Testing ChatGPT's Performance on Medical Diagnostic Tasks

Alexander Perko*
Franz Wotawa*
alexander.perko@tugraz.at
wotawa@ist.tugraz.at
Graz University of Technology, Institute of Software Technology
Graz, Austria

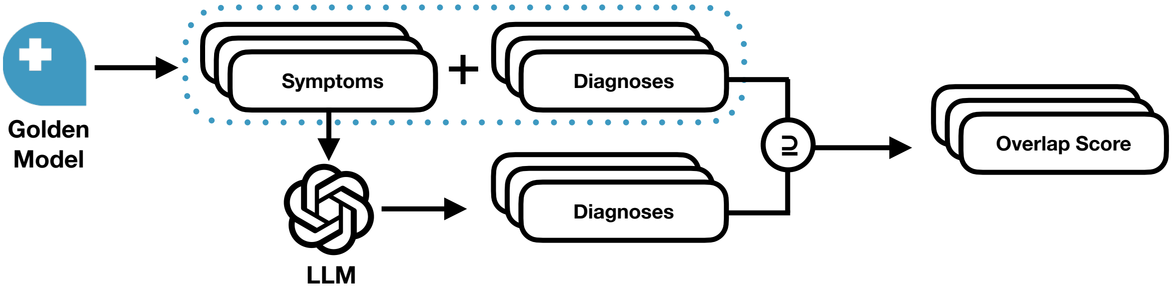


Figure 1: Semi-Automatic Evaluation of an LLM on Medical Diagnostics Using a Medical Expert System as a Golden Model.

Abstract

Large Language Models and chat interfaces like ChatGPT have become increasingly important recently, receiving a lot of attention even from the general public. People use these tools not only to summarize or translate text but also to answer questions, including medical ones. For the latter, giving reliable feedback is of utmost importance, which is hard to assess. Therefore, we focus on validating the feedback of ChatGPT and propose a testing procedure utilizing other medical sources to determine the quality of feedback for more straightforward medical diagnostic tasks. This paper outlines the problem, discusses available sources, and introduces the validation method. Moreover, we present the first results obtained when applying the testing framework to ChatGPT.

Keywords

Large Language Models, ChatGPT, NetDoktor, Testing, Validation

1 Introduction

Large Language Models (LLMs) are omnipresent in today's society, as they are used by a wide audience for a growing number of tasks. This study sheds light on one area of application in particular, which is asking for medical diagnoses. Assessing one's health and medical diagnostics are complex tasks, that fall into the domain of medical experts. However, since the dawn of search engines and medical websites, like NetDoktor [13], people have turned to the internet for getting health advice. Previously, users searching for answers had to consult multiple online resources, compare page contents, and evaluate whether their set of symptoms matched what they found. Nowadays it is seemingly easy to find answers in one spot as LLM-powered chatbots, like ChatGPT [8], are happy to respond with a diagnosis. This, of course, implies much risk of harm or misinterpretation. After all, the very reason many users - being non-experts - turn to chatbots is that they cannot assess symptoms themselves. Clusmann et al. [5] further point out that there is a lack of mechanisms to guarantee that the LLM's output is correct. All of this makes it important to test such systems on a practical level, which is close to the use cases of non-experts. As for its popularity, our evaluation focuses on ChatGPT [8], which is powered by OpenAI's most recent model, GPT-4o [9, 10]. The task of medical diagnostics shares many traits with the natural language processing (NLP) task of question answering (QA). Namely, this task tests for medical knowledge as well as basic reasoning facing medical language. MedQA [6] is a popular benchmark in literature, which is tailored to the medical domain. In recent years, open-domain LLMs such as GPT-3.5 [3], GPT-4 [9], and LLaMA-2 [16] as well as domain-specific LLMs like Med-PaLM 2 [15], Meditron [4] and Med-Gemini [14] have been evaluated on medical QA. The United States Medical Licensing Examination (USMLE) part of MedQA is used particularly often as a performance indicator in this domain. Table 1 shows reported scores of the mentioned LLMs and demonstrates GPT-4's and MedGemini's superiority, with GPT-4 performing marginally worse despite being an open-domain model.

Table 1: LLMs Evaluated on Medical Question Answering. Accuracy Results on the United States Medical Licensing Examination (USMLE) Part of MedQA [6], as Reported in [7, 14, 4, 15].

Model	Domain-Specific	MedQA USMLE
Med-Gemini	Yes	91.1
GPT-4	No	90.2
Med-PaLM 2	Yes	86.5
Meditron	Yes	75.8
LLaMA-2	No	63.8
GPT-3.5	No	60.2

*Both authors contributed equally to this research.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

Alongside ChatGPT's popularity, these results are a major reason why this paper focuses on GPT-4o in particular. This work contributes by introducing a semi-automated validation procedure for medical diagnostics performed with LLMs using an expert system as a golden model (compare to Figure 1). Specifically, we evaluate the performance of ChatGPT powered by GPT-4o with a focus on symptom descriptions in German and compare it to NetDoktor's Symptom-Checker [13], which is curated by medical professionals. Our setup is guided by the following questions regarding prompting ChatGPT:

- Does ChatGPT provide equivalent diagnoses when presented with the same symptoms as NetDoktor?
- Does the output quality - as measured by the overlap - change when asked for a specific amount of "most likely" diagnoses?
- Does the output increase in quality when ChatGPT is queried in English instead of German?

2 Validation Methodology

For the purpose of introducing our methodology, we use myocardial infarction (i.e. heart attack) as a guiding example. According to Statisik Austria's annual report, cardiovascular diseases, which include heart attacks, are the most common cause of death in Austria. The symptoms of a myocardial infarction include:

- Feeling of tightness or constriction
- Feeling of anxiety/panic attacks
- Sudden severe shortness of breath, unconsciousness, or severe dizziness
- Nausea and vomiting
- Blood pressure and pulse drop

These symptoms are now linked to an imaginary person's sex and age to form a persona whom for we want to retrieve diagnoses. Our exemplary set of symptoms shall be linked to an adult man and can be identified by ID 1 in all tables and plots. Besides this exemplary persona, where we first fixed a disease, all other sets of symptoms are picked at random. This can be done due to our assumption of a golden model, which we use as our baseline.

2.1 Golden Model

We use NetDoktor's "Symptom-Checker" [13] as a baseline for our evaluation. Symptom-Checker is a freely accessible, medical expert system for retrieving likely diagnoses corresponding to a person's symptoms. The system can be interacted with via a questionnaire but is only available in German. Parts of the questionnaire are static, such as questions regarding sex, age, and selecting the general area of one's body where symptoms occur most prominently, while others are adapting to the previously asked questions. The dynamically changing questions are always asked expecting an answer from the set: "Yes", "No" and "Skip". According to NetDoktor, the system is continuously validated by medical professionals and is based on the medical database AMBOSS [1] and follows the medical guidelines of professional societies [2]. We assume this expert system to be our golden model, as it comprises curated knowledge of high quality and is fully deterministic. The latter makes it possible, to generate a decision tree from a person's (or persona's) interaction with the system, that is reproducible across multiple calls [1]. Figure 2 shows the tree generated from the interaction of our exemplary persona having a heart attack. The tree is to be read from top to bottom, starting with the first question as the root node. It should be noted, that the very first question "Um wen geht es?" (i.e. "Who is it about?"), was always answered by "jemand anderen" (i.e. "somebody else") for this study. Rectangles represent questions and the ellipses represent the respective possible answers to choose from. The node at the second to last level, which is denoted by "Mögliche Erkrankungen" (i.e. "possible diseases") symbolizes the retrieval of diagnoses from the database, while the leaf nodes on the bottom level signify the results of the query. In this exemplary case, the questions were answered to correspond to the symptoms of a heart attack for demonstration. However, we can also use Symptom-Checker to automatically and randomly traverse the questionnaire's tree-like structure to retrieve sets of symptoms and corresponding diagnoses. This allows for a scaleable framework for comparing other methods against a strong and valid baseline. Sets of symptoms and corresponding "golden" diagnoses are extracted from such a tree as follows: Firstly, for each path from the root node to the bottom level nodes (i.e. the diagnoses), questions-answer-pairs are stored in a JSON data structure. Each full path represents one set of symptoms. Secondly, each set of symptoms is summarized in a textual representation in German taking special care not to lose or add information. This is then translated from German to English. The first rows of Tables 3 and 4 contain the textual descriptions of our example in German and English, respectively. Lastly, the diagnoses provided by the golden model are extracted from the bottom layer (i.e. the leaf nodes) of the tree, which is always a set of three diagnoses. These sets of diagnoses are referred to as NetDoktor diagnoses for the remainder of this paper.

2.2 Evaluation Metric

The main evaluation metric used in this work is the overlap of diagnoses as compared to NetDoktor. A set of diagnoses is considered as being good if it contains a large overlap with the golden model diagnoses of NetDoktor. Since the NetDoktor baseline always yields three diagnoses, the highest overlap any other system can achieve is 3/3. Thus, the score ranges from 0/3 to 3/3. We explicitly do not normalize, although we want to compare sets of diagnoses with varying cardinalities. The reason for this is that yielding more diagnoses should not be penalized (as they might be worth considering, as well), and yielding fewer should not lead to a better score automatically.

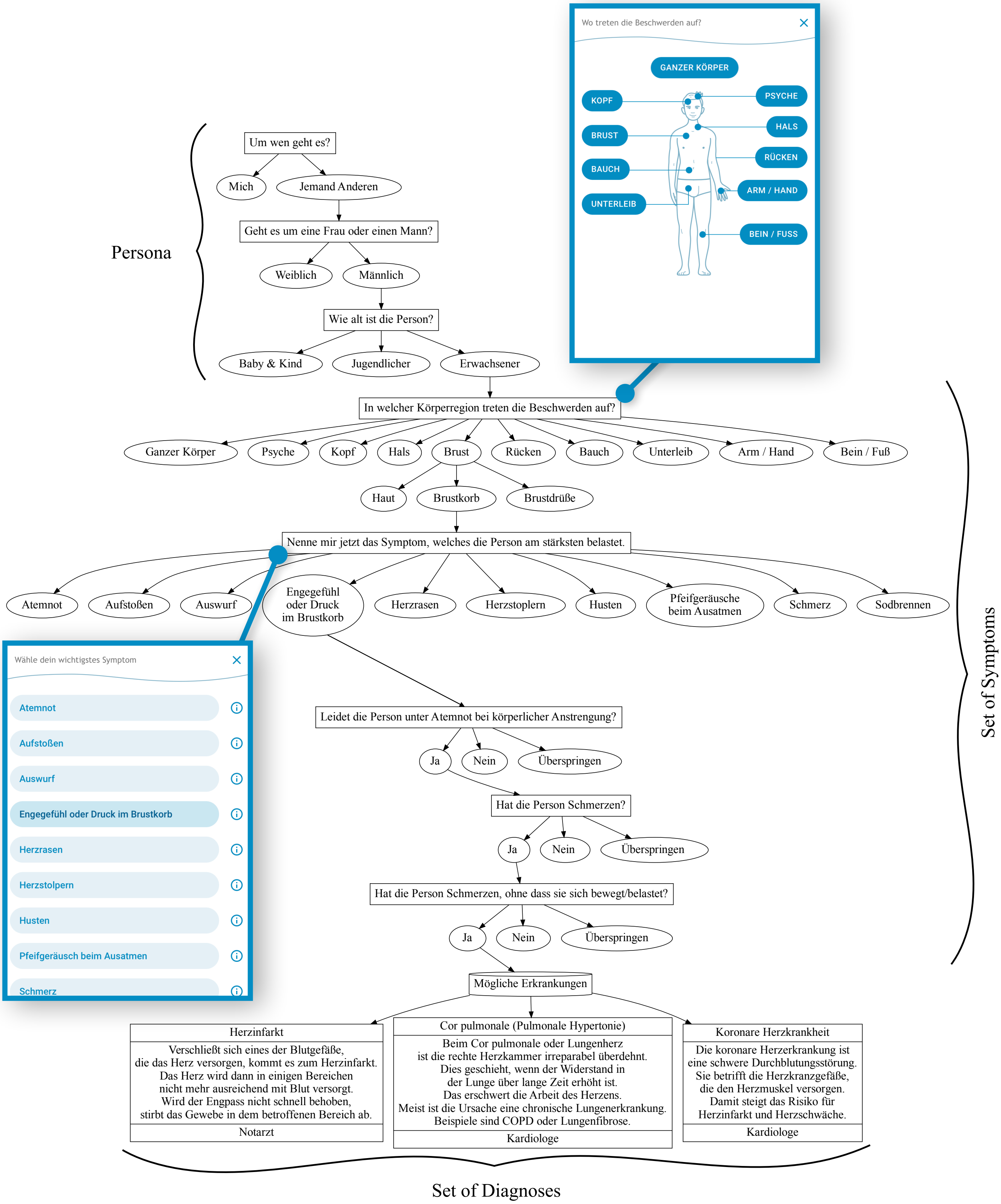
2.3 Equivalence of Diagnoses

This study compares systems designed for direct interaction with humans. These systems' output is presented to the users in natural language. A key feature of medical language is its interchangeable use of semantically equivalent terminology originating from different languages such as Latin, German, or English. Additionally, when talking to patients, medical personnel often have to use simplified terminology, which includes the use of colloquial synonyms, hypernyms, and hyponyms. Hence, the semantic equivalence of diagnoses must be considered to ensure the comparability of different systems.

- Synonyms are terms, which can be used interchangeably with one another.
- Hypernyms are superordinate or umbrella terms of a term.
- Hyponyms are describing subordinate terms (i.e. more specific) or another term.

Table 2 lists diagnoses that are treated as equivalents for this study.

[1] As long as the underlying knowledge base does not change.



Herzinfarkt

Verschließt sich eines der Blutgefäße, die das Herz versorgen, kommt es zum Herzinfarkt. Das Herz wird dann in einigen Bereichen nicht mehr ausreichend mit Blut versorgt. Wird der Engpass nicht schnell behoben, stirbt das Gewebe in dem betroffenen Bereich ab.

Notarzt

Cor pulmonale (Pulmonale Hypertonie)

Beim Cor pulmonale oder Lungenherz ist die rechte Herzkammer irreparabel überdehnt. Dies geschieht, wenn der Widerstand in der Lunge über lange Zeit erhöht ist. Das erschwert die Arbeit des Herzens. Meist ist die Ursache eine chronische Lungenerkrankung. Beispiele sind COPD oder Lungenfibrose.

Kardiologe

Koronare Herzkrankheit

Die koronare Herzerkrankung ist eine schwere Durchblutungsstörung. Sie betrifft die Herzkranzgefäße, die den Herzmuskel versorgen. Damit steigt das Risiko für Herzinfarkt und Herzschwäche.

Kardiologe

Figure 2: Golden Model: Exemplary Decision Tree Based on NetDoktor's Symptom-Checker Questionnaire [13] Filled-Out for a Persona Having a Heart Attack. Blue Boxes are Screenshots from Symptom-Checker Corresponding to Nodes in the Tree. We Set a Persona and Automatically Extract A) a Set of Symptoms and B) a Set of Diagnoses for Each Path From the Root Node to the Leaf Nodes on the Bottom-Most Level.

Table 2: Equivalent Diagnoses: Synonyms, Hypernyms, Hyponyms & Translations

Diagnosis	Equivalence (as Occurring in ChatGPT Output)
Herzinfarkt	Myokardinfarkt Akutes Koronarsyndrom Myocardial Infarction Heart Attack
Reiter-Syndrom	Reaktive Arthritis Morbus Reiter Reactive Arthritis Reiter's Syndrome
Kawasaki-Syndrom	Kawasaki Disease Kawasaki Syndrome
Blinddarmentzündung	Appendizitis
Vorhofflimmern	Herzrhythmusstörungen
Glutenunverträglichkeit	Zöliakie
Bakterielle Pharyngitis	Mild Bacterial Conjunctivitis with Pharyngitis
Krätze	Scabies
Erkältung	Virale Infekte
Pfeiffer-Drüsenfieber	Pfeiffersches Drüsenfieber Mononukleose
Blasenentzündung	Zystitis Harnwegsinfektion Urinary tract infection
Gürtelrose	Herpes Zoster
Mastopathie	Fibrozystische Mastopathie
Lipom	Lipoma

2.4 Sets of Symptoms & Personas

For this evaluation, we retrieved 12 sets of symptoms from NetDoktor - 6 for females and 6 for males, and for each sex, we used all of NetDoktor's 3 age categories (baby/child, adolescent, adult) twice. In addition, we used the exemplary set of symptoms for an adult man having a heart attack, as discussed in the previous section. This yields the 13 sets of symptoms listed in Tables 3 and 4. In Figure 2, the parts of the questionnaire are marked, which correspond to the persona and the set of symptoms respectively. In the following, both terms are used interchangeably.

2.5 Model, Prompts & Diagnose Retrieval

For all of our experiments, we used GPT-4o [9, 10] through ChatGPT [8]. More specifically, we used version GPT-4o-2024-08-06, which has been released in August 2024. We evaluate the same model in German and English and denote this with a trailing "[DE]" for German and "[EN]" for English for the respective results. We extended this convention to our golden model NetDoktor as well. The full list of prompts used can be found in the next section, Section 3. All LLM results were retrieved in a zero-shot methodology, without samples or additional context besides the prompt itself. Every symptom description is sent within a new chat to isolate individual queries. However, we cannot guarantee that we are indeed interacting with a "blank slate" as ChatGPT and GPT-4o are both black boxes and our user profile might interfere with the output.

Table 3: Sets of Symptoms per ID [DE]

ID	Description of Symptoms in German
1	Ein erwachsener Mann verspürt ein Engegefühl im Brustkorb. Er hat Schmerzen, auch wenn er sich nicht bewegt oder belastet. Außerdem leidet er unter Atemnot, wenn er sich anstrengt.
2	Ein Bub hat gerötete Augen und Fieber. Außerdem hat er Schmerzen beim Wasserlassen.
3	Ein kleiner Junge hat starke Bauchschmerzen, die bei Druck schlimmer werden. Es wurde auch festgestellt, dass er allgemein druckempfindlich ist.
4	Ein jugendlicher Junge verspürt ein Engegefühl oder Druck im Brustkorb. Er bemerkt, dass sein Herz sehr schnell schlägt und unregelmäßig arbeitet. Er fühlt sich oft müde und weniger leistungsfähig.
5	Ein jugendlicher Junge hat an Gewicht verloren und leidet unter anhaltender Müdigkeit. Er hat regelmäßig Durchfall, der besonders voluminös und übelriechend ist.
6	Ein erwachsener Mann hat eine Rachenentzündung und bemerkt gerötete Augen. Es gibt jedoch kein Fieber oder geschwollene Lymphknoten.
7	Ein erwachsener Mann hat entzündliche Hautveränderungen am Unterschenkel, die stark jucken, insbesondere nachts. Es wurde kein Zusammenhang mit Allergien festgestellt.
8	Ein kleines Mädchen hat seit einiger Zeit ihren Appetit verloren, fühlt sich ungewöhnlich müde und hat ungewollt an Gewicht verloren. Es wird auch über verminderten Urinfluss berichtet.
9	Ein kleines Mädchen hat Fieber, eine Rachenentzündung und geschwollene Lymphknoten. Sie fühlt sich abgeschlagen und schwitzt besonders nachts stark.
10	Eine jugendliche Mädchen hat Blut im Urin und Schmerzen beim Wasserlassen. Der Harndrang ist häufig, aber es wird nur eine geringe Urinmenge ausgeschieden. Zudem verspürt sie ein Brennen beim Wasserlassen.
11	Ein jugendliches Mädchen klagt über ausstrahlende Schmerzen im Nackenbereich und hat einen Hautausschlag mit kleinen Bläschen.
12	Eine erwachsene Frau hat Spannungsgefühle in der Brust und tastet schmerzlose Knoten. Die Haut ist nicht gerötet.
13	Eine erwachsene Frau verspürt Druckempfindlichkeit im Oberbauch, die Haut wölbt sich vor und die Region ist geschwollen.

3 Experimental Evaluation

Figure 1 depicts our experimental setup: NetDoktor is used as a golden model to automatically derive sets of symptoms and corresponding diagnoses as exemplified in Figure 2. The extracted symptoms are then used as input to the LLM GPT-4o via ChatGPT. ChatGPT diagnoses are then compared to NetDoktor diagnoses to compute an overlap score. Figure 4 gives an overview of our evaluation results. For each set of symptoms, NetDoktor results are shown, followed by four diagnosis strategies utilizing ChatGPT. The grey bars denote the cardinality of every resulting set of diagnoses. Blue overlays are used to show the overlap between NetDoktor diagnoses and ChatGPT diagnoses. These overlays correspond to the values in Table 5, which comprises the occurrences of overlaps in each category from 0/3 to 3/3. In addition to the 13 sets of symptoms, Figure 4 and Table 5 include averages computed over all sets for easier comparison of the prompts/-diagnosis retrieval methods. In the following, you can find the used prompts/methodologies corresponding to the depicted bars:

A **NetDoktor [DE]:** Diagnoses from NetDoktor were retrieved via the Symptom-Checker questionnaire as is documented in Subsection 2.1. This is our golden model and

Table 4: Sets of Symptoms per ID [EN]

ID	Description of Symptoms in English
1	An adult man feels a tightness in his chest. He experiences pain even when he is not moving or exerting himself. Additionally, he suffers from shortness of breath when he exerts himself.
2	A boy has red eyes and a fever. He also has pain when urinating.
3	A little boy has severe abdominal pain, which worsens with pressure. It was also found that he is generally sensitive to pressure.
4	A teenage boy feels a tightness or pressure in his chest. He notices that his heart beats very fast and irregularly. He often feels tired and less capable.
5	A teenage boy has lost weight and suffers from persistent fatigue. He has regular diarrhea that is particularly voluminous and foul-smelling.
6	An adult man has a throat infection and notices red eyes. However, there is no fever or swollen lymph nodes.
7	An adult man has inflammatory skin changes on his lower leg that itch intensely, especially at night. No connection with allergies was found.
8	A little girl has lost her appetite for some time, feels unusually tired, and has unintentionally lost weight. Reduced urine output is also reported.
9	A little girl has a fever, a throat infection, and swollen lymph nodes. She feels weak and sweats heavily, especially at night.
10	A teenage girl has blood in her urine and pain when urinating. The urge to urinate is frequent, but only a small amount of urine is passed. She also feels a burning sensation when urinating.
11	A teenage girl complains of radiating pain in the neck area and has a rash with small blisters.
12	An adult woman has a feeling of tension in her breast and can feel painless lumps. The skin is not reddened.
13	An adult woman feels tenderness in the upper abdomen, the skin bulges, and the area is swollen.

overlaps with its diagnoses are marked in blue, in Figure 4. The questionnaire and results are in German.

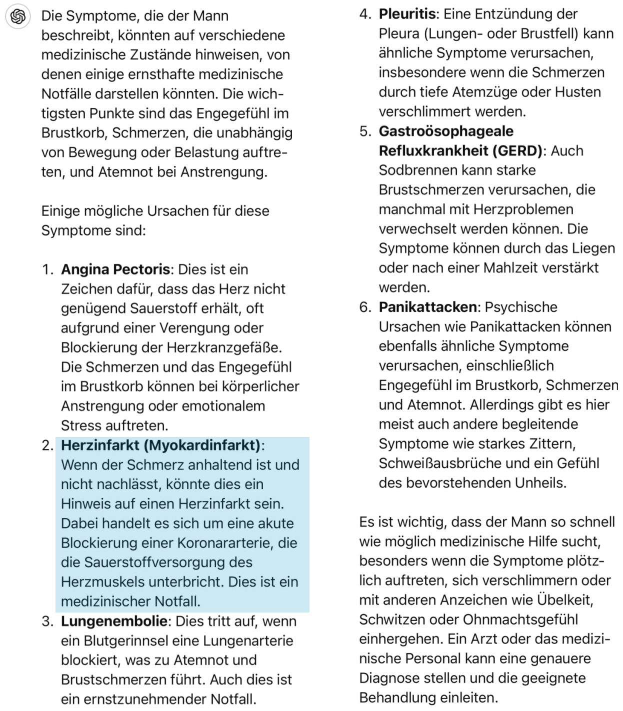
B **ChatGPT [DE]:** Ad-hoc query sent to ChatGPT using the symptom descriptions in German from Table 3 as is.

C **ChatGPT [DE] "3 Most Likely":** More elaborate query sent to ChatGPT using the symptom descriptions in German from Table 3, additionally requesting the "3 most likely" diagnoses.

D **ChatGPT [DE] "10 Most Likely":** More elaborate query sent to ChatGPT using the symptom descriptions in German from Table 3, additionally requesting the "10 most likely" diagnoses.

E **ChatGPT [EN]:** Ad-hoc query sent to ChatGPT using the symptom descriptions in English from Table 4 as is.

Out of convenience, the letters introduced in this list are used when referring to a specific prompt in the following paragraphs. The main takeaway from this evaluation is that none of the used prompts achieves a complete overlap of 3/3 with NetDoktor for any of the personas. Prompt B, achieves the highest score, with 0.92/3 i.e. 31%. For our small test set of 13 sets of symptoms, these results constitute from 2/3 for two, and 1/3 for seven sets of symptoms. The same prompting strategy in English, denoted by E, yields worse results, having overlaps of 2/3 for one, and 1/3 for three sets of symptoms. The two prompts giving ChatGPT the task of answering with the "n most likely" diagnoses are equally not performing as well as the simple prompt in German: C achieves 1/3 for 9 sets of symptoms and D yields 2/3 for three and 1/3 for five sets of symptoms. This means that prompt D, asking for the "10 most likely" diagnoses is the runner-up with an average of 0.85/3 i.e. 28%. Surprisingly, the simple prompt in English, E, performs poorest, which contradicts our hypothesis of English prompts performing better.



Die Symptome, die der Mann beschreibt, könnten auf verschiedene medizinische Zustände hinweisen, von denen einige ernsthafte medizinische Notfälle darstellen könnten. Die wichtigsten Punkte sind das Engegefühl im Brustkorb, Schmerzen, die unabhängig von Bewegung oder Belastung auftreten, und Atemnot bei Anstrengung.

Einige mögliche Ursachen für diese Symptome sind:

1. **Angina Pectoris**: Dies ist ein Zeichen dafür, dass das Herz nicht genügend Sauerstoff erhält, oft aufgrund einer Verengung oder Blockierung der Herzkranzgefäße. Die Schmerzen und das Engegefühl im Brustkorb können bei körperlicher Anstrengung oder emotionalem Stress auftreten.
2. **Herzinfarkt (Myokardinfarkt)**: Wenn der Schmerz anhaltend ist und nicht nachlässt, könnte dies ein Hinweis auf einen Herzinfarkt sein. Dabei handelt es sich um eine akute Blockierung einer Koronararterie, die die Sauerstoffversorgung des Herzmuskels unterbricht. Dies ist ein medizinischer Notfall.
3. **Lungenembolie**: Dies tritt auf, wenn ein Blutgerinnsel eine Lungenarterie blockiert, was zu Atemnot und Brustschmerzen führt. Auch dies ist ein ernstzunehmender Notfall.
4. **Pleuritis**: Eine Entzündung der Pleura (Lungen- oder Brustfell) kann ähnliche Symptome verursachen, insbesondere wenn die Schmerzen durch tiefe Atemzüge oder Husten verschlimmert werden.
5. **Gastroösophageale Refluxkrankheit (GERD)**: Auch Sodbrennen kann starke Brustschmerzen verursachen, die manchmal mit Herzproblemen verwechselt werden können. Die Symptome können durch das Liegen oder nach einer Mahlzeit verstärkt werden.
6. **Panikattacken**: Psychische Ursachen wie Panikattacken können ebenfalls ähnliche Symptome verursachen, einschließlich Engegefühl im Brustkorb, Schmerzen und Atemnot. Allerdings gibt es hier meist auch andere begleitende Symptome wie starkes Zittern, Schweißausbrüche und ein Gefühl des bevorstehenden Unheils.

Es ist wichtig, dass der Mann so schnell wie möglich medizinische Hilfe sucht, besonders wenn die Symptome plötzlich auftreten, sich verschlimmern oder mit anderen Anzeichen wie Übelkeit, Schwitzen oder Ohnmachtsgefühl einhergehen. Ein Arzt oder das medizinische Personal kann eine genauere Diagnose stellen und die geeignete Behandlung einleiten.

Figure 3: ChatGPT Output for ID 1, Using Prompt B. Overlap with NetDoktor Marked in Blue. Compare to Figure 2.

Table 5: Overlaps of Diagnoses with NetDoktor per Prompt

Score	Diagnosis Retrieval Method				
	A	B	C	D	E
3/3	13	0	0	0	0
2/3	0	2	0	3	1
1/3	0	8	9	5	3
0/3	0	3	4	5	9
Avg.	3/3	0.92/3	0.69/3	0.85/3	0.38/3
Avg.[%]	100%	31%	23%	28%	13%

Apart from the overlaps, other interesting observations can be made on closer inspection of the results: ChatGPT seems to rigorously follow the instruction to generate n diagnoses and as such, yields consistently 3 diagnoses for prompt C and 10 diagnoses for prompt D. However, it can be doubted that "most likely" is interpreted in a scientifically backed manner, as ChatGPT often does not include even one of the NetDoktor diagnoses and not once all of them. Equally interesting is the inclusion of the necessity to consult a doctor in one form or the other at the end of every result we received, which is likely due to being "hard-coded" for legal reasons on the part of OpenAI. This can also be seen in Figure 3. Although ChatGPT and GPT-4o are black boxes and LLMs are non-deterministic, we try to document our reported results as well as possible for replication. You can find all of our

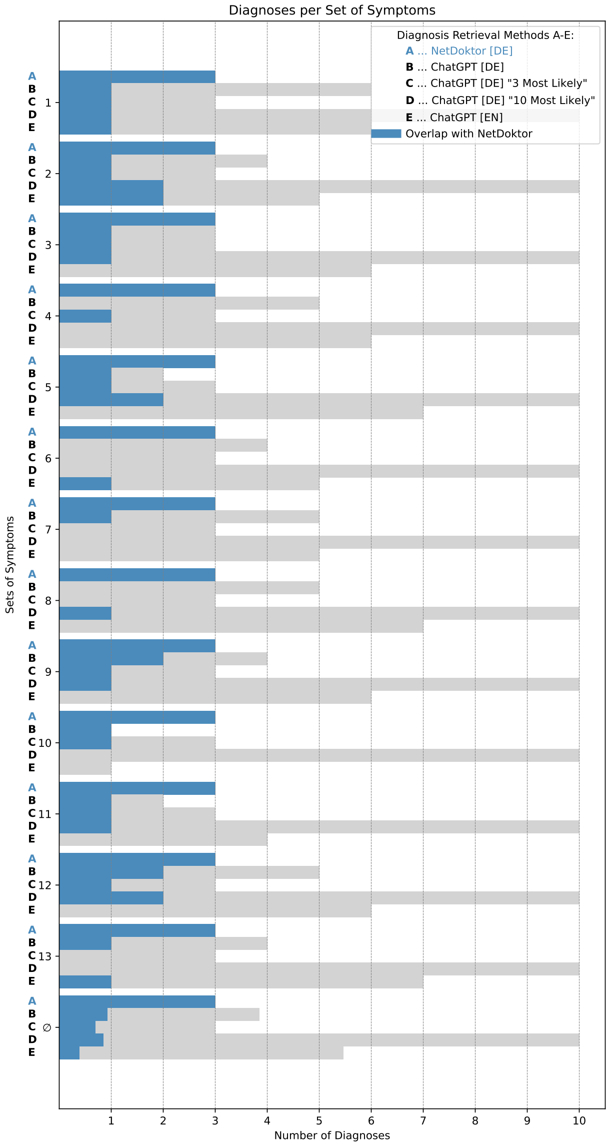


Figure 4: Comparison of Diagnoses for Symptoms Seen in Table 3

experimental results as a replication package under the provided URL [2].

4 Conclusions

In summary, ChatGPT diagnoses only partially match the diagnoses retrieved from our golden model NetDoktor. ChatGPT diagnoses are mostly well-structured and are seemingly valid but fail to include all NetDoktor diagnoses in any of the tested cases. This holds for all tested prompts and across all tested symptoms. The highest overlap, on average, could be achieved with the simplest prompt in German, giving only a description of the symptoms for a persona. When asked for a specific amount of "most likely" diagnoses, ChatGPT always delivered exactly the asked-for number of diagnoses. However, this does not benefit the quality of the output as measured by the overlap metric. Neither does an interaction in English change the output quality for the better. In our tests, ChatGPT always includes a notice to consult a doctor. Human assessment of the diagnoses cannot be fully bypassed by the proposed evaluation methodology. This is due to the immanent presence of semantic equivalence and the necessary medical knowledge to find those equivalences. Although such a task is automatable via LLMs as well, the authors of this paper underline the potential implications for undermining the quality of an evaluation, when fully automated. While our evaluation reports results achieved using ChatGPT and GPT-4o, the proposed methodologies transcend to other LLMs as well. As part of future work, we want to repeat our experiments at a larger scale to achieve representative results. Additionally, we want to consider stability metrics, as seen in [11]. Another interesting direction can be further analysis of the relationship between prompt (engineering) and the retrieval of matching diagnoses as well as their stability. Finally, it would be interesting to compile a corpus of medical symptoms corresponding to diagnoses including named entities and logical abstractions to perform evaluations as seen in [12] on the medical domain.

Acknowledgements

The work presented in this paper was partially funded in part by the European Union under Grand 101159214 – ChatMED. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] AMBOSS GmbH. 2024. Amboss. https://www.amboss.com. Accessed: 2024-09-03. (2024).

[2] AWMF. 2024. Arbeitsgemeinschaft der wissenschaftlichen medizinischen fachgesellschaften (awmf) - leitlinien. https://www.awmf.org/leitlinien. Accessed: 2024-09-03. (2024).

[3] Tom B. Brown et al. 2020. Language models are few-shot learners. arXiv: 2005.14165 [cs.CL]. (2020).

[4] Zeming Chen et al. 2023. Meditron-70b: scaling medical pretraining for large language models. (2023). https://arxiv.org/abs/2311.16079 arXiv: 2311.16079 [cs.CL].

[5] J. Clusmann et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3, 141. DOI: https://doi.org/10.1038/s43856-023-00370-1.

[6] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

[7] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. (2023). https://arxiv.org/abs/2303.13375 arXiv: 2303.13375 [cs.CL].

[8] OpenAI. 2023. ChatGPT. (2023). chat.openai.com/chat.

[9] OpenAI. 2023. GPT-4 technical report. arXiv: 2303.08774 [cs.CL]. (2023).

[10] OpenAI. 2024. Introducing gpt-4o and more tools to chatgpt free users. (May 2024). https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.

[11] Alexander Perko and Franz Wotawa. 2024. Evaluating openai large language models for generating logical abstractions of technical requirements documents. In *Proceedings to the 24th International Conference on Software Quality, Reliability, and Security (QRS)*. IEEE.

[12] Alexander Perko, Haoran Zhao, and Franz Wotawa. 2023. Optimizing named entity recognition for improving logical formulae abstraction from technical requirements documents. In *Proceedings to the 10th International Conference on Dependable Systems and Their Applications*.

[13] Jens Richter, Hans-Richard Demel, Florian Tiefenböck, Luise Heine, and Martina Feichter. 2024. Symptom-checker. https://www.netdoktor.at/symptom-checker/. Accessed: 2024-09-03. (2024).

[14] Khaled Saab et al. 2024. Capabilities of gemini models in medicine. (2024). https://arxiv.org/abs/2404.18416 arXiv: 2404.18416 [cs.AI].

[15] Karan Singhal et al. 2023. Towards expert-level medical question answering with large language models. (2023). https://arxiv.org/abs/2305.09617 arXiv: 2305.09617 [cs.CL].

[16] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. *ArXiv*.

[2] https://zenodo.org/doi/10.5281/zenodo.13765345