The accuracy and completeness of ChatGPT in Belgian general practice: a descriptive analysis of therapeutic cases.

Devolder Nathalie, KU Leuven

Reynders Emma, KU Leuven

van Bijsterveld Aline, KU Leuven

Promotor: Prof. Dr. Aertgeerts Bert

Prof. Dr. Delvaux Nicolas

Master of Family Medicine

Masterproef Huisartsgeneeskunde
Academiejaar: 2024 – 2025

©Copyright by KU Leuven

Deze masterproef is een examendocument dat niet werd gecorrigeerd voor eventueel vastgestelde fouten. Met uitzondering van het door de wet toegestaan gebruik, is zonder voorafgaande schriftelijke toestemming van de auteur(s) het overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie in verband met het overnemen en/of gebruik van gedeelten uit deze publicatie, wendt u zich tot de KU Leuven, Faculteit Geneeskunde, Campus Gasthuisberg O&N2, Herestraat 49 – bus 400, 3000 Leuven.

Voorafgaande schriftelijke toestemming van de promotor(en) is eveneens vereist voor het aanwenden van de in dit afstudeerwerk beschreven (originele) methoden, producten en programma's voor commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Abstract

Background: Chat Generative Pre-Trained Transformer (ChatGPT) is an advanced large language model-based generative artificial intelligence that has gained widespread attention since its introduction in 2022. Its potential applications in medicine are currently under investigation. In the context of general practice, ChatGPT may facilitate the creation of patient education materials, development of informational pamphlets (e.g., voiding diaries, headache diaries, translated questionnaires), and enhancement of patient communication. Furthermore, it has potential utility in clinical decision-making by assisting with differential diagnosis expansion, triage, and therapeutic option exploration. However, limitations exist, particularly with the publicly available version of ChatGPT 3.5, which is constrained by training data up to September 2021, potentially leading to outdated medical information. In contrast, ChatGPT 4.0, accessible via subscription, has direct access to the internet, which may mitigate this limitation. Given these constraints, this study aims to evaluate the reliability and applicability of ChatGPT in the general practice, where inaccuracies may have critical consequences.

Methods: This study assessed the therapeutic capabilities of ChatGPT 3.5 and ChatGPT 4.0 using five therapeutic cases derived from general practice in Belgium. Each model was presented with the same cases three times to investigate reproducibility, and clinical queries were posed accordingly. The responses generated by ChatGPT were systematically compared with established guidelines from clinical databases. A Likert scale was employed to assess response accuracy (scored out of 5) and completeness (scored out of 3).

Results: For therapeutic accuracy, ChatGPT 3.5 achieved a mean score of 3.867 (SD = 1.187), indicating reasonably accurate therapeutic recommendations but with considerable variability. In contrast, ChatGPT 4.0 had a higher score of 4.667 (SD = 0.816), reflecting improved accuracy and reduced variability compared to ChatGPT 3.5.

For therapeutic completeness, ChatGPT 3.5 achieved a mean score of 2.667 (SD = 0.617), indicating that while therapeutic recommendations were moderately complete, there was some inconsistency. GPT 4.0 demonstrated a slight improvement with a mean score of 2.733 (SD = 0.458), suggesting enhanced completeness and greater consistency relative to ChatGPT 3.5.

Conclusion: This study highlights the potential and limitations of ChatGPT in a general practice setting. ChatGPT demonstrated moderate therapeutic accuracy, with ChatGPT 4.0 having more accurate predictions. However, these differences were not statistically significant. Both models exhibited good completeness, with similar scores.

The small sample size of this study suggests caution when interpreting these results, as it may introduce bias. To improve the reliability and generalizability of these findings, future research should involve larger sample sizes before it is applied in medical practice.

Nederlandstalig abstract

Achtergrond: Chat Generative Pre-Trained Transformer (ChatGPT) is artificiële intelligentie, gebaseerd op “advanced large language models”, dat sinds de introductie in 2022 een veelbesproken onderwerp is. De potentiële toepassingen in de geneeskunde worden momenteel onderzocht. In de huisartsgeneeskunde kan ChatGPT helpen bij het creëren van educatiemateriaal, de ontwikkeling van informatiefolders (bijv. plasdagboeken, hoofdpijndagboeken, vertaalde vragenlijsten) en het verbeteren van de communicatie met patiënten. Bovendien heeft het potentieel nut in klinische besluitvorming door te ondersteunen bij het uitbreiden van een diagnostisch landschap, triage en het verkennen van therapeutische opties. Er zijn echter beperkingen. Zo is er geen transparante bronvermelding. Daarnaast zijn de trainingsdata van de openbaar beschikbare versie van ChatGPT 3.5 beperkt tot september 2021, wat kan leiden tot verouderde medische informatie. In tegenstelling hiermee heeft ChatGPT 4.0, beschikbaar via abonnement, directe informatie via het internet, waardoor deze beperking mogelijks wordt verholpen. Gezien deze beperkingen heeft deze studie als doel de betrouwbaarheid en toepasbaarheid van ChatGPT in de huisartsgeneeskunde te evalueren, waar onnauwkeurigheden ernstige gevolgen kunnen hebben.

Methode: Deze studie beoordeelde de therapeutische capaciteiten van ChatGPT 3.5 en ChatGPT 4.0 aan de hand van vijf therapeutische casussen uit de Belgische huisartsgeneeskunde. Aan elk model werden de casussen drie keer gepresenteerd en klinische vragen werden gesteld. De door ChatGPT gegenereerde antwoorden werden systematisch vergeleken met richtlijnen uit klinische databases. Een Likertschaal werd gebruikt om de nauwkeurigheid (beoordeeld op 5) en de volledigheid (beoordeeld op 3) van de antwoorden te evalueren.

Resultaten: Wat betreft therapeutische nauwkeurigheid behaalde ChatGPT 3.5 een gemiddelde score van 3.867 (SD = 1.187), wat wijst op een redelijk accurate therapeutische nauwkeurigheid met aanzienlijke variabiliteit. In tegenstelling daarmee behaalde ChatGPT 4.0 een hogere score van 4.667 (SD = 0.816), wat wijst op betere nauwkeurigheid en minder variabiliteit ten opzichte van ChatGPT 3.5.

Wat betreft therapeutische volledigheid behaalde ChatGPT 3.5 een gemiddelde score van 2.667 (SD = 0.617), wat aangeeft dat hoewel therapeutische aanbevelingen matig volledig waren, er enige inconsistentie bestaat. ChatGPT 4.0 vertoonde een lichte verbetering met een gemiddelde score van 2.733 (SD = 0.458), wat wijst op een betere volledigheid en betere consistentie ten opzichte van ChatGPT 3.5.

Conclusie: Deze studie benadrukt het potentieel en de beperkingen van ChatGPT in de huisartsenpraktijk. ChatGPT toonde een matige therapeutische nauwkeurigheid, waarbij ChatGPT 4.0 meer accurate antwoorden gaf. Deze verschillen waren echter statistisch niet significant. Beide modellen vertoonden goede vergelijkbare volledigheid. De kleine steekproefgrootte van deze studie vraagt om voorzichtigheid bij de interpretatie van deze resultaten, aangezien dit ruimte geeft voor bias. Toekomstig onderzoek moet gebeuren met grotere steekproeven om de betrouwbaarheid van deze bevindingen te verbeteren voordat het in de medische praktijk kan worden toegepast.

Table of contents

Abstract 2

Nederlandstalig abstract 3

Table of contents 4

Introduction 6

Methodology 8

- Case Selection and Standardization 8
- Therapeutic Questioning and AI Response Evaluation 9
- Descriptive analysis 9
- Reproducibility Analysis 9

Results 11

- Methodology 11
- Accuracy 12
- Completeness 13
- Statistical Analysis 13
  - Summary statistics 13
  - Outliers Analysis 14
  - Paired t-test Analysis 15
  - Paired line plots 16
  - Difference plots 17

Discussion 18

- Overall findings 18
- Reproducibility 19
- ChatGPT 3.5 versus 4.0 19
- Limitations 20
- Overall conclusions 21

Conclusion 24

Acknowledgments 25

Conflicts of interest 26

Financial Disclosure 26

References 27

Approval research ethics committee 30
Addenda 31
Addendum 1: Template 31
Addendum 2: Likert Scale 32
Addendum 3: All scores by reader 33
Addendum 4: All cases with answers 33

Introduction

Artificial intelligence (AI) driven tools are developing at an ever-increasing speed and applications of these AI driven tools keep expanding ever since. (1)

A well-known example of an AI tool is ChatGPT, an AI-driven chatbot, which uses a large language model (LLM) and natural language processing (NLP). This technology enables users to ask questions and receive responses in a conversational manner, mimicking human dialogue. (2) NLP uses machine learning technology to interpret and comprehend human language. Using a large, predetermined dataset, this NLP software can process input, analyze intent and sentiment, and generate real-time responses. (3)

Big data is the umbrella term for all the sources utilized by artificial intelligence programs. It consists of extremely large datasets that can be analyzed computationally to identify patterns. In 2023, ChatGPT was trained on 570 GB of data, consisting of books, web texts, articles, etc.., in a machine learning method. (4,5) However, the specific datasets used for training ChatGPT remain undisclosed. ChatGPT does not provide source citations, and the program itself reports that "the specific references used to train ChatGPT are not publicly available, [...]." As a result, ChatGPT cannot independently assess the reliability of its sources. The data selection process is managed by OpenAI's research team, which applies quality control measures. (4) This lack of explicit source verification creates the risk of so-called AI hallucinations-generated content that is inaccurate, non-evidence-based, or entirely erroneous. The latest version, ChatGPT 4.0, has reportedly reduced this risk. (5,6)

ChatGPT is taking the world by storm. From a medical point of view, its potential applications are currently being explored. Within general practice, multiple possibilities arise for both physicians and patients. (7) Patients may use it to quickly request documents for specialists or insurance purposes, or seek clarification on medical information they did not fully understand during a previous consultation. Additionally, they may use ChatGPT to ask sensitive questions they might hesitate to pose to their physician.

For general practitioners, this technology could assist in creating patient education materials, developing informational pamphlets (such as voiding diaries, headache diaries, or translated questionnaires), and generating content to enhance patient communication. (8) It may also support clinical decision-making by expanding differential diagnoses, assisting in triage, drafting sick certificates, and composing letters to insurers requesting coverage for costly medications. Furthermore, ChatGPT has the potential to improve adherence to medical prescriptions by simplifying language and minimizing miscommunication. (9)

Although ChatGPT is trained on extensive datasets, it remains a generative model, which means its responses can sometimes be irrelevant or nonsensical. Another limitation of ChatGPT, and all NLP-based systems, is that their performance is constrained by the quality and scope of the training data. (10,11) The publicly available version of ChatGPT 3.5 is limited to data collected up to September 2021, which can lead to outdated information in dynamic fields such as medicine. ChatGPT 4.0, which requires a subscription, can connect to the internet, potentially overcoming this limitation. (4) Given

these constraints, it is crucial to assess whether ChatGPT can be reliably applied in the medical field, where errors may have serious consequences.

In this study, we focus on evaluating ChatGPT's ability to address complex and less commonly encountered primary care issues. The intervention involves comparing ChatGPT's responses with established medical sources, including BCFI (Belgisch Centrum voor Farmacotherapeutische Informatie) (12), NHG (Nederlands Huisartsen Genootschap) (13), Domus Medica (14) and UpToDate (15). By systematically evaluating ChatGPT's therapeutic accuracy, adherence to clinical guidelines, and response reproducibility, this study aims to provide insight into its potential role as a decision-support tool in general practice. The comparative analysis between ChatGPT 3.5 and ChatGPT 4.0 will further clarify the impact of model advancements on clinical reliability and utility.

Methodology

In this study, a selected number of cases from general practice in Belgium were presented to ChatGPT 3.5 and ChatGPT 4.0. Clinical questions were posed to each chatbot, and their responses were compared with expert opinions and/or information from clinical guidelines and databases.

This thesis was conducted alongside a parallel study by the same research team investigating the diagnostic capabilities of ChatGPT. However, the present study specifically explores its therapeutic potential.

Case Selection and Standardization

The cases used in this study were drawn from the researchers' own clinical practice and subsequently anonymized. Each researcher is a general practitioner (GP) trainee working within a Belgian general practice. Common primary care issues were selected, particularly those complicated by factors such as comorbidities and/or polypharmacy. A diverse range of patient populations was included to reflect real-world demographics (table 1).

To ensure consistency, one researcher (VS) centralized, compiled, and formatted the cases into a predefined template, standardizing their presentation. Brand names of medications and abbreviations were avoided. The cases were then entered into ChatGPT using this standardized template (Addendum 1). The following data were provided in a fixed order: nationality, age, gender, general practitioner, main complaint, medical history, medication use, and relevant clinical findings, including history, physical examination results, and additional investigations such as medical imaging and laboratory tests. All data were entered in English.

Case 1	Rotator cuff tear
Case 2	Sialolithiasis parotid gland
Case 3	Toxoplasmosis infection in pregnancy
Case 4	Contraception
Case 5	Diabetes

Table 1: List of therapeutic cases

Therapeutic Questioning and AI Response Evaluation

A standardized prompt was used to elicit therapeutic recommendations from ChatGPT, requesting specificity and adherence to medical guidelines where applicable: *"Give the possible treatment options and the preferred treatment. Be specific and incorporate any applicable guidelines."*

The AI-generated therapeutic recommendations were compared with self-generated answers derived from established clinical databases, including NHG, Domus Medica, the Belgian Centre for Pharmacotherapeutic Information (BCFI) and UpToDate. Preference was given to Dutch-language sources, such as BCFI and NHG, to align the findings with Belgian guidelines. When Dutch-language sources were incomplete or unavailable, English literature was consulted.

The accuracy and completeness of the AI answers in comparison to the self-generated answers were assessed using the Likert scale (Addendum 2). The accuracy scale is a five-point Likert scale and the completeness scale is a three-point Likert scale. In accordance with the therapeutic query, there is a different interpretation of the scores. If a medical error was advised in the therapeutic cases, particularly advice that potentially could harm the patient, the accuracy scale received a score of maximum 2/5.

This scale was chosen as it evaluates both accuracy and completeness of answers. Another advantage is that the Likert scale is user-friendly and easily applicable by different researchers.

Descriptive analysis

Each case was independently scored by two researchers (VBA and SN). Discrepancies between scores were discussed, and if consensus could not be reached, a third researcher (KL) provided a final evaluation. The score assigned by the third researcher (KL) was considered definitive.

To ensure methodological rigor, external statistical advice was sought during the study design and analysis phase, particularly for the interpretation of Likert-scale scoring and the reproducibility assessment.

Reproducibility Analysis

To assess the reproducibility of ChatGPT's responses over time, each case and its corresponding therapeutic question were entered into ChatGPT 3.5 three times at one-month intervals by one researcher (VS). This interval was chosen arbitrarily, as this version of ChatGPT is not actively updated.

For ChatGPT 4.0, the same cases were entered three times at two-day intervals by the same researcher (VS). This interval differs from the interval of ChatGPT 3.5, as ChatGPT 4.0 updates daily. The Likert scores were compared across these repeated entries to evaluate response stability. Descriptive

statistical analyses were used to examine differences over time. In addition, the following statistical analysis was used to compare ChatGPT 3.5 and 4.0.

The outlier analysis identified in the dataset reveals instances where ChatGPT models significantly deviated from the gold standard scores.

The Paired t-test was employed to compare the performance of ChatGPT 3.5 and ChatGPT 4.0 across two categories: diagnostic accuracy and diagnostic completeness. This method is appropriate as it allows for paired comparisons of scores from the same cases, reflecting the relative performance of the two models. Statistics represent the magnitude of the difference between ChatGPT 3.5 and ChatGPT 4.0 in each category. Higher values indicate larger discrepancies between the models. P-value indicates the probability of observing a difference as extreme as the one measured, assuming the null hypothesis (no difference between ChatGPT 3.5 and ChatGPT 4.0 is true). A threshold of $p < 0.05$ was used to determine statistical significance. The estimate reflects the mean difference between ChatGPT 3.5 and ChatGPT 4.0 scores for each category. Positive estimates indicate that ChatGPT 3.5 scored higher on average, whereas negative estimates indicate that ChatGPT 4.0 performed better.

Bar charts, paired line plots and difference plots were generated to visually represent the scores and assess reproducibility.

Results

Methodology

The five therapeutic prompts were created and entered into OpenAI by one of the researchers (VS). This was carried out three times in ChatGPT 3.5 on January 13, 2024, February 27, 2024, and April 30, 2024. Subsequently, after the launch of ChatGPT 4.0, the prompts were run again on November 5, November 7, and November 9, 2024.
The responses from ChatGPT were collected separately by date. These responses were then independently evaluated by two researchers (VBA and SN). Both therapeutic cases and diagnostic cases from the complementary thesis were assessed by the same researchers. The evaluation was based on the guidelines from NHG, BCFI, Uptodate and Domus Medica, followed in this specific order. Eventually, no relevant information was used in the assessments from the latter source.

After all readings were assessed by the first two researchers, a third researcher (KL) was involved to provide a third evaluation for the cases where the first two researchers had not reached the same score (Addendum 3). The assessment from the third researcher was considered the final score (Table 2).

Therapeutic cases

Score accuracy with ChatGPT 3.5

	Reading 1	Reading 2	Reading 3
Case 1	5	4	5
Case 2	4	5	4
Case 3	1	2	4
Case 4	4	5	4
Case 5	5	3	3

Score accuracy with ChatGPT 4.0

	Reading 1	Reading 2	Reading 3
Case 1	4	5	5
Case 2	5	2	5
Case 3	5	4	5
Case 4	5	5	5
Case 5	5	5	5

Score completeness with ChatGPT 3.5

	Reading 1	Reading 2	Reading 3
Case 1	3	3	3
Case 2	3	3	3
Case 3	1	3	2
Case 4	3	3	3
Case 5	2	3	2

Score completeness with ChatGPT 4.0

	Reading 1	Reading 2	Reading 3
Case 1	3	3	3
Case 2	3	3	3
Case 3	3	3	2
Case 4	2	2	2
Case 5	3	3	3

Table 2: Final scores for accuracy (maximum score 5) and completeness (maximum score 3) after evaluation by the three researchers in all cases at three different times (corresponding reading) for both ChatGPT 3.5 (left) and 4.0 (right).

Accuracy

The accuracy was assessed by evaluating the correctness of ChatGPT's output in comparison to medical guidelines, expressed as a percentage. If medical errors were made, they were sanctioned with a lower score of 1/5 or 2/5.

A difference between the accuracy scores for the ChatGPT 3.5 and 4.0 runs was that, in general, there was more variability in the scores for the 3.5 runs. The scores ranged from the minimum to the maximum score, while for the 4.0 runs, the scores varied between 2/5 and the maximum score, with no cases reaching a score of 3/5. For the 4.0 runs, there was a higher frequency of the maximum score of 5/5, achieved twelve times compared to five times in the 3.5 runs.
In total, the third researcher assessed three readings for the ChatGPT 3.5 runs and four cases for the 4.0 runs.

Several specific cases were selected to illustrate key patterns and highlight notable findings.

Case 2: Sialolithiasis
In filling out the second prompt, the diagnosis of sialolithiasis was not explicitly stated. Instead, treatment plans were provided for diagnoses other than the desired diagnosis, such as sialadenitis or acute parotitis, which impacted the assessment. Both the 3.5 and 4.0 runs had one reading each (the first and third readings, respectively) where various diagnoses were listed with corresponding treatments, including sialolithiasis. In reading 2 of the 4.0 run, the lowest score of 2/5 was given, as more than half of the information was incorrect.

Case 3: Toxoplasmosis
The 3.5 runs contained the only case that received the minimum score of 1/5 for accuracy. This was the first reading of case 3, which handled the therapeutic management of toxoplasmosis infection during pregnancy. According to the Likert scale, this indicated that “more than one therapeutic mistake was made.” Reasons for this low score included incorrect advice on follow-up care, referencing irrelevant recommendations, and non-concrete medication advice. This case received also a lower score than other cases in the second and third readings but still scored higher than the minimum score received in the first reading.
In contrast, for the 4.0 runs, this case achieved the maximum score twice. The second reading scored 4/5, as it was less complete than the other readings.

Case 5: Diabetes
Case 5, related to diabetes management, required a third assessment in both the 3.5 and 4.0 runs for two out of three readings. In the 3.5 runs, only the first reading received the maximum score, while in the 4.0 runs, all reading achieved the maximum score. For the second and third readings of the 3.5 runs, the researchers assessed more variable scores: one researcher deemed the readings to be >90% correct enough for a maximum score, while the second researcher rated a score of 3/5 on both readings, which meant 50-70% of the information was correct. The third researcher provided the final score of 3/5.

Completeness

The completeness was evaluated by assessing the presence of relevant components of the proposed therapeutic options in comparison to the guidelines, again expressed in percentages (Addendum 2).

In total, both the 3.5 and 4.0 runs had eleven out of fifteen readings that achieved the maximum score of 3/3 (>75% of all therapeutic components listed, potentially with additional suggestions). The remaining readings scored 2/3 (50-75% complete therapeutic information).
In the 3.5 runs, the third researcher only had to evaluate one reading for completeness, namely the third reading of case 5 on diabetes. In contrast, the third researcher assessed four readings on the 4.0 runs, of which three were for case 4 on contraception.

Several specific cases were selected to illustrate key patterns and highlight notable findings.

Case 3: Toxoplasmosis
Only in the 3.5 runs did the first reading of case 3, concerning the treatment of toxoplasmosis infection in pregnancy, receive the minimum score. It received a score of 1/5 for accuracy as well as a 1/3 for completeness for the same reasons mentioned earlier. Furthermore, the completeness scores of the second and third readings fluctuated, respectively 3/3 and 2/3. In comparison, the 4.0 runs achieved higher scores for this case, with twice a maximum score of 3/3 and once a score of 2/3.

Case 4: Contraception
Case 4 about contraception, received higher completeness scores for all three readings in the 3.5 runs compared to the 4.0 runs. In the first, there was unanimity between the two researchers regarding the score of 3/3. In the latter, a third evaluation was required for each of the three readings, with 2/3 being the final score since not all potential contraceptives were mentioned.

Case 5: Diabetes
Case 5 on diabetes management received lower completeness scores in the 3.5 runs (twice 2/3) than in the 4.0 runs, where the maximum score was always achieved. Reasons for not achieving the maximum score included the omission of lifestyle recommendations for anticoagulation based on the CHA_2DS_2-VASc score, which were mentioned in the other readings.

Statistical Analysis

Summary statistics

Score results were listed descriptively (mean, median, standard deviation (SD), interquartile range (IQR), and were compared between groups using paired-t-test.

Category	Mean GPT 3.5 (SD)	Median GPT 3.5 (IQR)	Mean GPT 4.0 (SD)	Median GPT 4.0 (IQR)	Mean Gold Standard (SD)	Median Gold Standard (IQR)
Therapeutic Accuracy	3.867 (1.187)	4 (1.5)	4.667 (0.816)	5 (0)	5 (0)	5 (0)
Therapeutic Completeness	2.667 (0.617)	3 (0.5)	2.733 (0.458)	3 (0.5)	3 (0)	3 (0)

Table 3: Means and medians for diagnostic accuracy and completeness for both GPT 3.5 and 4.0.

For therapeutic accuracy, the mean score of ChatGPT 3.5 was 3.867, but with large standard deviation (SD = 1.187), indicating reasonably accurate therapeutic recommendations but with substantial variability. In contrast ChatGPT 4.0 achieved a higher mean score of 4.667, and with smaller standard deviation (SD = 0.816), indicating more accurate predictions with lower variability compared to ChatGPT 3.5.
The median (IQR) scores further support this observation. ChatGPT 3.5 achieved a median (IQR) score of 4 (IQR = 1.5), suggesting accurate therapeutic recommendations with noticeable variability. ChatGPT 4.0 improved upon this, achieving a median score of 5 (IQR = 0), demonstrating highly accurate and consistent therapeutic predictions.

For therapeutic completeness, ChatGPT 3.5 achieved a mean score of 2.667 (SD = 0.617). The moderate variability suggests that ChatGPT 3.5 produced therapeutic recommendations that are somewhat incomplete and inconsistent. ChatGPT 4.0 had a slightly higher mean score of 2.733 (SD = 0.458), indicating better completeness and greater consistency than ChatGPT 3.5.
The median (IQR) score of ChatGPT 3.5 was 3 (IQR = 0.5), indicating complete therapeutic recommendations with low variability. ChatGPT 4.0 matched this score, suggesting similar performance with equal variability.

ChatGPT 4.0 demonstrated overall improvements in therapeutic cases compared to ChatGPT 3.5 with higher mean scores. This suggests that ChatGPT 4.0 is more reliable for generating accurate and complete therapeutic recommendations.
The median scores highlight that ChatGPT 4.0 demonstrated greater consistency across all categories compared to ChatGPT 3.5, particularly in therapeutic accuracy where it achieves perfect scores (median = 5, IQR = 0).

Outliers Analysis

Case Id	Reading	GPT 3.5	GPT 4.0	Gold Standard	Category
Case 2	Reading 2	5	2	5	Therapeutic Accuracy

Case 3	Reading 1	1	5	5	Therapeutic Accuracy
Case 3	Reading 2	2	4	5	Therapeutic Accuracy

Table 4: Outliers analysis, the outliers identified in the dataset reveal instances where ChatGPT models significantly deviated from the gold standard scores.

Table 4 depicts an outliers analysis. For example, in case 2 (reading 2) under therapeutic accuracy, ChatGPT 3.5 provided a perfect score (score 5), whereas ChatGPT 4.0 scored only 2, deviating substantially from the gold standard score of 5. Similarly in case 3 (reading 1), ChatGPT 3.5 scored as low as 1 while ChatGPT 4.0 achieved a perfect score (score 5), aligning with the gold standard. These discrepancies highlight variability in ChatGPT models' responses, particularly in challenging cases or scenarios requiring nuanced clinical judgement. Further investigation is warranted to understand the reasons behind these outliers and improve model performance.

Paired t-test Analysis

Category	**statistic**	**p-value**	**estimate**	**method**
Therapeutic Accuracy	-1,976607382	0,068126928	-0,8	Paired t-test
Therapeutic Completeness	-0,323229968	0,751298496	-0,066666667	Paired t-test

Table 5: Paired t-test analysis.

The results for therapeutic accuracy did not reach statistical significance ($p = 0.068$), despite ChatGPT 4.0 achieved slightly higher scores than ChatGPT 3.5. The mean difference of -0.800 suggests a trend favoring ChatGPT 4.0, but the evidence is insufficient to confirm a meaningful difference. For therapeutic completeness, no significant difference was observed ($p = 0.751$), and the mean difference of -0.067 indicates that ChatGPT 4.0's marginally higher scores are negligible and lack practical relevance.

Overall, the results indicate that ChatGPT 4.0 showed comparable performance in therapeutic categories with no statistically meaningful difference observed.

Paired line plots

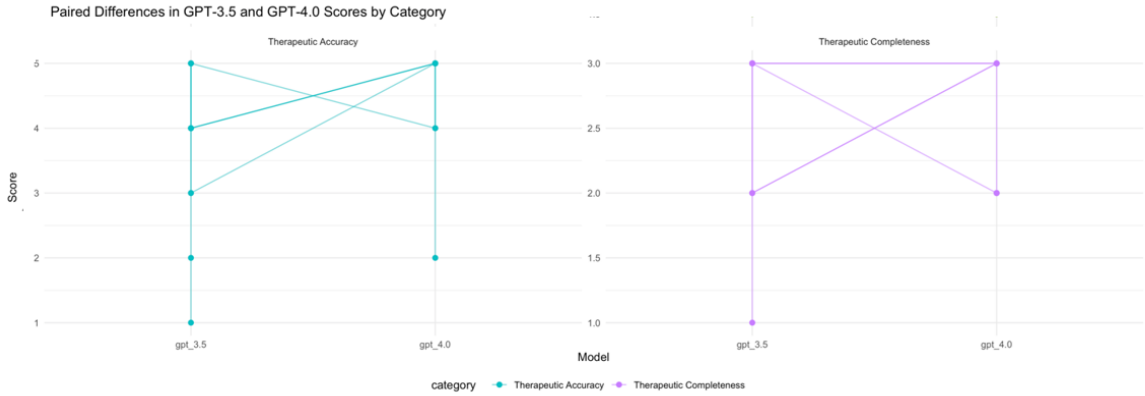


Figure 1: Paired differences in ChatGPT 3.5 and ChatGPT 4.0 scores by category. The paired line plot illustrates the differences between ChatGPT 3.5 and ChatGPT 4.0 scores across individual cases for each category. Each line represents a paired observation for the same case, linking the scores from ChatGPT 3.5 and ChatGPT 4.0, with the x-axis indicating the model and the y-axis representing the score.

The lines in therapeutic accuracy exhibit an upward trend indicating that ChatGPT 4.0 tends to outperform ChatGPT 3.5, but with greater variability. This aligned with the statistical analysis which showed a higher mean score for ChatGPT 4.0 but did not reach statistical significance reflecting insufficient evidence for a meaningful difference.

A more balanced distribution is shown in therapeutic completeness with lines showing both upward and downward trends, indicating that there is no clear advantage for either model in this category. The lack of systematic trends in the lines corresponded to the negligible and non-significant mean difference between the models in the statistical analysis.

Overall ChatGPT 4.0 demonstrated slightly better therapeutic accuracy without statistical significance. However, the greater variability observed in therapeutic categories suggests less consistent performance across cases.

Difference plots

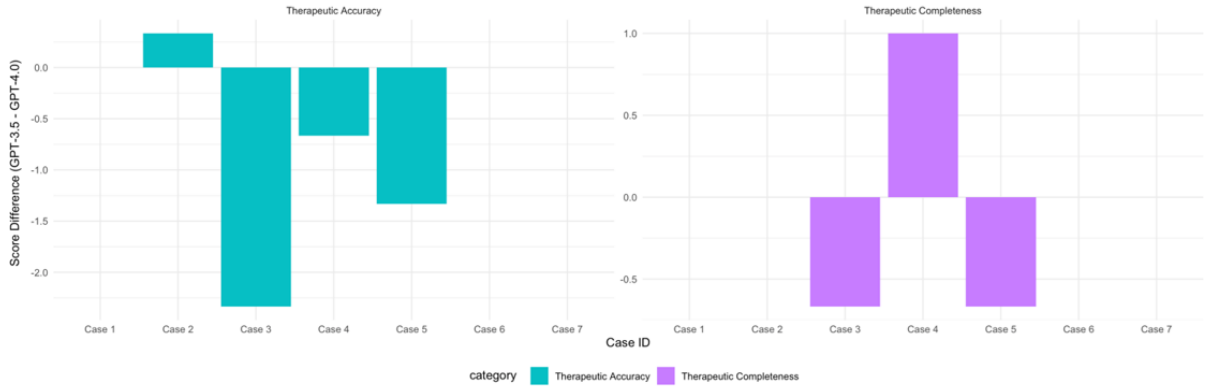


Figure 2: Differences between ChatGPT 3.5 and ChatGPT 4.0 scores by category. The difference plot highlights the variability in performance between ChatGPT 3.5 and ChatGPT 4.0 across individual cases within each category. Positive bars indicate higher scores for ChatGPT 3.5, while negative bars indicate higher scores for ChatGPT 4.0.

Therapeutic accuracy revealed mixed results, with several cases showing negative differences where ChatGPT 4.0 outperformed ChatGPT 3.5, reflecting its relative strength in therapeutic cases. Especially cases 3, 4 and 5 showed particularly large negative differences, highlighting ChatGPT 4.0 superior performance for these cases.

The differences in therapeutic completeness were more balanced with positive bars for some cases and negative bars for others. Case 4 showed a higher score for ChatGPT 3.5, while case 3 and 5 showed higher scores for ChatGPT 4.0. The relatively small magnitudes of the differences and their mixed directions reflected the statistical finding that there is no significant difference between the models in this category.

Overall, the plot emphasizes ChatGPT 4.0's strength in therapeutic accuracy, with a general consistency in completeness scores.

Discussion

Overall findings

ChatGPT demonstrated good accuracy in therapeutic cases, with a mean score of 3.867 for ChatGPT 3.5. The score for ChatGPT 4.0 was even higher, with a mean score of 4.667, indicating more accurate predictions than ChatGPT 3.5. For therapeutic completeness, the mean scores were 2.667 and 2.733 for ChatGPT 3.5 and ChatGPT 4.0, respectively.

In addition to examining its therapeutic capabilities, research is also being conducted on the diagnostic capabilities of the chatbot. Parallel to this study, a related study is conducted by the same research team, focusing on the accuracy and completeness of responses in diagnostic questioning. (16)

Overall, in this study the chatbot seems to perform well in answering therapeutic questions, often providing comprehensive recommendations that consider comorbidities effectively. However, in cases where it received lower scores, errors were primarily related to medication advice. Its recommendations did not always align with Belgian guidelines, which served as the basis for reviewer assessments. The lack of clarity regarding the chatbot's sources raises concerns about potential harm, making it challenging to rely on in clinical practice.

Other studies analyzing chatbot-generated medication advice highlight similar concerns. For example, a study evaluated ChatGPT's performance in clinical pharmacy practice comparing it to human pharmacists. (17) While ChatGPT performed well in drug counseling, it was significantly weaker in prescription review, patient medication education, adverse drug reaction (ADR) recognition, and ADR causality assessment. The AI demonstrated strong retrieval, information integration, and conversational abilities but struggled with complex reasoning, situational awareness, and personalized responses. Another similar study yielded more positive results. (18) The latter evaluated ChatGPT's ability to assist in medication therapy management by identifying drug interactions and providing management plans for complex clinical cases. Results showed that ChatGPT accurately identified drug interactions and offered reasonable management plans, though limitations existed in recommending alternative therapies and specific dosages.

In this current study, ChatGPT showed potential in providing accurate and comprehensive medical information. However, it still falls short of being entirely reliable. Numerous small-scale studies across various (sub)disciplines have already explored the potential of AI in managing different diagnoses. (19–23) Overall, based on the existing studies, it can be concluded that ChatGPT shows promise in providing accessible medical guidance but has limitations in accuracy, personalization, and handling complex cases. While it moderately aligns with clinical guidelines, it often lacks the nuanced judgment of experienced professionals. Caution is necessary, and its use should be supplemented by expert consultation, especially for severe or complex cases, until future models improve.

Similarly, a study evaluated ChatGPT's accuracy and completeness in answering 284 medical questions across various specialties. (24) Physicians rated responses on a Likert scale, finding that ChatGPT

provided largely accurate (median accuracy score: 5.5/6) and comprehensive (median completeness score: 3/3) answers. Accuracy varied slightly based on question difficulty but was similar for binary and descriptive questions. Notably, re-querying low-scoring answers after 8–17 days showed substantial improvement. While ChatGPT demonstrated strong performance, the study also highlights the need for further research and model development to address inaccuracies and enhance reliability in medical applications.

Reproducibility

Regarding reproducibility, it is important to note that identical prompts presented at different points in time can still yield varying responses from the chatbot. This variation may be reduced by configuring the chatbot to operate with a low temperature setting, which can enhance the consistency of its outputs. However, this parameter was not explored in the current study and default temperature was used.

The temperature setting in ChatGPT affects how creative or predictable its responses are. Low temperatures (0.0–0.3) produce more accurate and consistent answers, which is essential in medical contexts where reliability is critical. (4) Higher temperatures (0.7–1.0) lead to more creative but less predictable output, which may be unsuitable for healthcare. While temperature can be adjusted in ChatGPT 3.5 and ChatGPT 4 via the OpenAI API, it cannot be manually changed in the standard ChatGPT application, where it is set automatically by OpenAI.

ChatGPT 3.5 versus 4.0

Based on the current statistical analyses, ChatGPT 4.0 showed overall improvements in therapeutic cases compared to ChatGPT 3.5, with a higher mean score indicating more complete and more accurate responses. Studies comparing the clinical-decision-making abilities of ChatGPT 3.5 and ChatGPT 4.0 are limited. One such comparative study evaluated both models' responses to 176 real-world clinical questions, scoring them on accuracy, relevance, clarity, utility, and comprehensiveness and showed that ChatGPT 4.0 outperformed GPT 3.5 across all rating criteria. (25) Another study evaluated both ChatGPT versions in generating recommendations on critical care topics and concluded that ChatGPT 4.0 outperformed 3.5 in appropriateness, but there was no significant difference in consistency or readability. (26) Both models produced misinformation (“hallucinations”) and inconsistent responses, highlighting the risks of using LLMs without clinical oversight. It should be noted that the latter studies were conducted in secondary care settings. Primary care physicians and specialists tend to approach their patients from different perspectives, often following different sets of guidelines. There is a lack of research on ChatGPT's therapeutic and clinical-decision-making capabilities in primary care. Further studies are needed in this area before it can be considered applicable in this context.

Limitations

Despite the promising results of this current study, the scope of the conclusion is restricted by several limitations.

One major issue is the underpowered sample size and a lack of sufficient statistical expertise, which hinder robust analysis. Additionally, the absence of a formal validation mechanism to objectively assess the chatbot's accuracy introduces potential bias, as the evaluation relied on a self-developed Likert scale based on subjective ratings by physicians.

In line with this, another important consideration is interobserver bias, as two independent raters assessed each case, with a third assessor involved in cases of disagreement. Differences in strictness or leniency among raters could have influenced the score results. Interestingly, human errors—diagnoses overlooked by doctors but identified by the chatbot—were excluded from the scoring method, potentially impacting the overall evaluation.

The structured case templates employed in this study (Addendum 1) explicitly require that responses be formulated in accordance with current medical guidelines. These templates were developed by the researchers themselves, incorporating elements deemed essential to answering the study's research questions. Key components included the chief complaint, patient gender, nationality, medical history, medication use, clinical findings and technical investigations. However, whether this approach represents the optimal method for structuring case templates for ChatGPT remains uncertain.

This study reveals the importance of including a precise diagnosis when posing therapeutic questions to ChatGPT. One case within this study (case 2: Sialolithiasis) was characterized by an insufficiently defined diagnosis, which led to an overly broad response from the chatbot. This lack of clarity may have contributed to inconsistencies in the quality of responses and potentially compromised the chatbot's ability to generate accurate and comprehensive outputs. As a result, the accuracy and completeness scores in this study may have been adversely impacted. However, no standardized guidelines currently exist on how such templates should be constructed for chatbot-based medical questions.

Furthermore, the templates may not fully reflect real-world patients' queries. In practice, patients often lack the medical knowledge necessary to provide detailed and structured clinical information. As a result, chatbot responses in real-world scenarios may differ from those obtained using our predefined templates. This discrepancy highlights the need for further research into optimal structuring of case templates to ensure that they align more closely with real-life medical consultations.

A final important limitation of this study is the rapid and unprecedented pace at which the field of artificial intelligence is advancing. While these developments present exceptional opportunities, they also raise critical concerns regarding the long-term implications of AI in both research and practical applications. The accelerating evolution of AI models often exceeds the capacity of regulatory frameworks to keep pace. Consequently, the speed of progress frequently leads to the rapid

obsolescence of existing data.
Whereas ChatGPT 3.5 constituted the standard freely accessible version at the onset of the study, this has transitioned to a limited version of ChatGPT 4.0 in 2024.

Overall conclusions

Despite the great potential of LLMs such as ChatGPT, it is important to note that their integration into clinical practice and health-related services comes with significant technical, ethical, and regulatory challenges.

The reliability of LLMs is crucial in healthcare, where even minor errors can have serious consequences for patients. Concerns about their reliability stem from factors such as accuracy, consistency, interpretability, and the quality of the training data.

A major challenge is the "black box" nature of LLMs. (27,28) These models rely on complex neural networks with billions of parameters, and unlike traditional algorithms with explicit rules, they generate responses based on probabilistic patterns learned from vast datasets rather than engaging in true comprehension. In clinical settings, this lack of transparency is problematic, as medical decisions must be based on reliable and explainable reasoning. If an LLM misinterprets test results or provides inaccurate recommendations, there is often no clear way to audit or validate its decision-making process.

Another critical issue is the risk of hallucinations, where LLMs generate plausible but entirely false information. Unlike human experts who can verify claims against real-world knowledge, LLMs lack built-in fact-checking mechanisms which can be very dangerous in healthcare, where inaccurate medical advice or misinformation can have life-threatening consequences. (27–29) Moreover, LLMs can sometimes produce inconsistent answers when given slightly different prompts or contexts. This inconsistency is a result of how these models predict the most likely sequence of words based on patterns in the data, without having true comprehension or understanding. These variations in responses can make it difficult for healthcare professionals and patients to rely on these models for decision-making, highlighting the need for external validation when using LLMs in healthcare.

The quality and timeliness of the data used to train LLMs further complicate their reliability. (27,30,31) Many widely used models, including ChatGPT 3.5, are trained on outdated information, with some having a knowledge cutoff as far back as 2021. This makes them unreliable for accessing the latest medical advancements, a significant drawback in a field where rapid progress is constant. Moreover, much of the training data is sourced from publicly available websites rather than peer-reviewed, professionally validated materials. This increases the risk of incorrect or untrustworthy outputs, further underscoring the need for improved data curation and model transparency.

Bias is a significant issue when it comes to the application of LLMs in healthcare. (27,31) These models are trained on vast amounts of data generated by humans, which inherently includes human biases. The training data often reflects societal inequalities, such as gender, racial, and socioeconomic biases, which the models then learn and reproduce.

LLM biases can emerge at different stages. (30) First, there is bias in the data itself, such as underrepresentation of certain groups in medical research or the use of biased medical terminology. Second, the model's outputs can be biased due to the algorithms it uses, which may prioritize certain perspectives based on the structure of the training data. Additionally, the user base can influence biases, as healthcare providers or patients may unknowingly input data that steers the model toward biased outcomes. To mitigate these biases, it is essential to incorporate diverse and representative datasets and develop monitoring systems to detect and address bias in LLM outputs. One potential solution is the creation of domain-specific models, trained on more inclusive and specific medical data, which could help reduce these biases and improve the model's fairness and accuracy.

Privacy and security concerns are among the most pressing issues when using LLMs in healthcare. (27,32) Given the sensitive nature of health-related information, breaches of confidentiality could have serious legal and ethical consequences. Currently, most publicly available LLMs do not guarantee anonymity, which increases the risk of unauthorized access to patient data. The use of third-party LLMs, such as ChatGPT, to manage patient interactions introduces the possibility of inadvertently disclosing personal health information during processing. To address these concerns, healthcare organizations may need to implement strict protocols for anonymizing patient data before using it with LLMs. Alternatively, they may opt to use in-house, domain-specific models that do not require sharing external data. (32)

Medico-Legal complications could also arise by the lack of accountability for AI-generated medical advice, particularly if patients rely on inaccurate or harmful information. (33) Therefore, it is important not to blindly adopt the information provided by AI-generated systems like ChatGPT, but rather to approach these answers with a critical mindset and a solid knowledge of the subject.

Public acceptance and trust are also significant barriers to the widespread adoption of LLMs in healthcare. While some patients and healthcare professionals trust AI tools like ChatGPT for medical information, skepticism remains widespread. (30) A key factor contributing to this distrust is the lack of transparency about how these models operate and the fear of relying on technology that may produce inaccurate or biased results. Additionally, concerns about the security of personal health data further undermine confidence in LLMs. For LLMs to gain widespread adoption, both healthcare professionals and patients need to be educated about their benefits and risks. (29) It is crucial to provide proper training on using these tools safely and responsibly, ensuring that they complement human decision-making rather than replace it. Medical education should also include training on the limitations of AI tools and emphasize the importance of external validation when using these models.

Despite concerns surrounding their use, LLMs offer numerous potential benefits to the healthcare system. With further validation, LLMs have the potential to serve as valuable assets for quick medical information retrieval in fast-paced healthcare settings, improving healthcare efficiency and supporting complex decision-making. (29) While LLMs are not yet reliable enough for independent clinical decision-making, they can still aid healthcare providers by suggesting possible diagnoses or treatment options, which can be further validated by the clinician. (34) This can serve as a valuable tool for enhancing clinical decision support. LLMs can also be used to generate patient-friendly materials tailored to an individual's health literacy level and preferred language. (27,29) This helps improve

patient understanding and engagement, empowering them to take an active role in their own care. Additionally, LLMs can assist with administrative tasks by automating processes like generating clinical documentation, summarizing patient notes, or triaging patient queries. (32) This can reduce the administrative burden on healthcare professionals and improve efficiency within healthcare.

In conclusion, it is highly likely that LLMs, such as ChatGPT, will play a pivotal role in the evolution of medical decision-making within primary care. Currently, GP predominantly rely on the Evidence Linker to access relevant information. The Evidence Linker is an integrated tool within Electronic Health Records (EHR) that provides clinicians with evidence-based medical information during patient consultations. The potential of AI to enhance or even replace such tools lies in its ability to facilitate more rapid, context-specific data retrieval, offer real-time decision support, and continuously assimilate the latest medical research.

Further research should be focused on several key areas to maximize the potential of LLMs in healthcare.
Training LLMs on domain-specific datasets, including clinical notes, medical literature and even patient data could significantly enhance the accuracy and completeness in primary care medicine and prevent hallucinations. (27,30,34)
Rigorous clinical trials should be conducted to assess the performance and effectiveness of LLMs in improving healthcare outcomes. These trials should evaluate their impact on patient safety, diagnostic or therapeutic accuracy and completeness. Research into identifying and mitigating biases in LLMs is essential. It is crucial that these models do not perpetuate existing healthcare disparities, particularly with regard to race, gender, and socioeconomic status. Addressing biases will help ensure that LLMs provide equitable care and improve outcomes for all patients' populations. (27,30)

Conclusion

ChatGPT demonstrated good therapeutic accuracy and completeness for the presented medical cases. In this study ChatGPT 4.0 outperforms ChatGPT 3.5.

However, these differences were not statistically significant. The small sample size of this study suggests caution when interpreting these results, as it may introduce bias. As stated earlier, there are still numerous limitations of the chatbot itself. To improve the reliability and generalizability of these findings, future research is necessary.

Acknowledgments

We would like to express our gratitude to our promoters, Prof. Dr. Aertgeerts B. and Prof. Dr. Delvaux N. Thank you for the inspiration, advice and guidance through the steps of this project.

We would also like to thank the lectors for reading and evaluating this thesis.

Furthermore, we would like to thank and acknowledge Kerkhofs L., Stroobants N. and Vanpoecke S., who worked on a parallel thesis regarding the accuracy and completeness of ChatGPT in relation to diagnostic cases. We thank them for their smooth collaboration, which resulted in two theses providing valuable insights into both therapeutic and diagnostic challenges.

Special recognition goes to Pavlos Mamouris, who we thank for his insights on this project and helping us with the statistical analysis.

Conflicts of interest

None.

Financial Disclosure

The only financial support our thesis required was a one-month subscription to ChatGPT 4.0, which the researchers financed themselves.

References

1. Wiggers K, Corrall C, Silberling A, Park K. ChatGPT: Everything you need to know about the AI-powered chatbot [Internet]. 2023 [cited 2023 May 31]. Available from: https://techcrunch.com/2023/05/31/chatGPT-everything-you-need-to-know-about-the-open-ai-powered-chatbot

2. Ortiz S. What is ChatGPT and why does it matter? Here's everything you need to know [Internet]. 2024 [cited 2025 Apr 8]. Available from: https://www.zdnet.com/article/what-is-chatgpt-how-the-worlds-most-popular-ai-chatbot-can-benefit-you/

3. Amazon Web Services. What is NLP? - Natural Language Processing Explained [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://aws.amazon.com/what-is/nlp/

4. OpenAI. OpenAI [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://openai.com

5. Hughes A. GPT-3: Everything you need to know about OpenAI's GPT-3 tool [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://www.sciencefocus.com/future-technology/gpt-3

6. Ackerson N. ChatGPT insists I am dead: The problem with language models [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://towardsdatascience.com/chatgpt-insists-i-am-dead-and-the-problem-with-language-models-db5a36c22f11/

7. Weintraub K. ChatGPT is poised to upend medical information. For better and worse. [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://eu.usatoday.com/story/news/health/2023/02/26/chatgpt-medical-care-doctors/11253952002/

8. Keymanthri M, Rennie S. ChatGPT has many uses: Experts explore what this means for healthcare and medical research [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://theconversation.com/chatgpt-has-many-uses-experts-explore-what-this-means-for-healthcare-and-medical-research-200283

9. Will ChatGPT transform healthcare? Vol. 29, Nature Medicine. Nature Research; 2023. p. 505–6.

10. Culture Geek. ChatGPT: What is it and how does it work exactly? [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://medium.com/geekculture/chatgpt-what-is-it-and-how-does-it-work-exactly-62e7010524d3

11. Pentalog. ChatGPT fundamentals [Internet]. 2023 [cited 2025 Apr 8]. Available from: https://www.pentalog.com/blog/tech-trends/chatgpt-fundamentals/

12. BCFI. Belgisch Centrum voor Farmacotherapeutische Informatie [Internet]. [cited 2025 Apr 8]. Available from: www.bcfi.be

13. Nederlands huisartsen genootschap. NHG [Internet]. [cited 2025 Apr 8]. Available from: www.nhg.org

14. Domus Medica. Domus Medica [Internet]. [cited 2025 Apr 8]. Available from: www.domusmedica.be

15. UpToDate. UpToDate Clinical Decision Support [Internet]. [cited 2025 Apr 8]. Available from: www.uptodate.com

16. Kerkhofs L, Stroobants N, Vanpoecke S. The accuracy and completeness of ChatGPT in Belgian general practice: a descriptive analysis of diagnostic cases. 2025;

17. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol. 2024 Jan;90(1):232–8;

18. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003). 2024;64(2):422-428.e8.

19. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canouï E, et al. Can Chatbot Artificial Intelligence Replace Infectious Diseases Physicians in the Management of Bloodstream Infections? A Prospective Cohort Study. Clin Infect Dis. 2024 Apr 10;78(4):825–32.

20. Dagher T, Dwyer EP, Baker HP, Kalidoss S, Strelzow JA. "Dr. AI Will See You Now": How Do ChatGPT-4 Treatment Recommendations Align With Orthopaedic Clinical Practice Guidelines? Clin Orthop Relat Res. 2024 Dec 1;482(12):2098–106.

21. Saturno M, Mejia MR, Ahmed W, Yu A, Duey A, Zaidat B, et al. Can generative artificial intelligence provide accurate medical advice?: a case of ChatGPT versus Congress of Neurological Surgeons management of acute cervical spine and spinal cord injuries clinical guidelines. Asian Spine J. 2025 Mar 4;

22. Lazris D, Schenker Y, Thomas TH. AI-Generated Content in Cancer Symptom Management: A Comparative Analysis Between ChatGPT and NCCN. J Pain Symptom Manage. 2024 Oct;68(4):e303–11.

23. Kahlon S, Sleet M, Sujka J, Docimo S, DuCoin C, Dimou F, et al. Evaluating the concordance of ChatGPT and physician recommendations for bariatric surgery. Can J Physiol Pharmacol. 2025 Feb 1;103(2):70–4.

24. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023 Feb 28;

25. Lahat A, Sharif K, Zoabi N, Shneor Patt Y, Sharif Y, Fisher L, et al. Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4. J Med Internet Res. 2024 Jun 27;26:e54571.

26. Balta KY, Javidan AP, Walser E, Arntfield R, Prager R. Evaluating the Appropriateness, Consistency, and Readability of ChatGPT in Critical Care Recommendations. J Intensive Care Med. 2025 Feb;40(2):184–90.

27. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–40.

28. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton EW, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. medRxiv. 2024 Apr 27;

29. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023 Oct 2;6(10):e2336483.

30. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. J Med Internet Res. 2024 Nov 7;26:e22769.

31. 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 Feb 15;20(4).

32. Matulis J, McCoy R. Relief in Sight? Chatbots, In-baskets, and the Overwhelmed Primary Care Clinician. J Gen Intern Med. 2023 Sep;38(12):2808–15.

33. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595.

34. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health. 2024 Jan 30;12(Suppl 1).

Approval research ethics committee

Artificial intelligence, in de vorm van Chat-GPT, een troef in de klinische praktijk? Een vergelijking met EBM richtlijnen voor therapeutische vragen.

Dossiernr.:	MP028461
Masterproef:	Masterproef I HAIO, 2022-2025
Faculteit/opleiding:	Faculteit Geneeskunde - Master in de huisartsgeneeskunde
Studiefase:	eerste fase
Academiejaar:	2023-2024
Korte omschrijving / abstract:	**Deze literatuurstudie onderzoekt de bruikbaarheid en accuraatheid van ChatGPT, een AI gestuurde chatbot op basis van een large language model, in de huisartsenpraktijk. Dit wordt onderzocht aan de hand van fictieve casussen die worden getoetst aan het programma. De antwoorden worden beoordeeld op volledigheid en accuraatheid. Er wordt gekeken naar de toepasbaarheid in de praktijk, en of er geen belangrijke zaken gemist worden.**
Gewenste taal van communicatie:	nl
Promotor:	Prof Bert Aertgeerts (u0002409)
Copromotor:	Nicolas Delvaux (m9816943)
Organisatie:	KU Leuven
Dienst:	Academisch Centrum voor Huisartsgeneeskunde
Discipline:	Masterproef
Extra groepsleden:	Emma Reynders emma.reynders1@student.kuleuven.be r0611238 Aline van Bijsterveld aline.vanbijsterveld@student.kuleuven.be r0677547 Nathalie Devolder nathalie.devolder@student.kuleuven.be r0636540

De verstrekte informatie toont aan dat het onderzoek in het kader van de masterproef een zuivere literatuurstudie betreft en dat er op geen enkele wijze gebruik wordt gemaakt van proefdieren, proefpersonen of menselijk lichaamsmateriaal.

Dit type onderzoek vereist geen goedkeuring van een gemachtigde commissie voor medische ethiek.

Onder voorbehoud van de volledigheid en correctheid van de door u verstrekte gegevens, krijgt u hierbij het akkoord om het onderzoek in het kader van de masterproef te starten. Deze mail dient als bewijs van goedkeuring mocht u uw onderzoek wensen te publiceren.

Wij willen u erop attent maken dat u zelf verantwoordelijk blijft voor uw onderzoek. Bovendien doet elke wijziging aan de door u verstrekte gegevens omtrent de onderzoeksopzet deze goedkeuring vervallen. U dient in dat geval een amendement te maken aan uw huidig dossier.

Addenda

Addendum 1: Template

A Belgian #age year old #sex, visits the general practitioner with #complaint.

Symptoms and complaints:

Physical examination:

Technical examination:

Medical history:

Medication use:

Give the possible treatment options and the preferred treatment. Be specific and incorporate any applicable guidelines.

Addendum 2: Likert Scale

	THERAPEUTIC **(Control= guideline)**
Accuracy score	1. >1 medical error (*) 2. 1 medical error and further correct/more incorrect than correct information 3. > 50% correct 4. >70% correct 5. >90% correct (+-completely consistent with the guideline, nuance)
Completeness score	1. <50% complete (**) 2. 50-75% complete 3. >75% of all components listed, possibly +- additional proposals

(*) Medical error: advice that could potentially harm the patient

(**) Components of completeness/ accuracy therapeutic: non-drug therapy, drug therapy, indications for refer

Addendum 3: All scores by reader

Therapeutic cases

Score accuracy with chatgpt 3.5

	Reading 1			Reading 2			Reading 3		
	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.
Case 1	5	5		4	4		5	5	
Case 2	4	4		4	5	5	4	4	
Case 3	1	1		2	2		4	4	
Case 4	4	4		5	5		4	4	
Case 5	5	5		3	5	3	3	5	3

Score completeness with chatgpt 3.5

	Reading 1			Reading 2			Reading 3		
	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.
Case 1	3	3		3	3		3	3	
Case 2	3	3		3	3		3	3	
Case 3	1	1		3	3		2	2	
Case 4	3	3		3	3		3	3	
Case 5	2	2		3	3		2	3	2

Therapeutic cases

Score accuracy with chatgpt 4.0

	Reading 1			Reading 2			Reading 3		
	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.
Case 1	4	4		5	5		5	5	
Case 2	5	5		2	4	2	5	5	
Case 3	5	5		4	4		4	5	5
Case 4	5	5		5	5		5	5	
Case 5	4	5	5	4	5	5	5	5	

Score completeness with chatgpt 4.0

	Reading 1			Reading 2			Reading 3		
	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.	V.B.A.	S.N.	K.L.
Case 1	3	3		3	3		3	3	
Case 2	3	3		3	3		3	3	
Case 3	3	3		3	3		2	3	2
Case 4	3	2	2	3	2	2	3	2	2
Case 5	3	3		3	3		3	3	

Addendum 4: All cases with answers

See separated files