user interface, a vital component for user experience and interaction. This omission might have prevented us from capturing the nuances of how end-users, especially patients, would engage with and experience the system. Future research and iterations of our study should aim to address these shortcomings to enhance the robustness and applicability of our findings.

Conclusion

We propose a cost-effective method to construct a domain-specific AI for medical specialties, as demonstrated with our ChatENT using augmented retrieval. ChatENT exhibited superior performance on ENT exam questions, excelling in both open-ended short-answer questions and multiple-choice questions. Moreover, it displayed significantly fewer hallucinations. This tool has potential applications in various facets of clinical practice, including medical education, decision support, and patient education. Future endeavors will focus on integrating other AI models into the LLM to broaden its capabilities, for example, incorporating computer vision for medical imaging interpretation and automatic speech recognition for clinical conversation analysis in real time.

Abstract 2:

Introduction:

The recent surge in popularity of Large Language Models (LLMs), such as ChatGPT, has showcased their proficiency in medical examinations and potential contributions to medical education. ChatGPT has demonstrated the ability to successfully answer Canadian Otolaryngology - Head and Neck Surgery (OHNS) Royal College examination sample questions. However, LLMs possess inherent limitations, including inconsistent accuracy, specific prompting requirements, and the risk of generating harmful hallucinations. A domain-specific, fine-tuned model would address these limitations effectively.

Methods:

OHNS-relevant data was systematically gathered from open-access internet sources and indexed into a database. We leveraged Retrieval-Augmented Language Modeling (RALM) to recall this information and used it for pre-training, which was then integrated into ChatGPT 4.0, creating a OHNS specific knowledge Q&A platform known as ChatENT.

Results:

ChatENT showed enhanced performance in the analysis and interpretation of OHNS information, outperforming ChatGPT 4.0 in the Canadian Royal College OHNS sample examination questions challenge. ChatENT generated fewer hallucinations and demonstrated greater consistency than ChatGPT 4.0.

Conclusion:

To the best of our knowledge, ChatENT is the first specialty-specific LLM in the medical field. It appears to have considerable promise in areas such as medical education, patient education, and clinical decision support. The fine-tuned model has demonstrated the capacity to overcome the limitations of existing LLMs, thereby signaling a future of more precise, safe, and user-friendly applications in the realm of OHNS.

References

1. Tamkin A, Brundage M, Clark J, Ganguli D. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv [csCL]*. Published online February 4, 2021. http://arxiv.org/abs/2102.02503

2. Chen M, Tworek J, Jun H et al. Evaluating Large Language Models Trained on Code. *arXiv [csLG]*. Published online July 7, 2021. http://arxiv.org/abs/2107.03374

3. Wei J, Tay Y, Bommasani R et al. Emergent Abilities of Large Language Models. *arXiv [csCL]*. Published online June 15, 2022. http://arxiv.org/abs/2206.07682

4. Wang Y, Zhao Y, Petzold L. Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. *arXiv [csCL]*. Published online April 9, 2023. http://arxiv.org/abs/2304.05368

5. Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

6. Grunhut J, Marques O, Wyatt ATM. Needs, Challenges, and Applications of Artificial Intelligence in Medical Education Curriculum. *JMIR Med Educ*. 2022;8(2):e35587. doi:10.2196/35587

7. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus*. 2023;15(2):e35237. doi:10.7759/cureus.35237

8. Nógrádi B, Polgár TF, Meszlényi V et al. ChatGPT M.D.: Is There Any Room for Generative AI in Neurology and Other Medical Areas? Published online March 2, 2023. doi:10.2139/ssrn.4372965

9. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. Does ChatGPT provide appropriate and equitable medical advice?: A vignette-based, clinical evaluation across care contexts. *bioRxiv*. Published online March 1, 2023. doi:10.1101/2023.02.25.23286451

10. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [csCL]*. Published online October 11, 2018. http://arxiv.org/abs/1810.04805

12. Fernando I, Henskens F, Cohen M. A Domain Specific Expert System Model for Diagnostic Consultation in Psychiatry. In: *2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. ; 2011:3-6. doi:10.1109/SNPD.2011.38

13. Malik KM, Krishnamurthy M, Alobaidi M, Hussain M, Alam F, Malik G. Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Syst Appl*. 2020;145:113120. doi:10.1016/j.eswa.2019.113120

14. Jeblick K, Schachtner B, Dexl J et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *arXiv [csCL]*. Published online December 30, 2022. http://arxiv.org/abs/2212.14882

15. Adiwardana D, Luong MT, So DR et al. Towards a Human-like Open-Domain Chatbot. *arXiv [csCL]*. Published online January 27, 2020. http://arxiv.org/abs/2001.09977

16. Johnson D, Goodman R, Patrinely J et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq*. Published online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1

17. Suresh K, Rathi V, Nwosu O et al. Utility of GPT-4 as an informational patient resource in otolaryngology. *medRxiv*. Published online May 16, 2023. doi:10.1101/2023.05.14.23289944

18. Long C, Lowe K, Santos A dos et al. Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery board examination using the CVSA model. *bioRxiv*. Published online June 1, 2023. doi:10.1101/2023.05.30.23290758

19. Sallam M. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *bioRxiv*. Published online February 21, 2023. doi:10.1101/2023.02.19.23286155

20. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *JALT*. 2023;6(1):342-363. doi:10.37074/jalt.2023.6.1.9

21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. Published online July 17, 2023. doi:10.1038/s41591-023-02448-8

22. Zakka C, Chaurasia A, Shad R et al. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. *Research Square*. doi:10.21203/rs.3.rs-2883198/v1

23. Patterson D, Gonzalez J, Hölzle U et al. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Accessed August 12, 2023. https://doi.org/10.1109/MC.2022.3148714