

## Article

# From Language Models to Medical Diagnoses: Assessing the Potential of GPT-4 and GPT-3.5-Turbo in Digital Health

Jonas Roos <sup>1,\*</sup> , Theresa Isabelle Wilhelm <sup>2,3</sup> , Ron Martin <sup>4</sup>  and Robert Kaczmarczyk <sup>5</sup> 

- <sup>1</sup> Department of Orthopedics and Trauma Surgery, University Hospital of Bonn, 53127 Bonn, Germany  
<sup>2</sup> Eye Center, Faculty of Medicine, Albert-Ludwigs-University of Freiburg, 79106 Freiburg, Germany  
<sup>3</sup> Medical Graduate Center, School of Medicine, Technical University of Munich, 80337 Munich, Germany  
<sup>4</sup> Department of Plastic and Hand Surgery, Burn Care Center, BG Clinic Bergmannstrost, 06112 Halle (Saale), Germany  
<sup>5</sup> Department of Dermatology and Allergy, School of Medicine, Technical University of Munich, 80337 Munich, Germany  
\* Correspondence: jonas.roos@ukbonn.de; Tel.: +49-22828714170

**Abstract:** Background: Large language models (LLMs) like GPT-3.5-Turbo and GPT-4 show potential to transform medical diagnostics through their linguistic and analytical capabilities. This study evaluates their diagnostic proficiency using English and German medical examination datasets. Methods: We analyzed 452 English and 637 German medical examination questions using GPT models. Performance metrics included broad and exact accuracy rates for primary and three-model generated guesses, with an analysis of performance against varying question difficulties based on student accuracy rates. Results: GPT-4 demonstrated superior performance, achieving up to 95.4% accuracy when considering approximate similarity in English datasets. While GPT-3.5-Turbo showed better results in English, GPT-4 maintained consistent performance across both languages. Question difficulty was correlated with diagnostic accuracy, particularly in German datasets. Conclusions: The study demonstrates GPT-4's significant diagnostic capabilities and cross-linguistic flexibility, suggesting potential for clinical applications. However, further validation and ethical consideration are necessary before widespread implementation.

**Keywords:** AI; LLM; medical examination; ChatGPT



**Citation:** Roos, J.; Wilhelm, T.I.; Martin, R.; Kaczmarczyk, R. From Language Models to Medical Diagnoses: Assessing the Potential of GPT-4 and GPT-3.5-Turbo in Digital Health. *AI* **2024**, *5*, 2680–2692. <https://doi.org/10.3390/ai5040128>

Received: 4 October 2024  
Revised: 15 November 2024  
Accepted: 26 November 2024  
Published: 2 December 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Research in Context

### 1.1. Evidence Before This Study

We conducted a comprehensive review of the literature using PubMed, Embase, and the Cochrane Central Register up to 18 August 2023. Our search combined terms related to artificial intelligence (e.g., “language models”, “GPT”, “natural language processing”) with those indicating medical diagnosis (e.g., “medical diagnosis”, “clinical prediction”, “medical consultation”). We also limited our search to studies published in English and German. Our findings primarily consisted of studies emphasizing the potential of AI models in various medical applications. However, few detailed evaluations existed regarding the specific diagnostic capability of large language models (LLMs) like GPT-3.5-Turbo and GPT-4 in multiple languages.

### 1.2. Added Value of This Study

This manuscript presents, to our knowledge, the most comprehensive assessment of the GPT-3.5-Turbo and GPT-4 models in terms of diagnostic accuracy across the English and German languages. This study not only emphasizes the models' remarkable capability in accurate diagnosis predictions but also underscores their performance consistency across diverse languages, suggesting their broader applicability. Moreover, our examination

includes a comparison of the models' performances across different levels of question difficulty, providing nuanced insights into their potential limitations and strengths.

### 1.3. Implications of All the Available Evidence

The findings of this study reinforce the idea that AI-driven LLMs like GPT-4 possess significant potential as diagnostic tools, especially in scenarios demanding swift and accurate medical insights. Their consistent performance across languages indicates a promising avenue for global healthcare applications, particularly in regions facing a shortage of healthcare professionals. Nevertheless, as our research suggests, while the models show potential, their actual clinical integration requires further rigorous validation. Future research could focus on integrating these models into clinical systems, understanding their ethical implications, and tailoring them further based on real-world feedback.

## 2. Introduction

Large language models (LLMs) are language-based artificial intelligences (AIs) that potentially find application in education, research, and clinical practice [1]. The challenge indeed lies in recognizing pitfalls and promoting feasibility in application [2]. There are currently no standards for the use of LLMs in medicine [3]. One of the most well-known frameworks to use large language models is ChatGPT from OpenAI [4]. Currently, two models are available in ChatGPT [5]. The training datasets for GPT 3 consisted of 93% English language content [6]. In comparison to GPT-3.5-Turbo, the parameter size, a value describing the model size, is about six times larger in GPT-4. GPT-4 is thus even more capable and able to handle more demanding scenarios [7]. The most recent version of Microsoft's search engine Bing Chat is powered by GPT-4, merging extensive language comprehension abilities with real-time access to recent data available on the web [8].

ChatGPT has already shown good results on the United States Medical Licensing Exam (USMLE), reaching the passing mark of 60% [9]. ChatGPT also showed good to outstanding results in the German medical state examination, with the number of correct answers significantly differing between the model versions GPT-3.5-Turbo and GPT-4 [10]. This was also confirmed in further examinations in medical tests [11,12]. In contrast to the reality in clinics, the responses to patient complaints in the German Medical Licensing Examination and the United States Medical Licensing Exam are predetermined by multiple-choice answers [13,14]. This limits the response options and essentially does not reflect clinical daily life. So far, initial studies have shown promising results regarding the performance of LLMs in addressing open medical questions [15,16]. To further explore the limits of these models and identify areas for improvement, additional research is essential. If these artificial intelligences achieve similarly good results when responding to open questions that require correct diagnoses based on medical history, symptom descriptions, laboratory values, and imaging findings, they could potentially be more widely implemented in clinical settings in the future.

With platforms like TrueHealth [17], Google Health [18], and Ada [19] advocating for the democratization of healthcare—a term that remains contentious or undefined in this context [20,21]—the role of LLMs becomes increasingly significant. The urgent need for meticulous scientific scrutiny is amplified by the looming risks, notably misdiagnoses. The allure of “free assessment and treatment plans” emphasizes the pressing need to evaluate the capabilities and constraints of such AI-driven tools, prioritizing patient safety.

A crucial tool in the medical text mining toolbox is named entity recognition (NER), which emerged as a substantial approach to transforming natural language texts into structured machine-readable data, thereby automating the extraction of necessary information. Initially introduced in 1996 to identify various types of names and symbols [22], it has evolved to identify specific medical entities such as diseases, drugs, and treatments, facilitating informed medical decision-making and disease risk prediction. In recent years, advances in model architecture continuously improved the accuracy for biomedical entity linking [23,24].

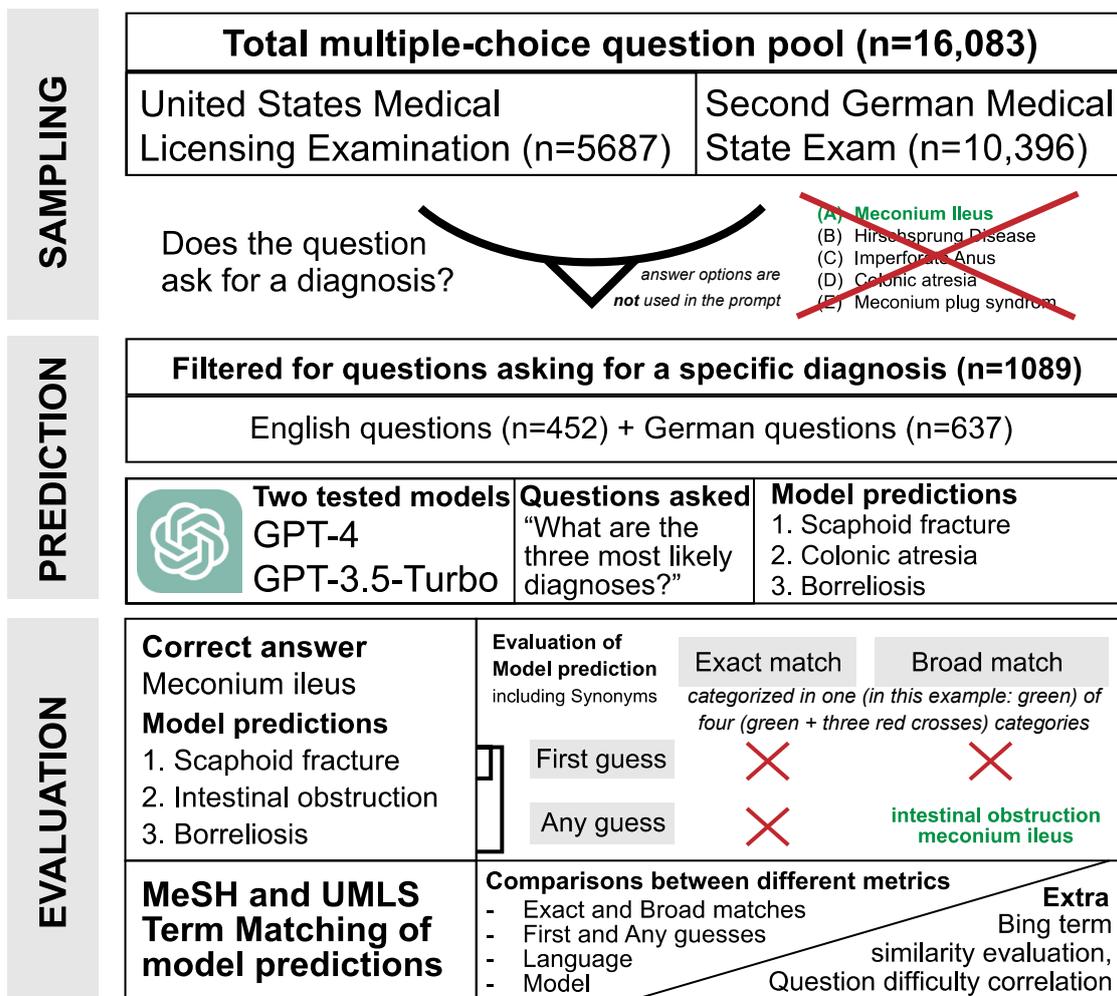
This study critically evaluates the capabilities of ChatGPT in medical education and clinical routine by analyzing the performance of GPT-4 and GPT-3.5-Turbo in answering open-ended medical questions from the German Medical Licensing Examination and the United States Medical Licensing Exam. Specifically, we examine the number of correct answers given by these large language models to both English and German questions—the latter not being their main training language—to assess their utility in the medical field.

In addition, we tested how well Bing Chat recognizes the answers determined by GPT-3.5-Turbo and GPT-4 as correct (exact match) or correct in the broader sense (broad match) compared to the human-created databases MeSH (Medical Subject Headings) and UMLS (Unified Medical Language System).

### 3. Methods

#### 3.1. Study Design

We conducted a comparative analysis of OpenAI’s large language models GPT-4 and GPT-3.5-Turbo (version of 24 May 2023, [25]), assessing their aptitude in predicting diagnoses based on exam questions in English and German, omitting the multiple-choice answers. MeSH and UMLS databases facilitated alignment between the correct answers and model responses. Furthermore, Bing Chat, with internet access during interactions, was employed for assessing the medical equivalency of terms (Figure 1).



**Figure 1.** Study design describing the sampling of the questions, prediction of diagnoses, and final evaluation after MeSH/UMLS term matching. The example shows a match within the any of the three model suggestions (“Any guess”) or in the broader sense (“Broad match”); for example, intestinal obstruction is not an exact match nor a synonym of meconium ileus but a broader concept.

### 3.2. Sampling of Exam Questions

Questions were sourced from the Amboss learning platform [26]. We manually curated multiple-choice questions seeking a diagnosis, specifically those with prompts like “What is the most likely diagnosis?”. The German subset comprised questions from the second clinical medical state exams spanning 2006–2023, while the English subset was derived from USMLE state exam preparation questions. Questions incorporating media (e.g., images) or having a series of queries for a single patient were excluded. From an initial pool of 16,083 questions, our final dataset included 637 (6.1%) German and 452 (7.9%) English questions, culminating in a sample size of 1089 questions. Depending on the percentage of correct student answers in the exam, the difficulty of each question was rated on a scale from 1 to 5 (Supplementary Table S1).

### 3.3. Prompting

The following standardized prompting sentences were used in OpenAI’s large language models GPT-4 and GPT-3.5-Turbo:

- English: “You are a medical expert (professor at a prominent US university hospital) tasked with determining the most likely diagnosis based on patient histories. Answer only in this format: <List of 3 diagnoses, the most probable first>”.
- German: “Du bist ein Medizinexperte (Professor an einer renommierten medizinischen Universitätsklinik) und möchtest die wahrscheinlichste Diagnose basierend auf Patientenanamnese bestimmen. Antworte nur in diesem Format: <Liste von 3 Diagnosen, die wahrscheinlichste zuerst>”.

### 3.4. Concept Detection and Diagnosis Matching

A total of 7623 medical terms were analyzed, with 1089 correct diagnoses and 6534 model-generated responses (three diagnostic suggestions per question from GPT-3.5-Turbo and GPT-4). In the next step, these responses/answer terms were matched with diagnoses (exact and broad matches) using the National Library of Medicine’s MeSH and UMLS Metathesaurus Browser. For this purpose, the German model-generated responses were translated into English via Google Translate. In cases of non-matching, we divided the responses into substrings (based on common patterns, Supplementary Table S2) and conducted a second search. The queries’ results were split into “exact matches” (exact string matches, abbreviations, and synonyms, e.g., atopic eczema and atopic dermatitis) and “broad matches” (e.g., beta-Thalassemia and alpha-Thalassemia, both terms located within the group of hemic and lymphatic diseases).

### 3.5. Bing Search Term Similarity Evaluation

To compare the AI-based matching of medical terms to the previously described matching using human-generated medical databases, we used Bing Chat to evaluate pairs of terms for synonymity or relatedness in a broader sense, employing a structured question template (Supplementary Table S3). If the models did not respond, we posed the question again. The Bing search mapping was then compared to the MeSH and UMLS mapping.

## 4. Statistical Analysis

Analyses were conducted on an Apple M1 Pro (16 GB) running macOS Ventura 13.4.1 (c) using Python 3.10.12. Libraries used include scipy (v1.11.1), Seaborn (v0.11.2), Matplotlib (v3.7.2), Pandas (v2.0.3), Numpy (v1.25.2), and Statannotations (v0.5.0).

We employed chi-squared tests to discern differences between groups with binary outcomes. Spearman’s rank correlation was utilized to associate question difficulty (Supplementary Table S4) with the accuracy of the LLM predictions. Graphical representations include 95% confidence intervals of the mean, with gray brackets denoting  $p$ -values  $\geq 0.5$ .

### *Declaration of Generative AI and AI-Assisted Technologies in the Writing Process*

While drafting this manuscript, we availed the services of Grammarly and GPT-4 to enhance linguistic clarity and rectify grammatical discrepancies. Thereafter, rigorous manual reviews were undertaken, and the authors assume complete responsibility for the content's integrity.

## **5. Results**

A comprehensive assessment of GPT models' accuracy was conducted in two prominent languages, namely English and German. The dataset for the English language comprised 452 questions, while the German dataset included 637 questions.

### *5.1. Matched Concepts*

Of all 7623 terms (correct exam answers + models' answers), 3749 remained after removing duplicates. We found 2727 (72.8%) matching concepts in the MeSH database and 3746 (99.9%) in the UMLS database. A total of 2724 (72.7%) matching concepts were found in both the UMLS and the MeSH databases. All 3749 (100%) terms, including synonyms, broader, and narrower concepts, were found in at least one of the two databases, thereby ensuring the complete set of 7623 terms was available for subsequent analyses.

### *5.2. Correct Diagnosis in LLM Predictions*

In the English dataset, when measuring broad accuracy, GPT-3.5-Turbo achieved a 75.0% accuracy rate on its first guess and 90.5% when any of the three predictions were considered. Meanwhile, GPT-4 demonstrated enhanced performance, with an 81.9% accuracy on the first guess and 95.4% when considering any prediction. When evaluating exact accuracy, GPT-3.5-Turbo and GPT-4 displayed 46.0% vs. 61.9% on the first guess and 61.1% vs. 76.8% on any of three predictions, respectively. For the German dataset, under broad accuracy metrics, GPT-3.5-Turbo displayed a 69.2% accuracy on its primary guess and an 82.6% accuracy on any of three predictions. In contrast, GPT-4 outperformed, with 79.0% on its initial guess and 90.0% on any guess. Examining the exact accuracy, GPT-3.5-Turbo achieved 47.4% on the first guess and 63.0% on any prediction, while GPT-4 reported a slight improvement, with 64.8% on its first guess and 76.6% considering any prediction.

### *5.3. Statistical Comparisons Between LLMs*

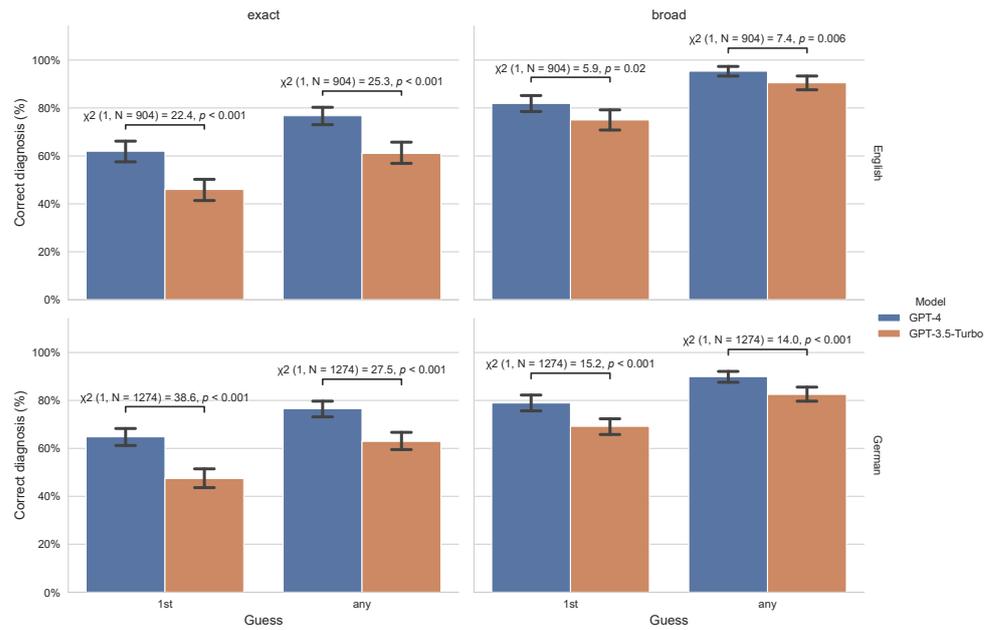
Upon a detailed examination of the exact and broad metrics for both languages, GPT-4 consistently demonstrated significant superiority over GPT-3.5-Turbo. For instance, in the exact metrics for the English language, any guess by GPT-4 was significantly better than GPT-3.5-Turbo ( $\chi^2(1, N = 904) = 25.3, p < 0.001$ ). Similar trends were observed across all analyses (Figure 2).

### *5.4. Differences Between Languages*

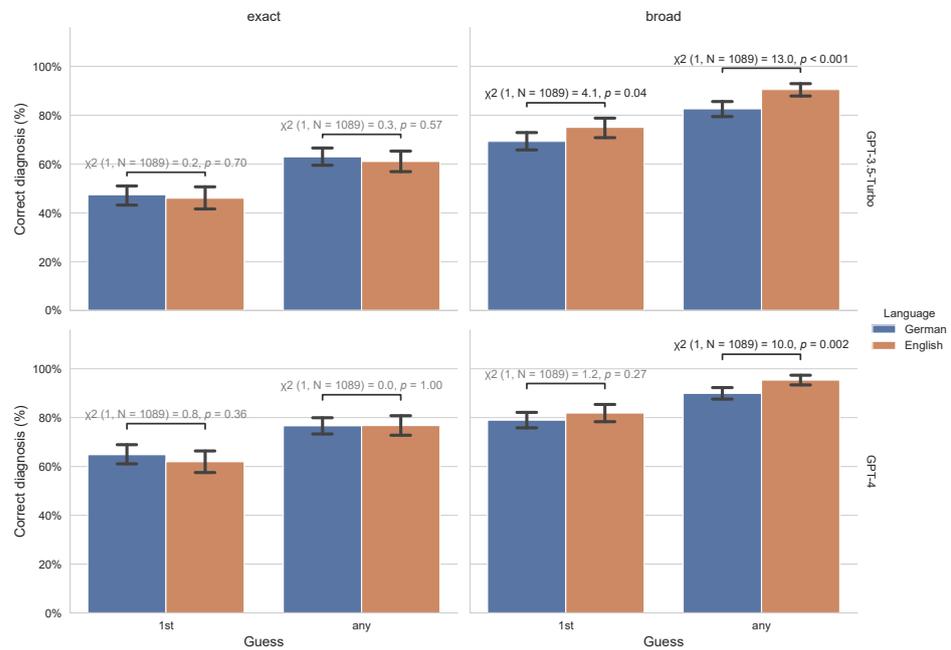
When isolating for the model type and comparing across languages, differences became apparent. For GPT-3.5-Turbo under the broad metric, the first and any guesses for the English language significantly surpassed that of German ( $\chi^2(1, N = 1089) = 4.1/13.0, p = 0.04/p < 0.001$ , respectively). However, when examining GPT-4's performance, while differences remained significant for any guess under the broad metric, the gaps narrowed, indicating the model's more consistent performance across languages (Figure 3).

### *5.5. Comparative Analysis Between First Guess and Any Guess*

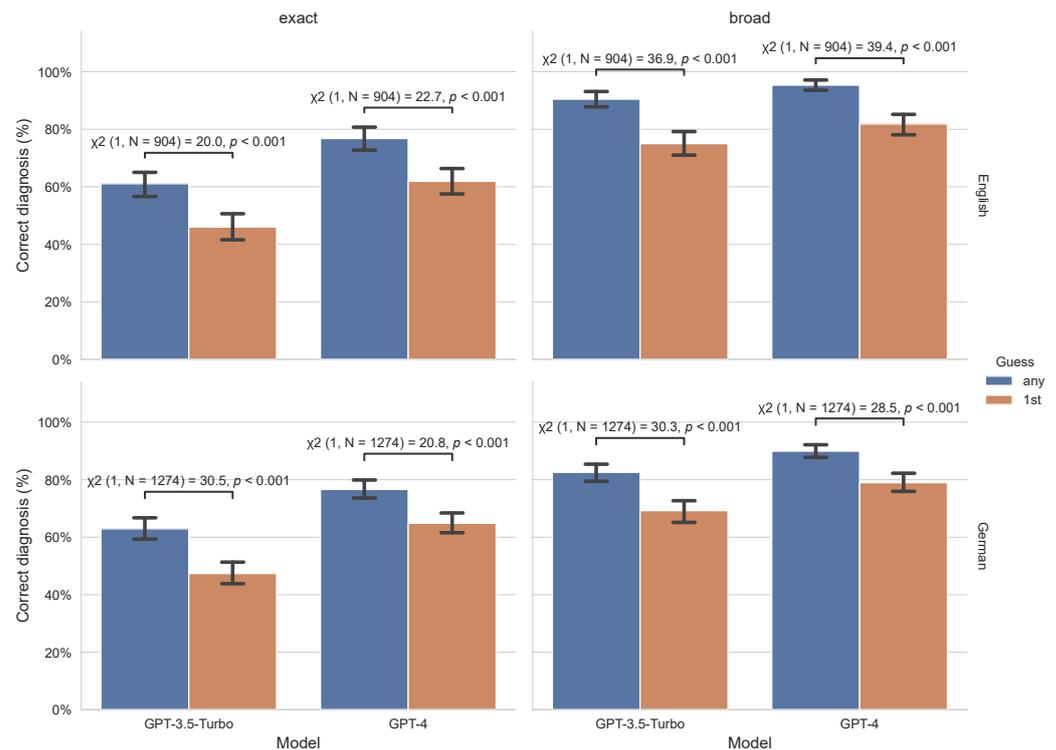
When comparing the first guess to any subsequent guesses, both models across both languages led to significantly improved accuracy, except for the first guesses made by GPT-4 considering the broad matching of diagnoses. The differences for the exact matches of model predictions and correct diagnoses were nullified for both types of guesses and both models (Figure 4).



**Figure 2.** A comparative analysis of diagnostic accuracy between the large language models GPT-4 and GPT-3.5-Turbo. The comparison is structured across two dimensions, namely the type of match—exact or broad (represented in the first and second columns, respectively)—and the language used—English or German (depicted in the first and second rows, respectively). Pairwise comparisons were conducted using chi-square tests to evaluate the differences in performance between the two models, with the results indicated above the respective bars in each subplot.



**Figure 3.** A comparative analysis of diagnostic accuracy for examination questions in the two languages of German and English. The comparison is structured across two dimensions, namely the type of match—exact or broad (represented in the first and second columns, respectively)—and the large language model used—GPT-3.5-Turbo or GPT-4 (depicted in the first and second rows, respectively). Pairwise comparisons were conducted using chi-square tests to evaluate the differences in performance between the languages, with the results indicated above the respective bars in each subplot. Comparisons not reaching statistical significance ( $p > 0.05$ ) are highlighted in gray.



**Figure 4.** A comparative analysis of diagnostic accuracy for the 1st vs. any guess. The comparison is structured across two dimensions, namely the type of match—exact or broad (represented in the first and second columns, respectively)—and the language of the examination question—English or German (depicted in the first and second rows, respectively). Pairwise comparisons were conducted using chi-square tests to evaluate the differences in performance between the 1st and any guess, with the results indicated above the respective bars in each subplot.

### 5.6. Question Difficulty and Correct LLM Diagnoses

In our assessment of the relationship between question difficulty and the accuracy of predicted diagnoses using Spearman's rank correlation, distinct patterns emerged contingent on the model, language, and metric applied. For English language questions, no discernible correlation was observed across all category combinations. On the other hand, for German language questions, the GPT-3.5-Turbo model, focusing on exact matches, showed a decline in correct answers as question difficulty increased, evident in both the primary guess ( $r = -0.13, p < 0.001$ ) and any subsequent guesses ( $r = -0.08, p = 0.04$ ). Similarly, the GPT-4 model presented this trend, but it was confined to exact matches on primary guesses, suggesting a decrease in accuracy with more challenging questions. Interestingly, for other category combinations associated with German questions, no clear link was found between question difficulty and the models' accuracy (Table 1).

### 5.7. Bing Evaluation of GPT-4 and GPT-3.5-Turbo Results

Next, we assessed Bing Chat's efficacy over 4356 diagnosis pairs, which comprised all unique combinations of correct exam answers and the LLM's diagnostic suggestions. Our evaluation underscores Bing Chat's commendable accuracy of 0.79 for both exact and approximate matches. Particularly noteworthy was the positive predictive value (PPV), which reached 0.90 for exact matches and 0.92 for approximate matches (Table 2), pointing to Bing Chat's strong potential in predicting correct diagnosis matches.

**Table 1.** Correct diagnosis [%] per question difficulty (very easy, easy, intermediate, difficult, and very difficult) among different parameters (language, metric, model, and guess). Spearman’s rank correlation between question difficulty and number of correct answers for each category are shown. *p*-values < 0.05 are in bold.

Language	Metric	Model	Guess	Percentage of Correct Answers for Each Difficulty Category					Spearman	
				Very Easy	Easy	Intermediate	Difficult	Very Difficult	<i>r</i>	<i>p</i>
English N = 452 df = 450	broad	GPT-3.5-Turbo	1st	77.1	74.5	72.2	78	73.7	−0.02	0.68
			any	91.5	89.2	91.7	92	84.2	−0.01	0.87
		GPT-4	1st	83.1	80.3	83.3	84	73.7	−0.01	0.89
			any	94.9	93.6	99.1	96	89.5	0.03	0.48
	exact	GPT-3.5-Turbo	1st	47.5	45.2	43.5	50	47.4	0	0.94
			any	66.1	56.7	59.3	66	63.2	−0.02	0.73
		GPT-4	1st	68.6	55.4	67.6	58	52.6	−0.04	0.38
			any	79.7	68.8	83.3	82	73.7	0.04	0.44
		N (%)	<b>118 (26.1%)</b>	<b>157 (34.7%)</b>	<b>108 (23.9%)</b>	<b>50 (11.1%)</b>	<b>19 (4.2%)</b>	<b>−0.01</b>	<b>0.72</b>	
German N = 637 df = 635	broad	GPT-3.5-Turbo	1st	72	66.8	71.4	62.9	70	−0.03	0.45
			any	84.5	80.3	85.1	80.6	75	−0.02	0.60
		GPT-4	1st	78.8	82.7	77.3	75.8	65	−0.04	0.30
			any	91.7	90.4	89.6	90.3	70	−0.06	0.13
	exact	GPT-3.5-Turbo	1st	56	46.2	46.8	30.6	35	<b>−0.13</b>	<b>&lt;0.001</b>
			any	68.4	60.6	66.9	48.4	50	<b>−0.08</b>	<b>0.04</b>
		GPT-4	1st	68.4	69.2	62.3	51.6	45	<b>−0.11</b>	<b>0.007</b>
			any	77.2	79.3	76.6	71	60	−0.05	0.22
		N (%)	<b>193 (30.3%)</b>	<b>208 (32.7%)</b>	<b>154 (24.2%)</b>	<b>62 (9.7%)</b>	<b>20 (3.1%)</b>	−0.06	<0.001	

**Table 2.** Evaluation metrics for Bing Chat’s abilities to predict the similarity of two medical terms.

Metric	Exact Matches	Approximate Matches
Accuracy	0.79	0.79
Sensitivity	0.74	0.83
Specificity	0.86	0.64
Positive Predictive Value (PPV)	0.90	0.92
Negative Predictive Value (NPV)	0.67	0.44

## 6. Discussion

The results from our assessment of the GPT models present compelling evidence on their diagnostic acumen. Remarkably, the models exhibit considerable capability in identifying accurate diagnoses, even in the absence of multiple-choice options. In the English dataset, GPT-4 demonstrated an impressive accuracy of 81.9% on its first guess and 95.4% when considering any of its three predictions under broad accuracy metrics. For the German dataset, GPT-4 achieved an admirable accuracy of 79.0% on its initial guess and 90.0% on any of its predictions. While the outcomes of the responses are not as strong as the results of the models with multiple-choice answers, overall, similarly positive results are evident when considering multiple answer options [27]. This efficacy is further accentuated when broader diagnostic categories are taken into account or when considering up to three model-generated guesses rather than only the primary one. However, it is important to note that in the case of German language questions, there was a decline in the accuracy of ChatGPT’s models as question difficulty increased. For the GPT-3.5-Turbo model, exact

match accuracy decreased with question difficulty. The GPT-4 model also saw a decline in exact match accuracy for primary guesses. This suggests that while the models perform well overall, there is room for improvement when dealing with more challenging questions in the German language.

The positive outcomes from this study set a promising precedent for real-world applications. Specifically, one could envision a scenario where a clinician feeds in patient histories and receives diagnostic suggestions “on the fly” from the model. Such a system could serve as an invaluable clinical advisory tool, assisting doctors in swiftly pinpointing time-sensitive diagnoses, which can be especially vital in critical care situations [28]. A comparison of performance in diagnosis and triage has already shown similarly good results in accurate diagnosis, albeit with room for improvement in triage performance [29]. Likewise, GPT-3 appears to generate differential diagnoses at a level comparable to that of medical professionals [30]. This is further corroborated by our findings, which demonstrate a high rate of accurate diagnoses across various medical conditions. With the addition of image recognition, this support can be further expanded. However, it still requires further scientific investigation. The integration of AI in healthcare is not just a task for computer scientists or AI experts but also requires close collaboration between AI practitioners, clinicians, ethicists, and policy makers.

Not only do the models need to be trained and adapted for further use, there is also a need for training medical staff to optimally use the programs and minimize potential sources of error. One way to improve this is through feedback mechanisms. If medical professionals could give feedback to the AI, this can be used for continuous improvement of the processes, as seen in reinforcement learning through human feedback (RLHF) [31].

It is essential to emphasize that AI-assisted diagnosis in its current form should be viewed as a complementary tool that supports, rather than replaces, physician expertise. The complexity of medical diagnosis requires a physician’s years of training, clinical experience, and holistic understanding of patient care—capabilities that extend far beyond pattern recognition.

A valid criticism is that the data on which GPT-3.5-Turbo and GPT-4 were trained on is only current up to September 2021 [32]. Nevertheless, medicine is a constantly evolving field, and the AI requires regular updates based on the latest medical research and findings to provide current recommendations. Another critical concern is the potential data privacy issues when using AI models [33]. Even if individual queries are not currently stored, additional regulations and protocols are needed to safely utilize the technology in the future. It is also essential to verify the source and quality of the data used to train these large language models before considering their general benefits. If the models were trained with biased or inaccurate data, this could lead to data unreliability and bias, which can be harmful in healthcare. To this day, OpenAI has not released their training dataset. However, they recently unveiled ChatGPT Enterprise to address security and privacy issues [34].

The incorporation of language models into healthcare can have significant economic implications for hospitals. Through appropriate integration into daily operations, these models have the potential to alleviate the burden on healthcare staff, reducing the need for overtime. Especially during times of staff shortages [35,36], this technology could serve as a valuable tool to streamline and facilitate work processes. Studies have yet to investigate how the widespread adoption of large language models might impact healthcare insurance costs, patient care expenses, and even the pharmaceutical industry through more accurate and streamlined diagnoses.

Our findings demonstrate that AI can rapidly propose suitable diagnoses, potentially facilitating the diagnostic process, yielding cost savings by avoiding misdiagnoses, and enhancing patient safety. However, a crucial element that currently cannot be replaced is the human touch, which is essential for a strong doctor–patient relationship [37–39]. Thus, it is hardly surprising that one of the major incentives for employing AI in healthcare is to enhance the efficiency of human medical staff [40]. This could free up physicians to

provide more attentive care and listen to their patients rather than being bogged down with electronic medical record documentation [41].

An important observation to highlight is the commendable performance of the models in languages other than their primary training language, English. This demonstrates not just the flexibility of the models but also suggests their potential applicability in diverse linguistic regions, especially in regions where healthcare professionals are scarce. The robust performance in the German language dataset indicates the models' readiness for deployment outside the English-speaking domain, paving the way for a broader global impact. Even though this study was able to show that the models have equally good results in languages they were not mainly trained in, it is important to consider the cultural nuances in healthcare in different regions. Symptoms might be described differently and cultural practices could influence health outcomes. This requires further investigation in the future.

Bing Chat demonstrates promising capabilities in detecting similarities between medical terms, showcasing high accuracy in both exact and broader matches. While prior research utilizing only GPT-3.5-Turbo highlighted limited efficacy in general and medical named entity recognition [42,43], our study breaks new ground by evaluating medical similarity detection using Bing Chat, leveraging OpenAI's most recent LLM GPT-4 enhanced with web access. This preliminary success suggests its potential to become a pivotal tool in medical diagnostics, leveraging deep learning and real-time web insights to bridge the gap between technology and nuanced medical understanding. The system adeptly accommodates near-perfect matches, paving the way for practical applications where exact matches are not mandatory and presenting a promising avenue in the diagnostic landscape through enriched data access. Bing Chat also has its limitations. Its performance depends on internet connectivity and access to accurate up-to-date medical information. Furthermore, medical terms can vary significantly across languages and cultural contexts, potentially leading to inconsistencies in term recognition. These limitations highlight the need for further testing and refinement before integrating Bing Chat into high-stakes clinical environments. Nonetheless, its ability to recognize related terms efficiently is a promising step toward more robust AI support in healthcare.

Our study, however, has some inherent limitations. Primarily focusing on English and German datasets may limit the applicability of our findings to other less common languages and thus a truly global context. The subjective categorization of question difficulty based on students' correct answers could introduce variability, possibly affecting the perceived performance of the models against different levels of challenge. Additionally, the evaluation was limited to GPT-3.5-Turbo and GPT-4, without considering the nuances of other iterations or open-source models. Furthermore, we did not evaluate key performance metrics such as model runtime, response time for diagnostic suggestions, and usability by clinical professionals. It is crucial to emphasize that these general-purpose LLMs, in their current form, are not intended or validated for direct clinical diagnosis or treatment decisions within doctor-patient relationships. And while the findings are encouraging, real-world clinical validation of these models is essential before any firm conclusions on their applicability can be drawn. And finally, diagnostic reasoning involves not only identifying disease patterns but also the integration of clinical, social, and personal factors of the patient. The models' performance might differ depending on the complexity of a real-world clinical context.

Future studies are crucial to validate GPT-4 and GPT-3.5-Turbo's diagnostic capabilities in real-world clinical settings. We propose prospective trials comparing AI-generated suggestions with final diagnoses and outcomes while prioritizing ethical considerations such as patient safety, AI's impact on decision-making, and integration into electronic health record (EHR) systems. These steps will ensure rigorous validation for a safe and effective use in healthcare.

In conclusion, while further validations and iterations are necessary, the promise shown by the models in providing accurate diagnostics across languages suggests a promising horizon for AI in healthcare.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai5040128/s1>, Supplementary Table S1: Raw data of correct answers, model responses, Bing Chat evaluation, and MeSH and UMLS matching. Supplementary Table S2: Patterns to dismantle strings without an exact match into substrings. Supplementary Table S3: Question template for the evaluation of pairs of medical terms. Supplementary Table S4: Automatically determined difficulty category by how often it is answered correctly or incorrectly in AMBOSS. All questions are then sorted by difficulty.

**Author Contributions:** All authors contributed to the study conception and design. R.K. and J.R. collected, verified, and analyzed the data and had access to the raw data. R.K. developed the statistical methods. J.R. and R.K. wrote the first draft of the manuscript, which was edited by R.M. and T.I.W. All processes were supervised by R.K. and J.R. R.K. and J.R. took final responsibility for the decision to submit for publication. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors received no financial support for the research, authorship, and/or publication of this article. This work was supported by the Open Access Publication Fund of the University of Bonn.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The Supplementary Materials contain the raw response data (Supplementary Table S1). For the German exam questions, inquiries can be directed to the Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP). Both German and English questions are also available at Amboss.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
2. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J. Med. Syst.* **2023**, *47*, 33. [CrossRef]
3. Kim, J.K.; Chua, M.; Rickard, M.; Lorenzo, A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J. Pediatr. Urol.* **2023**, *19*, 598–604. [CrossRef]
4. Introducing ChatGPT. Available online: <https://openai.com/blog/chatgpt> (accessed on 22 June 2023).
5. Timothy, M. ChatGPT with Browsing vs. ChatGPT Plugins: Which Version of ChatGPT Should You Use? MUO. 4 July 2023. Available online: <https://www.makeuseof.com/which-version-of-chatgpt-should-you-use/> (accessed on 3 September 2023).
6. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Amodei, D. Language Models are Few-Shot Learners. 2020. Available online: <https://arxiv.org/abs/2005.14165> (accessed on 22 June 2023).
7. Learnprompt.org. Know the Difference—Chat GPT 3.5 vs GPT 4 [UPDATED]. LearnPrompt.Org. 2023. Available online: <https://www.learnprompt.org/chat-gpt-3-vs-gpt-4/> (accessed on 3 September 2023).
8. Confirmed: The New Bing Runs on OpenAI's GPT-4. 14 March 2023. Available online: [https://blogs.bing.com/search/march\\_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4/](https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4/) (accessed on 11 September 2023).
9. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef] [PubMed]
10. Ärzteblatt, D.Ä.G. Redaktion Deutsches ChatGPT Passes German State Examination in Medicine with Picture Questions Omitted (30 May 2023). Deutsches Ärzteblatt. Available online: <https://www.aerzteblatt.de/int/archive/article?id=231006> (accessed on 3 September 2023).
11. Ali, R.; Tang, O.Y.; Connolly, I.D.; Sullivan, P.L.Z.; Shin, J.H.; Fridley, J.S.; Asaad, W.F.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. *Neurosurgery* **2023**, *93*, 1353–1365. [CrossRef] [PubMed]
12. Friederichs, H.; Friederichs, W.J.; März, M. ChatGPT in medical school: How successful is AI in progress testing? *Med. Educ. Online* **2023**, *28*, 2220920. [CrossRef] [PubMed]

13. Medizin-www.impp.de. Available online: <https://www.impp.de/pruefungen/medizin.html> (accessed on 3 September 2023).
14. Step 3 Exam Content | USMLE. Available online: <https://www.usmle.org/step-exams/step-3/step-3-exam-content> (accessed on 3 September 2023).
15. Strong, E.; DiGiammarino, A.; Weng, Y.; Kumar, A.; Hosamani, P.; Hom, J.; Chen, J.H. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Intern. Med.* **2023**, *183*, 1028–1030. [[CrossRef](#)] [[PubMed](#)]
16. Long, C.; Lowe, K.; Santos, A.D.; Zhang, J.; Alanazi, A.; O'Brien, D.; Wright, E.; Cote, D. Evaluating ChatGPT-4 in Otolaryngology–Head and Neck Surgery Board Examination using the CVSA Model. *MedRxiv* **2023**. [[CrossRef](#)]
17. TrueHealth | Free Online Digital Clinic. Available online: <https://www.truehealthapp.com/> (accessed on 3 September 2023).
18. What Is Google Health?—Google Health. Available online: <https://health.google/> (accessed on 3 September 2023).
19. Health. Powered by Ada. Available online: <https://ada.com/> (accessed on 3 September 2023).
20. Rubeis, G.; Dubbala, K.; Metzler, I. “Democratizing” artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Front. Genet.* **2022**, *13*, 902542. [[CrossRef](#)]
21. Roth, P.H.; Bruni, T. Participation, Empowerment, and Evidence in the Current Discourse on Personalized Medicine: A Critique of “Democratizing Healthcare”. *Sci. Technol. Hum. Values* **2022**, *47*, 1033–1056. Available online: <https://journals.sagepub.com/doi/abs/10.1177/01622439211023568> (accessed on 3 September 2023). [[CrossRef](#)]
22. Grishman, R.; Sundheim, B. Message Understanding Conference-6: A brief history. In Proceedings of the 16th Conference on Computational Linguistics, Kiev, Ukraine, 21–23 April 2021; Association for Computational Linguistics: Monroe County, PA, USA, 1996; pp. 466–471. [[CrossRef](#)]
23. Zhang, S.; Cheng, H.; Gao, J.; Poon, H. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning. *arXiv* **2023**. Available online: <http://arxiv.org/abs/2208.14565> (accessed on 3 September 2023).
24. Zhang, S.; Cheng, H.; Vashishth, S.; Wong, C.; Xiao, J.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Knowledge-Rich Self-Supervision for Biomedical Entity Linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Monroe County, PA, USA, 2022; pp. 868–880. [[CrossRef](#)]
25. ChatGPT—Release Notes | OpenAI Help Center. Available online: <https://help.openai.com/en/articles/6825453-chatgpt-release-notes> (accessed on 3 September 2023).
26. Medizinwissen, Auf Das Man Sich Verlassen Kann | AMBOSS. Available online: <https://www.amboss.com/de> (accessed on 3 September 2023).
27. Roos, J.; Kasapovic, A.; Jansen, T.; Kaczmarczyk, R. Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. *JMIR Med. Educ.* **2023**, *9*, e46482. [[CrossRef](#)] [[PubMed](#)]
28. Schultebrucks, K.; Chang, B.P. The opportunities and challenges of machine learning in the acute care setting for precision prevention of posttraumatic stress sequelae. *Exp. Neurol.* **2021**, *336*, 113526. [[CrossRef](#)] [[PubMed](#)]
29. Levine, D.M.; Tuwani, R.; Kompa, B.; Varma, A.; Finlayson, S.G.; Mehrotra, A.; Beam, A. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. *medRxiv* **2023**. [[CrossRef](#)]
30. Hirosawa, T.; Harada, Y.; Yokose, M.; Sakamoto, T.; Kawamura, R.; Shimizu, T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3378. [[CrossRef](#)] [[PubMed](#)]
31. Zheng, R.; Dou, S.; Gao, S.; Hua, Y.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Zhou, Y.; et al. Secrets of RLHF in Large Language Models Part I: PPO. *arXiv* **2023**. [[CrossRef](#)]
32. GPT-4. Available online: <https://openai.com/research/gpt-4> (accessed on 3 September 2023).
33. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **2019**, *25*, 30–36. [[CrossRef](#)]
34. Introducing ChatGPT Enterprise. Available online: <https://openai.com/blog/introducing-chatgpt-enterprise> (accessed on 3 September 2023).
35. Schnack, H.; Uthoff, S.A.K.; Ansmann, L. The perceived impact of physician shortages on human resource strategies in German hospitals—A resource dependency perspective. *J. Health Organ. Manag.* **2022**, *36*, 196–211. [[CrossRef](#)]
36. Winter, V.; Schreyögg, J.; Thiel, A. Hospital staff shortages: Environmental and organizational determinants and implications for patient satisfaction. *Health Policy* **2020**, *124*, 380–388. [[CrossRef](#)]
37. Karches, K.E. Against the iDoctor: Why artificial intelligence should not replace physician judgment. *Theor. Med. Bioeth.* **2018**, *39*, 91–110. [[CrossRef](#)]
38. Gala, D.; Makaryus, A.N. The Utility of Language Models in Cardiology: A Narrative Review of the Benefits and Concerns of ChatGPT-4. *Int. J. Environ. Res. Public Health* **2023**, *20*, 6438. [[CrossRef](#)] [[PubMed](#)]
39. Bianchi, D.W. The power of human touch in the era of artificial intelligence. *Pediatr. Res.* **2019**, *86*, 670–671. [[CrossRef](#)] [[PubMed](#)]
40. Shuaib, A.; Arian, H.; Shuaib, A. The Increasing Role of Artificial Intelligence in Health Care: Will Robots Replace Doctors in the Future? *Int. J. Gen. Med.* **2020**, *13*, 891–896. [[CrossRef](#)] [[PubMed](#)]
41. Aminololama-Shakeri, S.; López, J.E. The Doctor-Patient Relationship with Artificial Intelligence. *Am. J. Roentgenol.* **2019**, *212*, 308–310. [[CrossRef](#)]

- 
42. Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G. GPT-NER: Named Entity Recognition via Large Language Models. *arXiv* **2023**. [[CrossRef](#)]
  43. Zhou, W.; Zhang, S.; Gu, Y.; Chen, M.; Poon, H. Universal NER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv* **2023**. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.