Mayra Spizzo
Engehaldenstrasse 8, 3012 Bern
mayra.spizzo@unibe.ch

Data Science Project

Unravelling the Human Interaction with Generative AI-Based Decision Support in Healthcare

Conceptual Design Report

6 October 2024

Abstract

Artificial intelligence (AI) systems to augment the diagnostic process and to reduce diagnostic errors are increasingly implemented in the field of medicine. Likewise, generative AI systems are progressively used by physicians in their decision-making process. When it comes to generative AI systems, the quality of the system's output relies highly on the input and the interaction of the user with the system. However, the actual usage of a generative AI system in this context, including underlying reasons and patterns in the interaction, remains so far unknown. In this project, chat interactions of individuals with ChatGPT and a human expert from an online experiment are investigated in order to understand how individuals interact with ChatGPT in a medical context. In the online experiment, individuals solve two diagnostic tasks and can use the chat as a support in their decision-making process. For this project, a final dataset of 316 observations and 14 variables will be generated from the data of the experiment. Unsupervised clustering models will then be applied, in order to detect patterns in the interactions and to generate clusters of individuals with similar interaction behavior.

Table of Contents

1 Project Objectives

In general, 5-15% of patients receive an incorrect diagnosis, because the physician made an error in the diagnostic decision-making process[1][2][3]. The consequences of this incorrect decision are crucial[4] - for example, the patient gets the wrong treatment, has to stay in the hospital for a longer time, and possibly has to endure pain until the diagnostic error is discovered and corrected. To reduce this number of diagnostic errors, AI-based systems are increasingly being implemented to support the decision-making process of physicians. Also the generative AI tool ChatGPT shows potential in augmenting diagnostic decisions of physicians[5]. However, when it comes to generative AI tools, the input of the user in the system is an important factor that determines the quality of the output of the system which, in turn, influences the support to reduce diagnostic errors. In the interaction with a generative AI tool, the user does not just evaluate the system's output but can steer the process of the output generation through the repeated interaction with the generative AI tool and, with its input, influence the quality of the output.

Research evaluating the potential of ChatGPT to support the diagnostic decision-making process mainly focuses on the output of the decision of a physician[6]. Consequently, a thorough understanding of how physicians interact with the generative AI is missing, besides the importance of the interaction for the quality of the system's output. Therefore, the question arises: How does a user interact with generative AI in medical diagnostics? Furthermore, are there groups of individuals with similar interaction patterns?

To examine this research question, the data from an experiment conducted with *N* = 158 fourth-year medical students is used. In the online experiment, students are asked to generate differential diagnoses and a final diagnosis for two patient cases. During the task, they can use a chat to ask questions as support. One randomly assigned group of participants has a physician as their interaction partner replying in real time, while the other group has ChatGPT (version gpt-4-0613) as their chat partner. Throughout the task, all clicks on patient information, chat interactions, and noted differential diagnoses are logged with timestamps.

The collected interaction parameters, such as, the type of questions being asked in the chat, the discreteness of the first question, the information acquisition prior to the beginning of the interaction, the average accuracy of the differential diagnoses, and the accuracy of the final

[1] Berner & Graber (2008)
[2] Newman-Toker et al. (2023)
[3] Singh et al. (2014)
[4] Hautz et al. (2019)
[5] Ferdush et al. (2024)
[6] Rao et al. (2023)

diagnosis are used as explanatory variables to detect patterns in the interactions and to generate clusters of individuals with a similar interaction behavior.

2 Methods

The data used for this project will also be used for additional studies. Therefore, the data cannot yet be made publicly available. The data collection is still in progress, after the data collection has been finished and the data analysis of the additional studies has been finalized, the anonymized data will be made available for download on the OSF platform[7]. Until then, the data will be stored locally on private devices. To code the chat interactions, the software MAXQDA[8] is used.

In order to find patterns in the chat interactions and generate clusters, unsupervised clustering models will be used. In the first step, the K-means clustering technique will be implemented. To reduce the dimensions of the dataset, a principal component analysis (PCA) will be performed. Different numbers of clusters are tested, and to assess the quality of the clusters, the silhouette score will be calculated and compared for the different models. In the next step, other clustering techniques will be tested, depending on the distribution of the data after the PCA (see Figure 1). In particular, the clustering models in each step will be generated for participants in the ChatGPT condition, in the human coach condition, and for both conditions together. With this, the clusters can be compared between the conditions, and it can be assessed whether the interaction patterns differ, depending on the assistant in the chat.

The following python libraries will be used:

- Pandas: To clean and preprocess the data.
- Numpy: To perform array operations for the needed variables.
- Matplotlib: To generate visualizations.
- Statsmodels.api and scipy: To perform statistical tests.
- Scikit-learn: To generate machine learning models.

[7] Center for Open Science (2024)

[8] MAXQDA (2024)

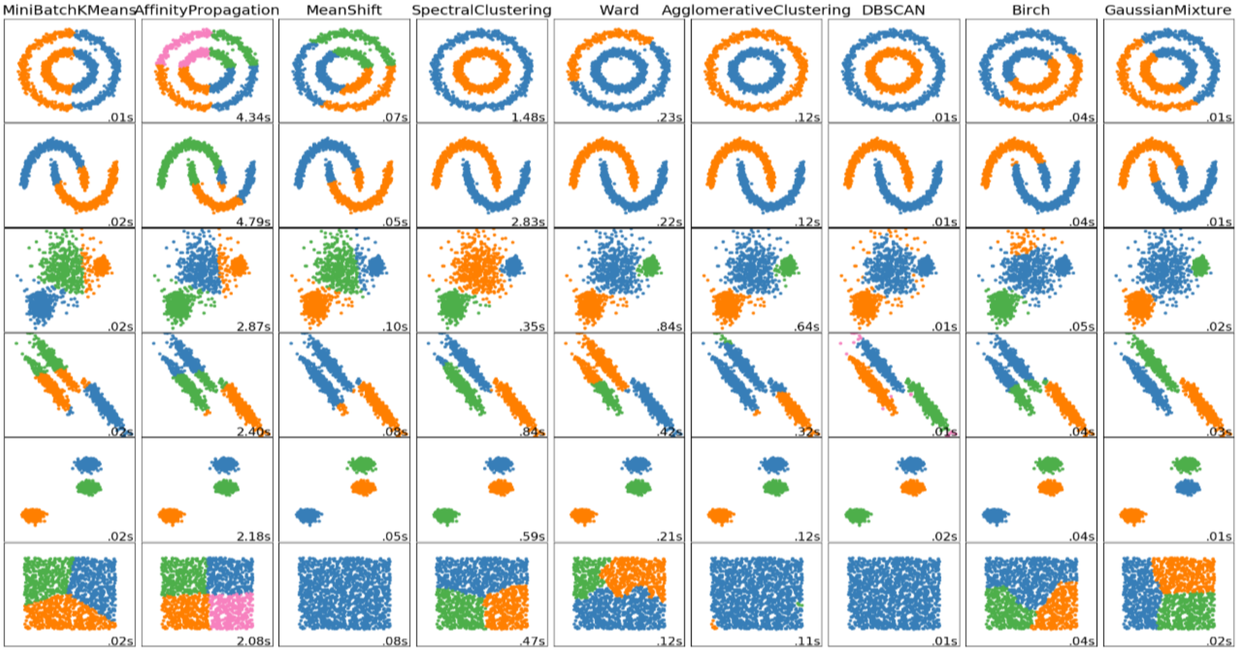


Figure 1: A comparison of different clustering techniques[9].

3 Data

The data of the experiment, which is described in further detail in chapter “6 Data Flow”, is used for the analyses of this project. Since the data is still being collected and is also used for additional studies, it cannot yet be published. For the descriptive analyses of this chapter, the data of the chat interactions that have been collected and coded so far was used (*N* = 36). The descriptive analyses for the accuracy of the differential and final diagnoses cannot yet be provided because the respective data will only be transmitted and made accessible to the author of this report when the data collection is completed and all participants finished the study.

To clean the data set, first all technical errors in the chat interaction logs are removed. During the recording of the chat interactions, on rare occasions, chat messages were recorded twice or a blank message was recorded (the reason for these errors was when the person sending the message clicked on “Send” before finishing the message or when he/she sent the message before writing anything). Second, after the coding of the chat interactions, the data set is controlled and it is checked that all chat segments are assigned to a code (i.e., it is assured that the coder did not miss coding a segment). In the next step, the qualitative codes of the segments are transformed to numeric values. A numeric value results for each message in the chat and for each coding category, indicating if a segment of the message was coded in the respective

[9] Seif (2018)

category (dummy-coded as 0 = no coded segment for this category and 1 = there is a coded segment for this category, respectively).

Afterward, the variables needed for this project are generated. An overview of the variables and the calculations made for each variable is provided in Table 1. The final dataset (after having finished the data collection) will contain 316 observations of 14 variables.

Table 1: Overview of the important variables for the project.

Variable	How it is calculated
Participant ID	These values are already existing and do not have to be calculated.
Case	Assignment: 1 = Case 1, 2 = Case 2
Condition	Binary variable: 0 = Human condition, 1 = ChatGPT condition
Type of question being asked in the chat: Technical questions	Amount of technical questions asked by the participant in the respective case. (Counting the amount of questions being asked in the chat messages of the respective case for each question type category, according to the coding scheme in Appendix 1)
Type of question being asked in the chat: Request	Amount of request questions asked by the participant in the respective case.
Type of question being asked in the chat: Statement	Amount of statements made by the participant in the respective case.
Type of question being asked in the chat: Differentiation	Amount of differentiation questions asked by the participant in the respective case.
Type of question being asked in the chat: Verification	Amount of verification questions asked by the participant in the respective case.

Type of question being asked in the chat: Diagnostics	Amount of diagnostic support questions asked by the participant in the respective case.
Type of question being asked in the chat: Management	Amount of management support questions asked by the participant in the respective case.
Discreteness of first question	Binary variable: Tells if the first question in the chat interaction is developed by the participant itself or if information from the patient record is copied and pasted into the chat; 0 = Copied information from the patient record, 1 = Self-developed question.
Information acquisition prior to the beginning of the interaction	The amount of patient information (in percent) the participant looked at before he/she asked the first question in the chat.
Average accuracy of differential diagnoses	The mean value of the proximities of the ICD-10 (International Classification of Diseases, 10th revision) codes of all differential diagnoses to the ICD-10 code of the correct diagnosis is calculated.
Accuracy of final diagnosis	The proximity of the ICD-10 code of the final diagnosis to the ICD-10 code of the correct diagnosis is calculated.

Regarding the distribution of the type of questions being asked in the chat (see Figure 2), questions for verification were asked the most (34.17%), followed by request questions (29.17%), statements (17.5%), technical questions (14.17%), and differentiation questions (5%). In the chat interactions that were coded so far, no questions of the categories diagnostics or management questions were noted.

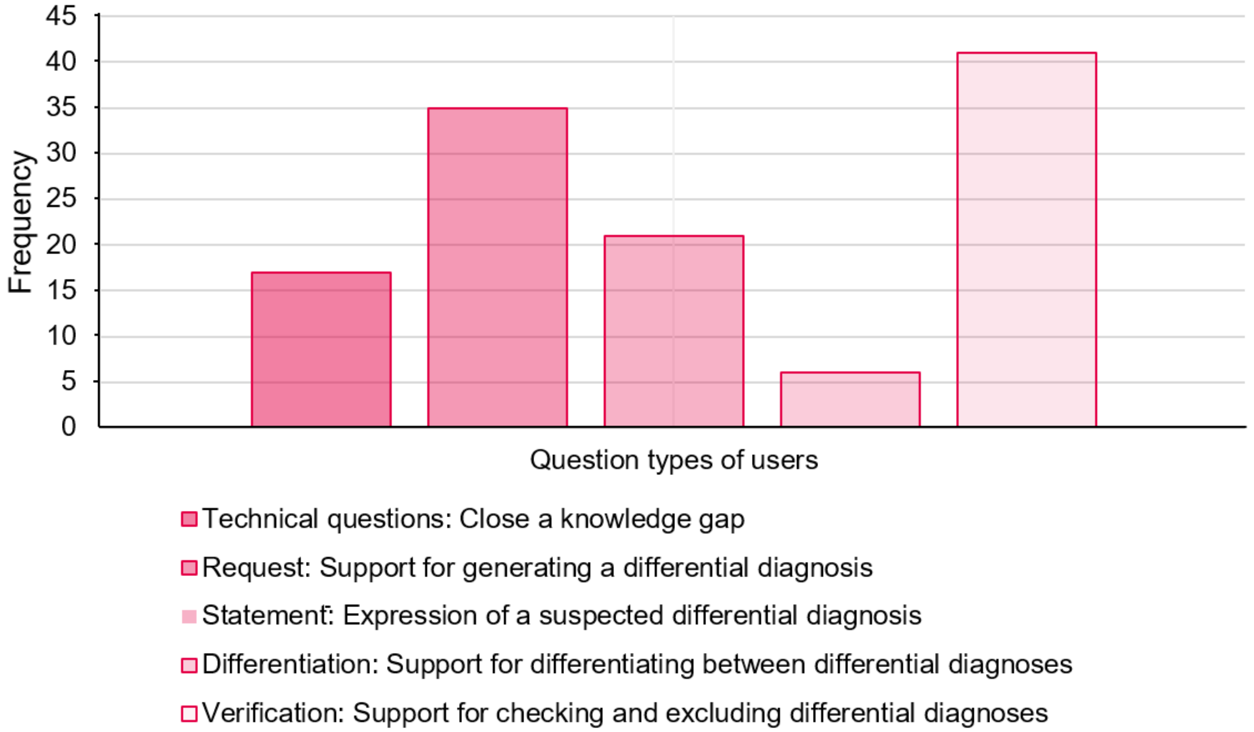


Figure 2: Distribution of question types being asked in the chats.

For an additional comparison of the question types, they were analyzed in an aggregated form. Since the question types technical questions and request questions are not based on an existing differential diagnosis or hypothesis of the correct solution, they can be assigned to the group “questions to build hypotheses”. The category statement does not include questions and was used as a category for neutral messages. The other two categories that include questions (i.e., differentiation questions and questions for verification) build on an existing differential diagnosis or hypothesis of the correct solution for the patient case and can therefore be allocated to the group “questions to test hypotheses”. The analysis of the two groups of questions showed that during one chat interaction, participants asked, on average, more frequently questions to test hypotheses (see Figure 3).

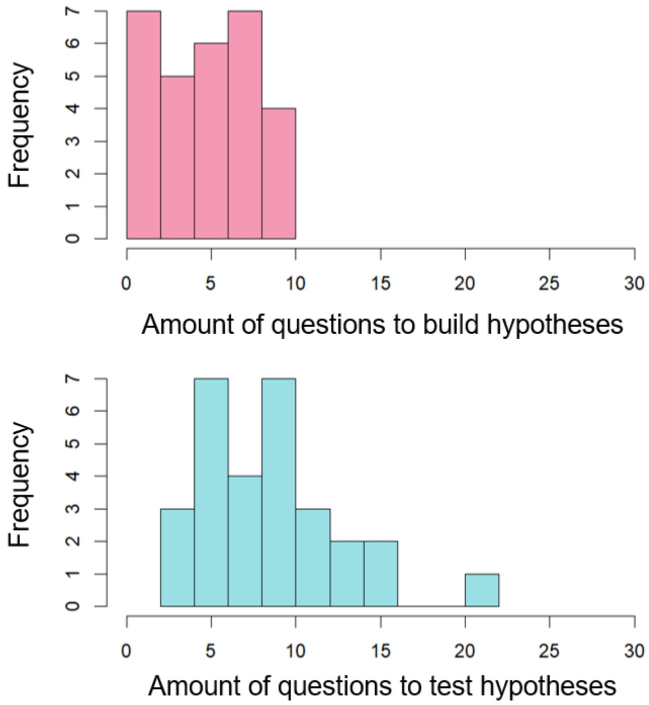


Figure 3: Distribution of the groups of questions being asked during one chat interaction.

When we look at the amount of patient information participants look at before they start the interaction in the chat, a U-shaped distribution is noticeable (see Figure 4). Participants tend to most frequently either look at very little information or at all the patient information before they initiate the chat interaction.

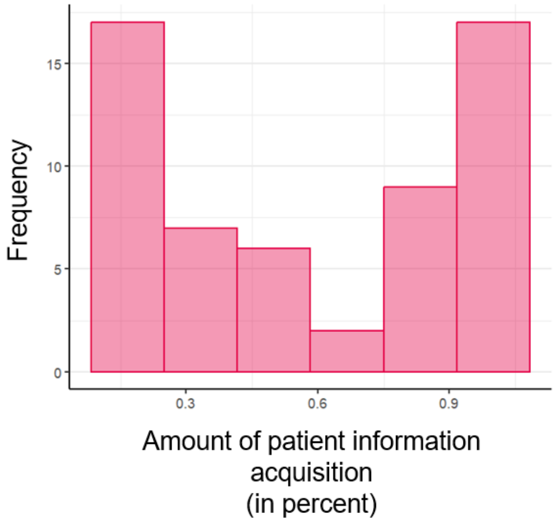


Figure 4: Distribution of the patient information acquisition prior to the beginning of the chat interaction.

4 Metadata

The information on the already collected data, including which data points from which datasets are already available, information about the participants, and timestamps of the recording of the respective data points are stored in an additional csv file. In this file, it is also noted with timestamps which chats were sent to the coders, which chats were received back with the codes, and which chats are prepared for analysis.

Since data is merged from multiple datasets, the variable “Participant ID” is used to match the observations of the different datasets.

As already mentioned, the datasets cannot yet be provided to the public and are currently stored as csv files locally on private devices. After the analyses of the additional studies are finalized, the data will be made available for download on the OSF platform. If someone wishes to access the data prior to the upload on OSF, he/she can contact the author of this conceptual design report to receive an anonymized version of the data.

5 Data Quality

To determine the size of the sample, the target sample size was calculated to be $N = 158$ using G*Power 3.1.9.7[10] for an analysis of variance to detect a medium effect size with $\alpha = .05$, and $\beta = .80$. In order to further increase the amount of observations, the task in the experiment consists not only of diagnosing one but two patient cases.

To further ensure the quality of the data, inclusion criteria for the participants were defined. The participation in the study is possible for all medical students at the Charité Medical School in Berlin that are in the fourth year of their studies ($N = 640$), are at least 18 years old, and have given their written consent for participation.

To assure the quality of the codings of the question types, the chat interactions of 20% of the participants were coded by two medical master students who were trained as coders, and the intercoder agreement was calculated. Detailed information about the intercoder agreement of the different question type categories can be found in Appendix 2. To ensure intercoder reliability, disagreeing coding segments will be discussed with the coders and afterward, the rest of the chat interactions are randomly divided between the two coders.

In general, the quality of the data can be classified as high, since the data used in this project consists of primary data that is collected by one institute of the emergency unit of the university

[10] Faul et al., 2007

hospital Inselspital. This institute also coordinates the experiment and prepared the required materials for it (e.g., the surveys, the acquisition and training of the human coaches, as well as the development of the task interface which are explained in more detail in chapter “6 Data Flow”).

6 Data Flow

The data flow can be divided into three process steps, namely the experiment, the data extraction, and the data analysis (see Figure 5).

Experiment

The online experiment is conducted with medical students from the Charité Medical School in Berlin. The data collection started on 22 April, 2024 and until 6 October, 2024 the answers of two thirds of the sample have been collected. A between-subjects design is used with source of assistance as a factor. The factor consists of the two levels human coach and ChatGPT. The conditions of the factor are randomly assigned.

The medical students participating in the study watch a general introduction video of large language models (LLMs) in the beginning. A short survey follows the video in which participants sign consent on the participation and answer baseline questions regarding their attitude and experience with AI tools, as well as demographic questions.

Afterward, the participants are randomly assigned to one condition (i.e., either to the human coach or ChatGPT). To start the diagnostic tasks, participants pass a get-to-know phase in which they get to know their respective chat assistant, as well as his strengths and weaknesses. The diagnostic tasks consist of diagnosing two patient cases that are presented in random order. The patient cases in each of the two diagnostic tasks are based on real emergency cases[11][12]. The equivocal cases have a correct diagnosis as the solution of the task, but also a main incorrect competing diagnosis. In one case, the correct diagnosis is pulmonary embolism with myocardial infarction as the incorrect competing diagnosis, and in the other case, the correct diagnosis is pulmonary embolism with stroke as the incorrect competing diagnosis, respectively.

Participants see information about the patient case (i.e., patient history, blood samples, laboratory results, medical imaging, and ECGs) by clicking on the equivalent tab in the interface (see Figure 6). Below the patient information, participants enter all differential diagnoses they consider during the diagnostic task. On the right side of the interface, participants have the possibility to chat with their assigned assistant in real time (i.e., a human coach or ChatGPT) as a support to solve the

[11] Kumar et al. (2011)
[12] Kunina-Habenicht et al. (2015)

diagnostic task. During the diagnostic tasks, all clicks, noted differential diagnoses, and chat interactions are logged with timestamps.

After each diagnostic task, participants choose the final diagnosis out of their differential diagnoses and fill in a survey regarding the perception of the difficulty of the case, the familiarity of the case, the competence of the assistant, and the support of the assistant. In the next step, participants complete a final survey that collects information regarding the perception of the usefulness, satisfaction, and credibility of the assistant. Finally, participants are informed about the debriefing.

Data Extraction and Data Analysis

Two researchers from the emergency unit of the university hospital Inselspital have access to the data from the experiment. Together with additional data received from the Charité Medical School about the participants (e.g., student records, grades of the last two semesters, etc.), they anonymize the data and create three datasets. Two datasets contain information gathered in the surveys and one dataset includes the relevant chat information for this project. By using the SWITCH filesender[13], the data is transmitted securely to the author of this conceptual design report. In the next step, the author cleans the data and coordinates the coding of the chats with the two medical master students before the data is analyzed and the results are interpreted.

[13] University of Bern (2024)

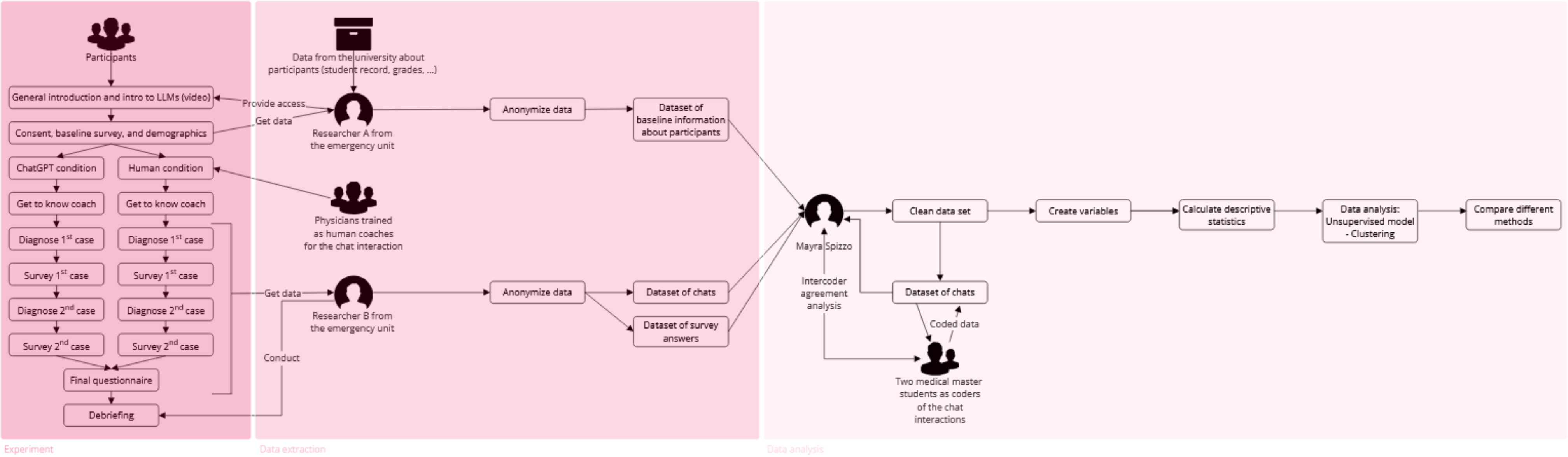
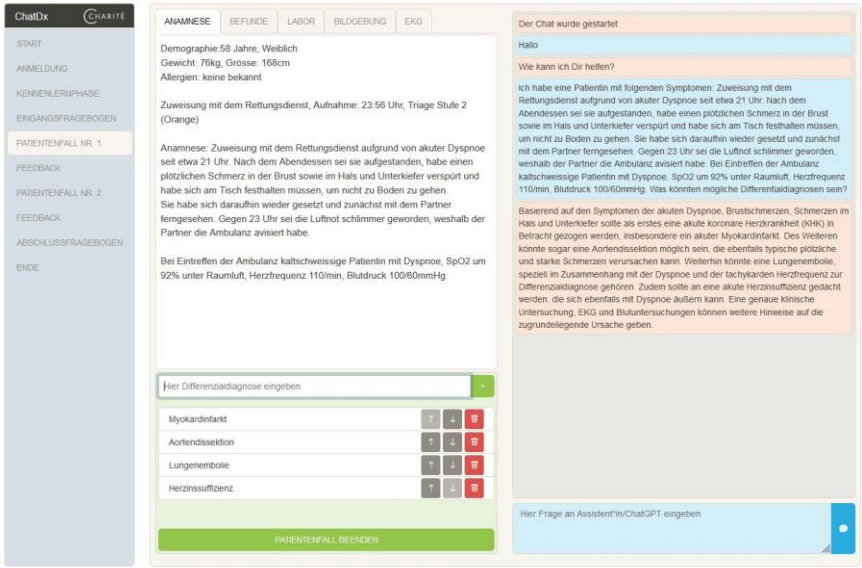

Participants
General introduction and intro to LLMs (video)
Consent, baseline survey, and demographics
ChatGPT condition
Human condition
Get to know coach
Get to know coach
Diagnose 1st case
Diagnose 1st case
Survey 1st case
Survey 1st case
Diagnose 2nd case
Diagnose 2nd case
Survey 2nd case
Survey 2nd case
Final questionnaire
Debriefing
Experiment
Data from the university about participants (student record, grades, ...)
Provide access
Get data
Researcher A from the emergency unit
Anonymize data
Dataset of baseline information about participants
Physicians trained as human coaches for the chat interaction
Get data
Researcher B from the emergency unit
Conduct
Anonymize data
Dataset of chats
Dataset of survey answers
Data extraction
Mayra Spizzo
Intercoder agreement analysis
Clean data set
Create variables
Calculate descriptive statistics
Data analysis: Unsupervised model - Clustering
Compare different methods
Dataset of chats
Coded data
Two medical master students as coders of the chat interactions
Data analysis

Figure 5: The data flow model divided into the three process steps experiment, data extraction, and data analysis.



Demographie:58 Jahre, Weiblich
Gewicht: 76kg, Grösse: 168cm
Allergien: keine bekannt

Zuweisung mit dem Rettungsdienst, Aufnahme: 23:56 Uhr, Triage Stufe 2 (Orange)

Anamnese: Zuweisung mit dem Rettungsdienst aufgrund von akuter Dyspnoe seit etwa 21 Uhr. Nach dem Abendessen sei sie aufgestanden, habe einen plötzlichen Schmerz in der Brust sowie im Hals und Unterkiefer verspürt und habe sich am Tisch festhalten müssen, um nicht zu Boden zu gehen.
Sie habe sich daraufhin wieder gesetzt und zunächst mit dem Partner ferngesehen. Gegen 23 Uhr sei die Luftnot schlimmer geworden, weshalb der Partner die Ambulanz avisiert habe.

Bei Eintreffen der Ambulanz kaltschweissige Patientin mit Dyspnoe, SpO2 um 92% unter Raumluft, Herzfrequenz 110/min, Blutdruck 100/60mmHg.

Hier Differenzialdiagnose eingeben

- Myokardinfarkt
- Aortendissektion
- Lungenembolie
- Herzinssuffizienz

Der Chat wurde gestartet

Hallo

Wie kann ich Dir helfen?

Ich habe eine Patientin mit folgenden Symptomen: Zuweisung mit dem Rettungsdienst aufgrund von akuter Dyspnoe seit etwa 21 Uhr. Nach dem Abendessen sei sie aufgestanden, habe einen plötzlichen Schmerz in der Brust sowie im Hals und Unterkiefer verspürt und habe sich am Tisch festhalten müssen, um nicht zu Boden zu gehen. Sie habe sich daraufhin wieder gesetzt und zunächst mit dem Partner ferngesehen. Gegen 23 Uhr sei die Luftnot schlimmer geworden, weshalb der Partner die Ambulanz avisiert habe. Bei Eintreffen der Ambulanz kaltschweissige Patientin mit Dyspnoe, SpO2 um 92% unter Raumluft, Herzfrequenz 110/min, Blutdruck 100/60mmHg. Was könnten mögliche Differentialdiagnosen sein?

Basierend auf den Symptomen der akuten Dyspnoe, Brustschmerzen, Schmerzen im Hals und Unterkiefer sollte als erstes eine akute koronare Herzkrankheit (KHK) in Betracht gezogen werden, insbesondere ein akuter Myokardinfarkt. Des Weiteren könnte sogar eine Aortendissektion möglich sein, die ebenfalls typische plötzliche und starke Schmerzen verursachen kann. Weiterhin könnte eine Lungenembolie, speziell im Zusammenhang mit der Dyspnoe und der tachykarden Herzfrequenz zur Differenzialdiagnose gehören. Zudem sollte an eine akute Herzinsuffizienz gedacht werden, die sich ebenfalls mit Dyspnoe äußern kann. Eine genaue klinische Untersuchung, EKG und Blutuntersuchungen können weitere Hinweise auf die zugrundeliegende Ursache geben.

Hier Frage an Assistent*in/ChatGPT eingeben

Figure 6: A screenshot of the patient case interface in the diagnostic task.

7 Data Model

At the conceptual level, I develop a tool to identify patterns in chat interactions of chats from a diagnostic task between a human and ChatGPT (or two humans) and create clusters of individuals with similar chat interaction behavior in the context of medical decision-making processes.

At the logical level, I apply different unsupervised clustering models, namely the K-means clustering technique and other clustering techniques, depending on the data distribution after the PCA, to explain differences in chat interactions. The models are applied for chat interactions in the ChatGPT condition, the human coach condition, and for both conditions together. In the case of the K-means clustering, the silhouette score is used to determine the quality of the clusters.

The following features are defined to be used as explanatory variables to detect patterns and clusters in the chat interactions: The type of questions being asked in the chat, the discreteness of the first question, the information acquisition prior to the beginning of the interaction, the average accuracy of the differential diagnoses, and the accuracy of the final diagnosis.

Regarding the physical level, there are no further requirements needed to store and process the data because the dataset is not exceptionally big, and the software MAXQDA does not require any specific infrastructure too (the data is stored in a cloud which is provided by the University of Bern for this application).

8 Documentation

The documentation of the data collection (e.g., the data of which participants is already available) and the coding of the chat interactions (e.g., which chats are already coded by the coders) is stored in a csv file. Further detail about this file is provided in chapter “4 Metadata”.

To ensure the traceability and reproducibility of the python code, comments will be provided throughout the script. In addition, multiple notebooks will be created for each step of the project (e.g., one notebook to clean the data set, one notebook for the descriptive analysis, one notebook for the clustering models, etc.).

9 Risks

The following three risk factors could be identified.

First, the honest and dedicated participation of the individuals in the experiment is crucial for the success of this project. If participants only click through the experiment without being engaged in the task and trying to find the correct diagnosis the data analysis does not result in a meaningful outcome. To incentivize a dedicated participation in the experiment, each participant receives a financial reimbursement of 35 Euros for his/her participation in the study (the average duration of the experiment is 40 minutes).

Second, technical errors could occur during the data collection, leading to data not being recorded. In this case with missing data for parts of the experiment, the respective participants will be excluded from the final dataset.

Third, the coding of the questions in the chat interactions could be different between the two coders. Since the categorization of the questions is an important feature in the clustering models, big differences in the codings between the coders could influence the quality of the patterns and clusters detected in the dataset. To ensure the quality of the codings, the intercoder agreement was calculated, as described in chapter “5 Data Quality”. If big differences in codings are still found in the final dataset, the chat interactions could additionally be classified using a trained model (for example by adapting a pre-trained model from Hugging Face[14]).

10 Preliminary Studies

This dataset was not used for the assignment in Module 2, but two preliminary analyses were conducted nonetheless. For the preliminary analyses, the chat interactions that have been coded so far were used ($N = 36$).

The first preliminary analysis investigates whether individuals differ in the initial patient information acquisition, depending on the chat partner. Since the human aspect is missing in ChatGPT, possible influencing factors, like for example psychological safety, may be less present, and together with the aspect of curiosity towards the AI technology, leading to more exploratory questions being asked to the ChaGPT assistant in comparison to the human assistant. This exploratory behavior could be observed by how much the individuals inform themselves about a patient case before asking questions in the chat. Accordingly, we expect that the amount of patient information acquisition, measured with the duration and the amount of patient information individuals look at, will be higher in the human condition than in the ChatGPT condition.

To preliminarily test this hypothesis, two linear mixed effects models with random intercepts for participants and tasks were conducted. The dependent variables included duration (measured in seconds) and amount of initial patient information acquisition, the independent variable consisted

[14] Guthikonda (2023)

of the variable source of assistance (dummy-coded as 1 = ChatGPT and 0 = human coach, respectively). Results show that individuals in the human coach condition looked significantly longer (b = 221.11, p = .00) and at more (b = 0.53, p = .03) patient information before they entered the chat (see Figure 7).

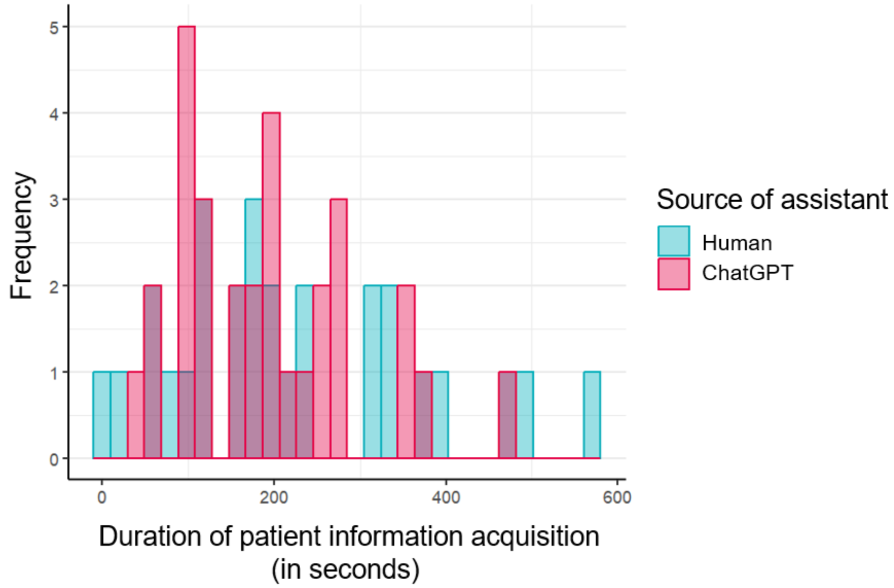


Figure 7: Distribution of the duration of patient information acquisition for the two conditions ChatGPT and human coach.

In the second preliminary analysis, I examine the influence of the chat partner on the types of questions being asked during the whole chat interaction. For this analysis, the groups of question types “questions to build a hypothesis” and “questions to test a hypothesis” are used. Following the same explanation as in the first preliminary analysis, it is expected that participants in the human condition ask more frequently questions during the chat interaction to test hypotheses, whereas in the ChatGPT condition, participants ask more frequently questions to build hypotheses.

To analyze the hypothesis, the data was examined with two linear mixed effects models, including random intercepts for participants and tasks. The absolute amount of questions being asked to build or to test hypotheses formed the dependent variables, and the independent variable included the variable source of assistance (dummy-coded as 1 = ChatGPT and 0 = human coach, respectively). Results show that individuals in the ChatGPT condition asked less frequently questions to test hypotheses (b = -4.01, p = .01) than individuals in the human coach condition (b = 10.23, p = .01; see Figure 8).

Since the results of the two preliminary analyses show significant differences between the two conditions, generating individual clustering models for the conditions seems to be reasonable.

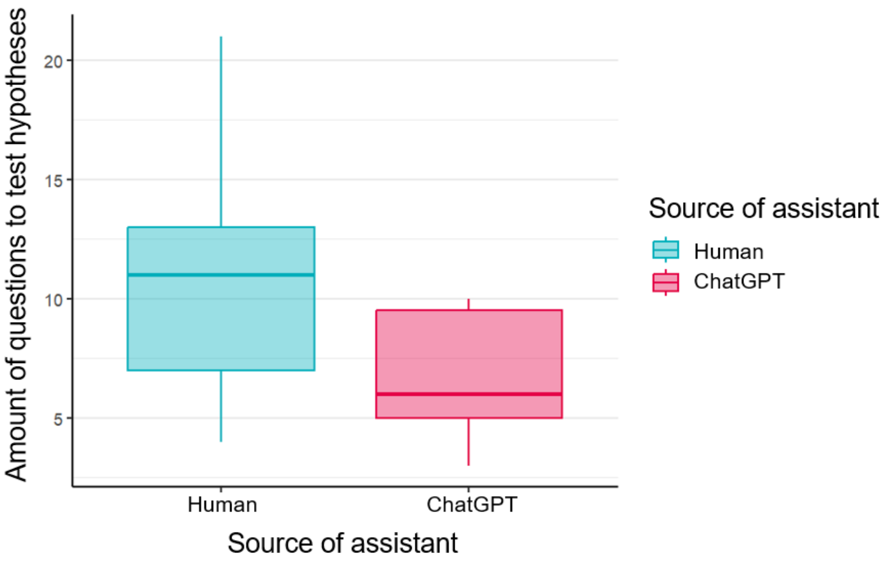


Figure 8: Boxplot to illustrate the amount of questions being asked during the chat interaction to test hypotheses, depending on the two conditions ChatGPT and human coach.

11 Conclusions

With the generation of this conceptual design report and the analysis of the already available data, I am confident that the project objective can be achieved. However, the project is dependent on the final dataset. It is important that the data collection is finalized and all the chat interactions are coded, in order to execute the project and to receive valuable patterns and clusters in the data analysis.

This project contributes to the research on human-AI interaction in the medical context by investigating chat interactions in a comprehensive analysis which allows a deeper understanding of the usage of ChatGPT support in the diagnostic process.

Hospitals or professionals in medicine could use the findings of this project to adapt the design of LLM interfaces for different clusters of physicians and for the development of measurements to reduce diagnostic errors.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 06.10.2024 Signature:

M. Spizzo

Appendix 1

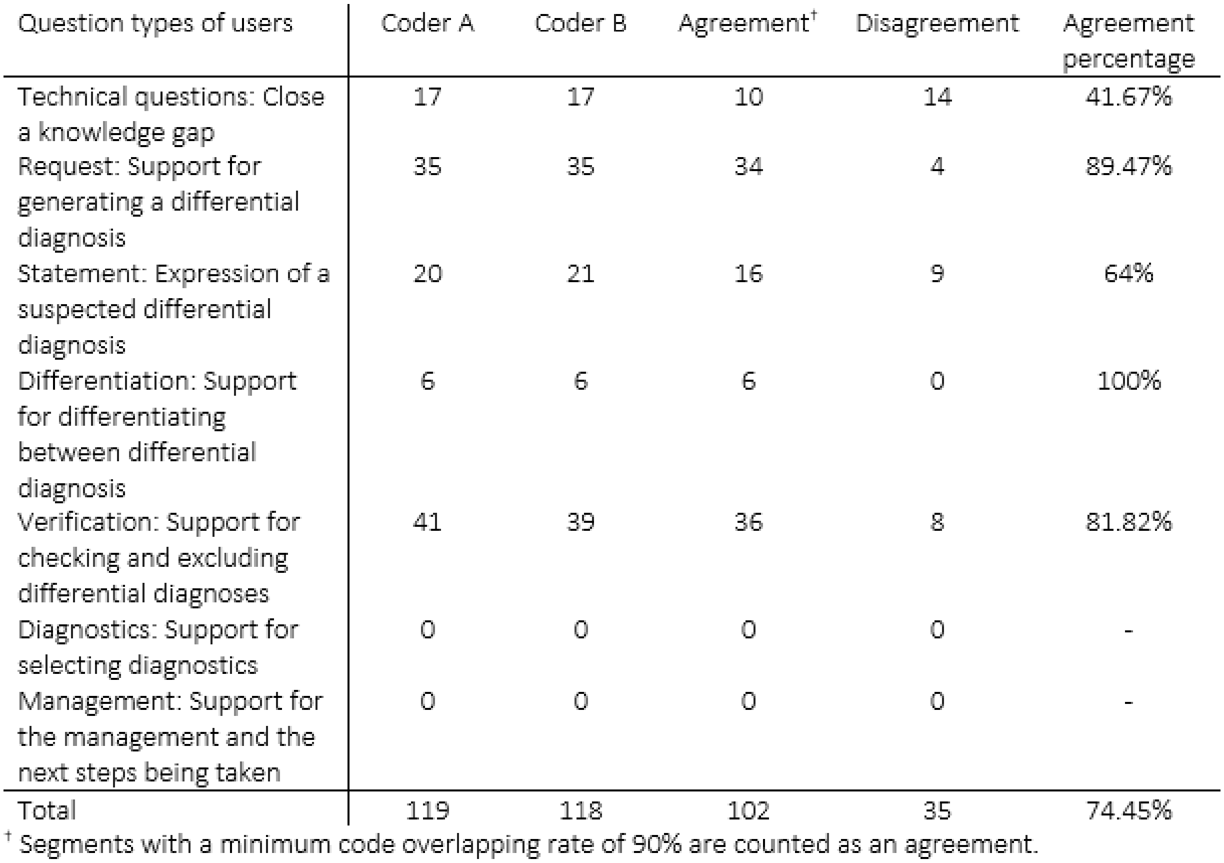
Table 2: Coding scheme for the type of questions being asked in the chats.

Category	Explanation	Example from chats
Technical questions: Close a knowledge gap	A general question without connection to a diagnosis	'Bitte den Score für Lungenembolien anzeigen' 'Wofür steht die T-negativierung im EKG?' 'Was kann Dyspnoe und Brustschmerzen auslösen?' 'Ich brauche Hilfe beim EKG auswerten'
Request: Support for generating a differential diagnosis	A request to the assistant that he generates/suggests a new differential diagnosis	'Liste mir mögliche Diagnosen' 'Worauf weisen diese Symptome / der Befund hin?' 'Welche Diagnosen sind häufig für akute Dypnoe?' 'Welche Differenzialdiagnosen gibt es für armbetonte Hemiparese rechts und links hängende Mundwinkel?'
Statement: Expression of a suspected differential diagnosis	A statement by the user on its own or as an answer to a question from the assistant	'Ich vermute einen Myokardinalinfarkt.' 'Assistent: Woran hast du gedacht? User: Myokardinfarkt, LAE.'
Differentiation: Support for differentiating between differential diagnosis	A minimum of two differential diagnoses are mentioned - question about differences between the differential diagnoses	'Ich vermute einen Myokardinfarkt oder Vorhofflimmern. Was unterscheidet die beiden?' 'Wie unterscheide ich hämorrhagischer und ischämischer Stroke?'
Verification: Support for checking and	Based on a diagnosis the question is asked about diagnostic findings (wanting	Worauf sollte ich achten, wenn ich X vermute?'

excluding differential diagnoses	a diagnosis to be verified out of the context)	
Diagnostics: Support for selecting diagnostics	Based on a diagnosis the question is asked about further diagnostics	‘Wären Blutkulturen für Pneumonie ggf. relevant?’ ‘Welche weitere Diagnostik bräuchte ich für ein kautes Nierenversagen?’
Management: Support for the management and the next steps being taken	A general question about further steps needed to being taken, or about special measurements after a certain diagnosis	‘Patient hat Vorderwandinfarkt und ischämischen Schlaganfall. Was mache ich jetzt?’ ‘Welche Massnahmen kämen in den ersten Stunden nach Einsetzen der Symptome bei ischämischem Schlaganfall in Betracht?’

Appendix 2

Table 3: Intercoder agreement.



Question types of users	Coder A	Coder B	Agreement[†]	Disagreement	Agreement percentage
Technical questions: Close a knowledge gap	17	17	10	14	41.67%
Request: Support for generating a differential diagnosis	35	35	34	4	89.47%
Statement: Expression of a suspected differential diagnosis	20	21	16	9	64%
Differentiation: Support for differentiating between differential diagnosis	6	6	6	0	100%
Verification: Support for checking and excluding differential diagnoses	41	39	36	8	81.82%
Diagnostics: Support for selecting diagnostics	0	0	0	0	-
Management: Support for the management and the next steps being taken	0	0	0	0	-
Total	119	118	102	35	74.45%

[†] Segments with a minimum code overlapping rate of 90% are counted as an agreement.

References and Bibliography

[1] Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5), S2-S23.

[2] Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., ... & Robinson, K. A. (2023). Diagnostic errors in the emergency department: a systematic review.

[3] Singh, H., Meyer, A. N., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ quality & safety, 23*(9), 727-731.

[4] Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., ... & Schauber, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. *Scandinavian journal of trauma, resuscitation and emergency medicine, 27*, 1-12.

[5] Ferdush, J., Begum, M., & Hossain, S. T. (2024). ChatGPT and clinical decision support: scope, application, and limitations. *Annals of Biomedical Engineering, 52*(5), 1119-1124.

[6] Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dreyer, K. & Succi, M. D. (2023). Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *Journal of Medical Internet Research, 25*, e48659.

[7] Center for Open Science. (2024). *There's a better way to manage your research*. https://osf.io/ (Accessed on 6 October, 2024)

[8] MAXQDA. (2024). *The #1 qualitative data analysis software with the best AI integration*. https://www.maxqda.com/ (Accessed on 6 October, 2024)

[9] Seif, G. (2018). *The 5 Clustering Algorithms Data Scientists Need to Know*. Towards Data Science. https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 (Accessed on 6 October, 2024)

[10] Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

[11] Kumar, B., Kanna, B., & Kumar, S. (2011). The pitfalls of premature closure: clinical decision-making in a case of aortic dissection. *Case Reports, 2011*, bcr0820114594.

[12] Kunina-Habenicht, O., Hautz, W. E., Knigge, M., Spies, C., & Ahlers, O. (2015). Assessing clinical reasoning (ASCLIRE): Instrument development and validation. *Advances in Health Sciences Education, 20*, 1205-1224.

[13] University of Bern. (2024). *Send large files*. Online tools. https://www.unibe.ch/research/services/online_tools/collaboration/sending/index_eng.html (Accessed on 6 October, 2024)

[14] Guthikonda, S. (2023). *Text Classification Using Hugging Face(Fine-Tuning)*. https://medium.com/@sandeep.ai/text-classification-using-hugging-face-fine-tuning-43c7416b049b (Accessed on 6 October, 2024)