ISSN 2007-9737

Application of Large Language Models to the Diagnosis of Respiratory Diseases

Anastasiia A. Shamrikova[1], Svetlana V. Krasilnikova[2], Georgii S. Ignatov[3], Nailya Kubysheva[4], Imre Rudas[5], Muhammad Ahmad[6], Ildar Batyrshin[6,*]

[1] Onecta, Inc., Moscow, Russia

[2] Privolzhsky Research Medical University of Nizhny Novgorod, Russia

[3] Effective Technologies, Inc., Nizhny Novgorod, Russia

[4] Kazan Federal University, Kazan, Russia

[5] Obuda University, Budapest, Hungary

[6] Instituto Politecnico Nacional, Centro de Investigación en Computación, Mexico

aanastasish@gmail.com, mashkovasv@mail.ru, ignatovgrg@yandex.ru, aibolit70@mail.ru, rudas@uni-obuda.hu, mahmad.riaz102@gmail.com, batyr1@cic.ipn.mx

Abstract. The implementation of large language models (LLM) using artificial intelligence can currently become extremely popular for solving various medical problems. Eight publicly available AI systems were prompted to make an otolaryngological diagnosis based on known symptoms obtained using the standard SNOT-22 medical questionnaire. The aim of the study was to find out to what extent modern AI systems can make a diagnosis without prior training. The results showed that most systems, with one exception, performed satisfactorily, achieving an accuracy of 70-80% compared to an accuracy of 84% achieved by a human specialist using various machine learning methods. The advantages and disadvantages of AI systems for medical diagnostics are discussed in the paper.

Keywords. Large language models, artificial intelligence, medicine, respiratory diseases, SNOT-22.

1 Introduction

Making a diagnosis by a professional doctor using endoscopic equipment requires significant time and money from both the patient and the healthcare system. Reducing these costs is possible due to modern cybernetic systems, including neural network models.

Allergic rhinitis is a common disease characterized by chronic allergic inflammation of the nasal mucosa with involvement of the paranasal sinuses in the pathological process [1]. In some patients, the course of allergic rhinitis is aggravated by the formation of hypertrophic and polypous changes in the sinonasal mucosa (SMM), which is considered to be a result of its pathological remodeling due to persistent inflammation. Detection of hypertrophic and polypous changes in the SMM is an important stage in providing medical

care to patients with impaired nasal breathing [2]. The “gold standard” for diagnosing these pathological changes in the SMM is the use of modern visualization methods - endoscopic research methods and radiation diagnostic methods, including high-resolution computed tomography and magnetic resonance imaging. [3] However, despite the high information content and diagnostic value of these diagnostic methods, their use is associated with significant time and material costs, requires highly qualified personnel, expensive equipment and consumables [4].

A rapid alternative method for clinical assessment of the condition of the nose and paranasal sinuses is the use of validated questionnaires such as the Sinonasal Outcome Test – 22 (SNOT-22), Total Nase Symptom Score (TNSS), which allow standardizing the severity of subjective complaints of patients [5]. However, the use of these questionnaires gives a large spread in the assessment of the patient's condition, since it leads to the need to simultaneously take into account a large number of indicators. The use of modern machine learning algorithms [6] allows you to accomplish this task, but requires high qualification of medical personnel and additional training. We assume that it is possible to significantly simplify and improve the results of patient assessment using modern artificial intelligence systems. Since a large number of AI systems are currently publicly available, it is possible to use them for healthcare needs, which eliminates the need to develop and train your own neural networks.

In this study, we examined the capabilities of several publicly available AI systems by providing them with a training set and testing their responses on a test set without training or validation. Although the diagnosis made by AI systems should be treated with caution, it can be used for disease monitoring, which can be carried out by nursing staff or even the patient themselves, relieving highly trained specialists and busy equipment and ensuring that they are consulted only for patients that the AI system classifies as being at risk.

2 Procedure

350 pediatric patients with bronchial asthma (complicating diagnosis) were surveyed using the professional validated and standardized questionnaire SNOT-22, recommended by EPOS2020 [1] to assess the impact of chronic rhinosinusitis and other diseases of the nose and paranasal sinuses on the quality of life of patients. After that, each patient was examined by an otolaryngologist using nasal endoscopy (rigid and flexible). The doctor diagnosed the presence or absence of polypous and hypertrophic changes in the sinonasal mucosa in patients with bronchial asthma. The survey results were coded as Symptoms lines with an additional indication of the diagnosis.

The study was approved by the Ethical Committee of the Volga Region Research Medical University (protocol No.13, dated 10.10.2016). All participants and all primary caregivers gave written informed consent.

Based on the generated database, training data sets of 221 records were randomly selected from the full database with maintaining the ratio of diagnoses in the database (approximately 69:31). The limited size of the test sample (96 records) was determined by the fact that some of publicly available AI systems cannot accept queries that are too long, and when a query length exceeds 100 records, errors or system failures occur. The training data set was sent to several publicly available AI systems without any training. Verification was carried out based on the test set, which was generated from the same database and included survey results and diagnoses for 30 patients. The diagnosis made by the doctor was also encoded as a number in the training set and was absent from the test set. Based on the system responses (predictions of diagnoses in the test set) and their comparison with known diagnoses, statistical characteristics of each of the AI systems were obtained. Along with the responses of the AI systems, their messages were recorded, allowing us to judge on what basis the AI systems arrived at the predicted diagnosis value.

The following 8 AI systems were queried: Grok-3 (reasoning) [7], Gemini 2.0 (reasoning) [8], Gemini 2.5 Pro (thinking) [9], Gemma 3 27B [10], DeepSeek-R1 (reasoning) [11], GPT o3-mini (reasoning) [12], Claude 3.7 (reasoning) [13], Qwen2.5-Max (reasoning) [14]. Example of a prompt given to an AI system:

Hi! I have a list {Symptoms={list},Diagnosis} of 221 entries:
{Symptoms={1,5,0,1,1,0,0,0,0,1,2,0,0,0,0,0,0,0,1,0,0,0,0,0,3,3,0,2,9},{Diagnosis=1}},
{Symptoms={1,15,1,2,2,3,1,4,1,2,1,2,2,5,0,1,0,0,1,1,3,3,3,3,3,3,0,0,20},{Diagnosis=2}},
{Symptoms={2,12,1,0,1,2,1,2,0,1,4,0,1,2,2,1,3,1,0,0,4,2,0,0,1,2,3,2,14},{Diagnosis=2}},
(215 additional records placed here)
{Symptoms={1,5,0,0,0,3,0,3,0,0,0,2,0,0,0,0,0,0,0,1,0,0,2,1,0,0,0,0,4},{Diagnosis=1}},
{Symptoms={1,17,0,0,1,3,0,4,0,1,0,0,0,2,0,0,0,0,3,0,1,0,0,0,0,0,0,0,4},{Diagnosis=2}},
{Symptoms={2,17,2,1,1,1,3,1,1,1,2,3,0,0,0,0,0,0,0,2,1,0,0,0,0,4,0,0,7},{Diagnosis=1}}.
(95 additional records placed here)
What is Diagnosis for symptoms: Symptoms={1,12,2,1,1,2,1,2,1,1,3,2,1,0,0,0,0,0,0,0,2,0,0,0,1,0,1,1,5}.

At the same time, statistical processing of the database was performed by a human (a specialist in the application of statistical methods in medicine) in order to determine what maximum accuracy can be expected using this data set. The methods used were logistic regression (LR), K-means clustering (KM), support vector machines (SV), decision tree (DT), random forest (RF), and gradient boosting (GB). Training without cross-validation and with cross-validation (5-fold cross-validation with approximately 177 training records and 44 test records per fold) were considered.

3 Results

The response results of the eight surveyed AI systems are presented in Table 1. Table 2 shows the corresponding results achieved by a human specialist using six different methods of machine learning.

All AI systems, except Qwen2.5-Max, demonstrated satisfactory accuracy in making a diagnosis based on professional questionnaires, which was 70-80%. The cross-validation estimation results in lower values of accuracy in the ranges of 60-70%.

After cross-validation, the best performance indicators of the systems considered were shown by Gemma 3 27B (80.2%), Gemini 2.0 (79.2%) and Grok-3 (77.1%). The worst performance was shown by the Qwen2.5-Max system (57.3%). Subject to the cross-validation, the best performance was shown by Claude 3.7 (71.5), DeepSeek-R1 and GPT o3-mini (71.1%). The lowest scores were obtained from Grok-3 (62.9%) and Qwen2.5-Max (63.8%).

It should be noted that these indicators vary significantly across multiple calls to AI systems. For example, the Grok-3 system demonstrated an indicator of 62.5% (without cross-validation) in some cases, although its best indicator was significantly higher.

Although none of the AI systems reached the indicators of a human specialist using machine learning methods, these values differ only slightly from the best human indicators.

In case of these methods, the best accuracy of 84.4% was achieved using the logistic regression method. The performance of other methods was in the range of 68-79%, which is comparable or even lower than the performance of many AI systems.

Under cross-validation condition, the best performance was achieved by random forest and gradient boosting methods (75.2 and 74.7%, respectively). It should be noted that other methods, taking into account cross-validation, gave similar accuracy parameters, and even in the worst cases (LR and SVM), they slightly exceeded the indicators of the AI outsider systems.

The response time also varied significantly from 29 to 327 seconds, although this may have depended on how thoroughly the system tried to explain its actions. In the cases of Claude 3.7 and Qwen2.5-Max timing was unavailable. Gemini 2.0 looked for similar combinations of symptoms in each list and, based on these similarities, it determined which diagnosis (1 or 2) was most likely.

Gemini 2.5 Pro chose a diagnosis by finding similar combinations of symptoms in the examples.

Gemma 3 27B analyzed the vector of symptoms and, based on the patterns found in the provided data, determined the most likely diagnosis. This model could not calculate the diagnosis for all records at once, so the test data was loaded in chunks. DeepSeek-R1 determined the diagnosis based on the analysis of patterns in the data, including: (1) The sum of symptom values (high sums more often indicated diagnosis 2); (2)

Table 1. Results of the diagnosis predictions carried out by various AI systems

Parameter	AI system							
	Grok-3	**Gemini 2.0**	**Gemini 2.5 Pro**	**Gemma 3 27B**	**DeepSeek-R.1**	**GPT o3-mini**	**Claude 3.7**	**Qwen2.5-Max**
True positive, TP	15	13	20	13	22	22	19	16
True negative, TN	59	63	52	64	48	48	53	39
False negative, FN	7	3	14	2	18	18	13	27
False positive, FP	15	17	10	17	8	8	11	14
Accuracy, %	77.1	79.2	75.0	**80.2**	72.9	72.9	75.0	57.3
Precision, %	68.2	81.3	58.8	**86.7**	55.0	55.0	59.4	37.2
Recall, %	50.0	43.3	66.7	43.3	**73.3**	**73.3**	63.3	53.3
F1-score, %	57.7	56.5	62.5	57.8	**62.9**	**62.9**	61.3	43.8
After 5-fold cross-validation								
Accuracy, %	62.9	67.4	70.2	65.2	71.1	71.1	**71.5**	63.8
Precision, %	46.1	51.3	53.4	44.0	53.2	53.2	**56.3**	44.3
Recall, %	49.1	46.6	52.3	43.3	**58.0**	**58.0**	50.8	50.8
F1-score, %	46.1	47.1	52.6	43.3	**55.2**	**55.2**	53.2	46.8
Average time for answer, s	189	**29**	186	181	327	59	145	216

Table 2. Results of the diagnosis predictions carried out by human specialist using various machine learning methods

Parameter	AI system					
	Logistic Regression	K Nearest Neighbors	Support Vector Machine	Decision Tree	Random Forest	Gradient Boosting
True positive, TP	18	19	20	21	19	18
True negative, TN	63	57	53	47	47	55
False positive, FP	3	9	13	19	19	11
False negative, FN	12	11	10	9	11	12
Accuracy, %	**84.4**	79.2	76.0	70.8	68.8	76.0
Precision, %	**85.7**	67.9	60.6	52.5	50.0	62.1
Recall, %	60.0	63.3	**66.7**	70.0	63.3	60.0
F1-score, %	**70.6**	65.5	63.5	60.0	55.9	61.0
After 5-fold cross-validation						
Accuracy, %	69.7	70.1	68.8	71.0	**75.2**	74.7
Precision, %	72.3	74.4	77.1	**82.1**	80.3	81.1
Recall, %	**91.5**	86.9	78.4	75.0	85.0	83.0
F1-score, %	80.6	80.0	77.3	78.1	**82.2**	81.6

Key positions in the list (certain indices with high values were associated with diagnosis 2).

GPT o3-mini compared each set of symptoms with the available examples, assessing the structure and relative magnitudes of the features (e.g. high or low values at certain positions). If the distribution of values was closer to examples with diagnosis 1 or 2, the corresponding diagnosis was selected. This was an intuitive, heuristic approach without a formal model.

Claude 3.7 selected diagnoses by analyzing patterns in the distribution of symptom values based on several key principles: (1) high values in the middle of the vector (approximately from positions 3 to 15) corresponds to diagnosis 2 more frequently; (2) high values at the end of the vector (positions 20-29) correspond to diagnosis 1 more often; (3) if the last number (position 29) is large (>10), it usually corresponds to diagnosis 1; (4) as for the total sum of values in the symptoms vector, the model has found the dependency between “peaks” in the first half of the vector and the diagnosis 2; (5) when the number of symptoms is small (many zeros), the location of non-zero values becomes important. Qwen2.5-Max analyzed combinations of numbers in the symptom list. The decision was made based on comparison with examples from the knowledge base, symptom lists

were compared with existing samples where diagnoses were indicated. Key positions were also analyzed, i.e. certain indices in the list (for example, high values in the middle or at the end could indicate diagnosis 2), and diagnosis frequencies (if similar combinations corresponded to diagnosis 1 more often, it was chosen).

4 Conclusion

In general, existing publicly available AI systems are quite capable of performing medical diagnostic tasks, but to achieve these goals, they lack the “focus” on these tasks to ensure reliable answers to questions from a medical specialist. The disadvantages of publicly available AI systems are the inability to access them with large amounts of data. In addition, in some cases, the accuracy of prediction varies significantly from one call to another, as was the case with Grok-3. Perhaps the future use of AI for medical diagnostics requires the correct "focusing" of the system through additional prompts, which can be done with the help of special applications that carry out a preliminary dialogue with the system. In addition, it may make sense in the future to create specialized systems on a national scale that carry out medical consultations and diagnoses based on specially trained AI systems.

Acknowledgements

This work has been partially supported by the project SIP 20250738 of IPN.

References

1. **Fokkens, W.J., Lund, V.J., Hopkins, C, et al. (2020).** European Position Paper on Rhinosinusitis and Nasal Polyps 2020. Rhinology, 58 (Suppl S29), 1–464.
2. **Jankowski, R., Favier, V., Saroul, N., Lecanu, J.-B., Nguyen, D.T., de Gabory, L., Verillaud, B., Rumeau, C., Gallet, P., Béquignon, E., Vandersteen, C., Patron, V. (2023).** Critical review of diagnosis in rhinology and its therapeutical implications. European Annals of Otorhinolaryngology, Head and Neck Diseases, 140 (6), 271–278.
3. **Rimmer, J., Hellings, P., Lund, V. J., et al. (2019).** European position paper on diagnostic tools in rhinology. Rhinology, 57 (Suppl S28), 1–41.
4. **Nangia, S., Giridher, V., Chawla, P. (2019).** Evaluation of the Role of Nasal Endoscopy and Computed Tomography Individually in the Diagnosis of Chronic Rhinosinusitis. Indian J Otolaryngol Head Neck Surg, 71 (Suppl 3), 1711–1717.
5. **Koenraads, S.P.C, Aarts, M.C.J., van der Veen, E.L., Grolman, W., Stegeman, I. (2016).** Quality of life questionnaires in otorhinolaryngology: a systematic overview. Clin. Otolaryngol, 41, 681–688.
6. **Meghazi, H. M., Mostefaoui, S. A., Maaskri, M., Aklouf, Y. (2024).** Deep Learning-Based Text Classification to Improve Web Service Discovery. Computación y Sistemas, 28(2), 529–542.
7. https://x.ai/news/grok-3
8. https://ai.google.dev/gemini-api/docs/models
9. https://deepmind.google/technologies/gemini/pro/
10. https://deepmind.google/technologies/gemini/
11. https://github.com/deepseek-ai/DeepSeek-R1
12. https://openai.com/index/introducing-o3-and-o4-mini/
13. https://www.anthropic.com/news/claude-3-7-sonnet
14. https://qwenlm.github.io/blog/qwen2.5-max/

Article received on 06/04/2025; accepted on 28/07/2025.
**Corresponding author is Ildar Batyrshin.*