IDRE: AI Generated Dataset for Enhancing Empathetic Chatbot Interactions in Italian language.

Simone Manai[1,2,*,†], Laura Gemme[2,†], Roberto Zanoli[3] and Alberto Lavelli[3]

[1] *University of Trento, 38123 Trento, Italy*
[2] *Lutech-Softjam, 16148 Genova, Italy*
[3] *Fondazione Bruno Kessler, 38123 Trento, Italy*

Abstract

This paper introduces IDRE (**I**talian **D**ataset for **R**ephrasing with **E**mpathy), a novel automatically generated Italian linguistic dataset. IDRE comprises typical chatbot user utterances in the healthcare domain, corresponding chatbot responses, and empathetically enhanced chatbot responses. The dataset was generated using the Llama2 language model and evaluated by human raters based on predefined metrics. The IDRE dataset offers a comprehensive and realistic collection of Italian chatbot-user interactions suitable for training and refining chatbot models in the healthcare domain. This facilitates the development of chatbots capable of natural and productive conversations with healthcare users. Notably, the dataset incorporates empathetically enhanced chatbot responses, enabling researchers to investigate the effects of empathetic language on fostering more positive and engaging human-machine interactions within healthcare settings. The methodology employed for the construction of the IDRE dataset can be extended to generate sentences in additional languages and domains, thereby expanding its applicability and utility. The IDRE dataset is publicly available for research purposes.

Keywords

Empathy, LLMs, Llama2, Dataset, Chatbot, Healthcare

1. Introduction

Emotional intelligence has been widely recognized as a crucial factor influencing human communication, impacting aspects such as behavioral choices and the interpretation of information [1]. Consequently, there has been a growing interest in developing chatbots capable of exhibiting empathetic responses [2] [3] [4]. While significant strides have been made in this direction, the integration of empathy into commercial chatbots remains challenging due to the rigid constraints imposed by business rules such as the response must not lose the original meaning and the dialogue must maintain structure.

To address this limitation, one possible approach is to build a layer that rephrases the bot's response by increasing empathy without altering the structure or meaning of the underlying dialogue. This strategy offers the potential to enhance user experience and create a foundation for more sophisticated empathetic dialogue systems.

To facilitate the development of such systems, a robust dataset containing empathetic responses is essential. Despite the increasing body of research on emotion recognition and generation in human-computer interaction, there is a notable absence of publicly available datasets specifically focused on empathy in chatbot interactions.

This paper introduces the IDRE dataset, a new Italian language resource comprising human-bot interactions within the healthcare domain. The dataset is available publicly, and the address is provided in the Online Resource section. The dataset includes the user questions, original bot responses and corresponding empathetic reformulations for a total of 480 sentences, providing a valuable foundation for research and

---

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

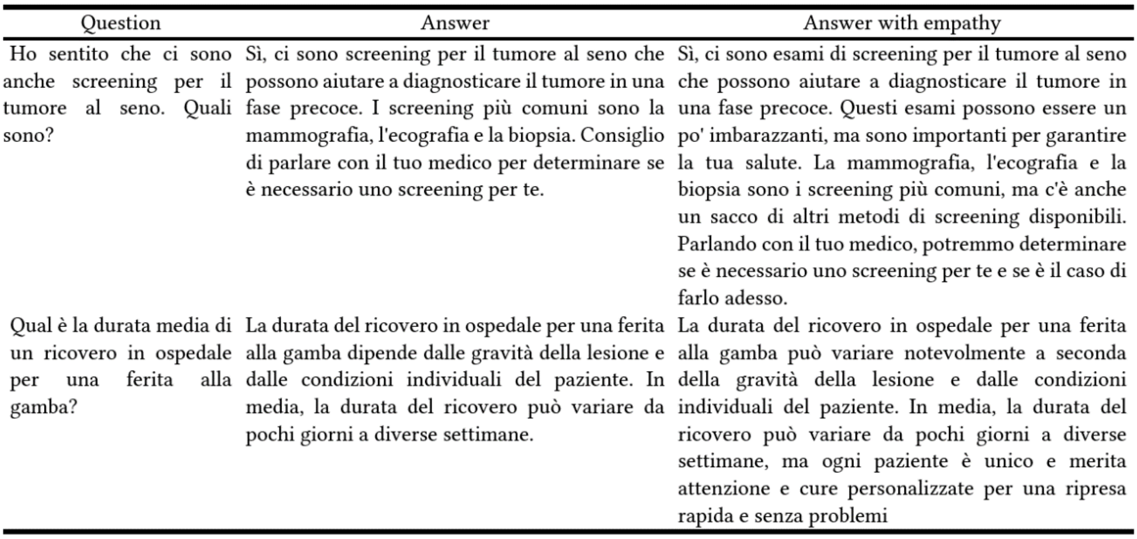
[†] These authors contributed equally.

simone.manai@unitn.it (S. Manai); l.gemme@softjam.it (L. Gemme); zanoli@fbk.eu (R. Zanoli); lavelli@fbk.eu (A. Lavelli)

0000-0002-7175-6804 (A. Lavelli); 0000-0003-0870-0872 (R. Zanoli)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1
Examples of Question, Answer and Answer with empathy



Question	Answer	Answer with empathy
Ho sentito che ci sono anche screening per il tumore al seno. Quali sono?	Sì, ci sono screening per il tumore al seno che possono aiutare a diagnosticare il tumore in una fase precoce. I screening più comuni sono la mammografia, l'ecografia e la biopsia. Consiglio di parlare con il tuo medico per determinare se è necessario uno screening per te.	Sì, ci sono esami di screening per il tumore al seno che possono aiutare a diagnosticare il tumore in una fase precoce. Questi esami possono essere un po' imbarazzanti, ma sono importanti per garantire la tua salute. La mammografia, l'ecografia e la biopsia sono i screening più comuni, ma c'è anche un sacco di altri metodi di screening disponibili. Parlando con il tuo medico, potremmo determinare se è necessario uno screening per te e se è il caso di farlo adesso.
Qual è la durata media di un ricovero in ospedale per una ferita alla gamba?	La durata del ricovero in ospedale per una ferita alla gamba dipende dalle gravità della lesione e dalle condizioni individuali del paziente. In media, la durata del ricovero può variare da pochi giorni a diverse settimane.	La durata del ricovero in ospedale per una ferita alla gamba può variare notevolmente a seconda della gravità della lesione e dalle condizioni individuali del paziente. In media, la durata del ricovero può variare da pochi giorni a diverse settimane, ma ogni paziente è unico e merita attenzione e cure personalizzate per una ripresa rapida e senza problemi

development in empathetic chatbot technology, see Table 1 for an example. The paper also elaborates on the methodology employed for dataset generation, highlighting its applicability to diverse domains and languages.

2. Related Works

The development of empathetic chatbots capable of understanding and responding to human emotions represents a research area of growing interest [5]. However, building such systems requires high-quality datasets that include examples of human-machine interactions with empathic components.

Despite the growing availability of datasets for machine learning and natural language processing, the lack of resources dedicated specifically to empathetic Italian-language chatbots represents a significant challenge.

There are datasets that contain emotional information, such as [6] [7] [8] [9]. However, these resources focus primarily on labelling words or sentences with generic emotions and do not provide the context for complex, nuanced conversational interactions like those required for developing empathetic chatbots.

3. Dataset

This chapter details the methodology employed for the construction of the IDRE dataset and outlines the evaluation process implemented.

The dataset created consists of 480 sentences and roughly 18k total tokens divided as follows: 2k for the question, 7k for the bot's response and 9k for the response with empathy.

3.1. Dataset Creation

The IDRE dataset comprises triplets of sentences, the first sentence represents a user query, the second sentence is the corresponding response generated by a chatbot, and the third sentence is a transformed version of the second sentence intended to enhance its empathetic tone.

The sentence generation process was done by the Llama2 13B language model [11], operating on an Azure Virtual Machine equipped with four NVIDIA Tesla V100 GPUs. The choice of Llama 2 was motivated by its open-source nature, which allowed flexible and provider-independent access.

The dataset generation process consists of two phases as illustrated in Figure 1:

QnA Sentence Generation: To ensure the generation of empathetic and compassionate responses, the healthcare domain was selected as the focus for the initial set of bot-human sentence pairs. This domain, characterized by sensitive topics, is well-suited for evaluating the model’s ability to generate empathetic responses.

The thirteen specific topics chosen for the sentence pairs were invented for the purposes of the experiment: 'information on breast cancer', 'breast cancer prevention', 'therapies for breast cancer', 'psychological support after a cancer diagnosis', 'life expectancy after a cancer diagnosis', 'psychological support after surgery', 'hospital admissions', 'post-operative care', 'information on leukemia', 'psychological support', 'anti-cancer therapies', 'information on stroke', and 'preparation for surgeries.'

An initial set of bot-human sentence pairs was generated using the Llama2 model. These pairs simulated a typical chatbot interaction concerning a specific health issue or domain. For instance, a human query such as "What are the symptoms of COVID-19?" would elicit a corresponding chatbot response like "The most common symptoms of COVID-19 are fever, dry cough, and tiredness".

Empathy Enhancement: After the generation of the initial sentence pairs, an empathy enhancement process was undertaken. Leveraging the Llama2 model once more, the chatbot responses were modified to convey a more empathetic tone. This was achieved by prepending expressions of concern or appreciation, and by substituting specific words to engender a supportive demeanor. To illustrate, the aforementioned chatbot response could be transformed into "I understand that you're concerned about COVID-19. Some common symptoms include fever, dry cough, and fatigue".

Both prompts are included in the Appendix.

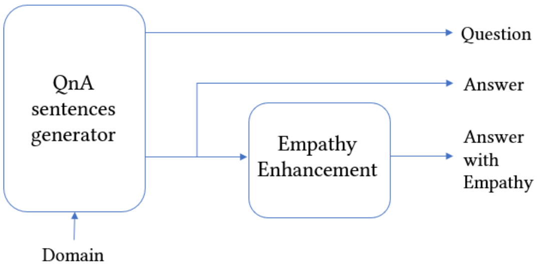


Figure 1: Dataset generation process

3.2. Evaluation Methodology

To ensure the quality of the generated sentences, a rigorous evaluation process was implemented. Twelve volunteer annotators from Lutech-Softjam, experienced IT developers and project managers with a solid understanding of chatbot domain, participated. Despite lacking prior experience in linguistic annotation, their familiarity with chatbots significantly accelerated the evaluation process. Before start, they underwent comprehensive training on the evaluation task.

Each evaluator was assigned 70 sentences for assessment. To ensure diverse evaluations, 40 sentences were unique to each evaluator and used for dataset creation, while 30 common sentences were evaluated by all evaluators, solely for measuring agreement and will not be part of the dataset. This approach ensured that each sentence received focused evaluation while also providing a consistent assessment across evaluators.

The evaluation process involves the administration of a metric-specific question, which requires a response on a scale of 1 to 5.

The rating scale used is the following:

1. Totally disagree
2. Disagree
3. Neutral
4. Agree
5. Totally agree

The specific metrics used in this evaluation are:

- **Bot sentence correctness:** measures the absence of spelling, grammatical, or punctuation errors in the question and the bot's answer. The question used is: "Il testo della risposta con empatia è corretto sia dal punto di vista grammaticale che semantico."
- **Absence of English words in bot sentences**: checks if there are any words or sentences in English within the sentences generated by the model. The question used is: "Nel testo della domanda dell'utente e della risposta del bot (colonne QUESTION e ANSWER) non sono presenti parole o frasi in lingua inglese, a meno che non siano di uso comune in italiano (ad esempio "badge", "sport", ecc.)"
- **Empathic answer correctness:** measures the absence of spelling, grammatical, or punctuation errors in the bot's answer with the insertion of empathy. The question used is: "Il testo della risposta con empatia è corretto sia dal punto di vista grammaticale che semantico."
- **Absence of English words in empathic sentences:** checks if there are any words or sentences in English within the sentences with empathy generated by the model. The question used is: "Nel testo della risposta con empatia non sono presenti parole o frasi in lingua inglese, a meno che non siano di uso comune in italiano (ad esempio "badge", "sport", ecc.)"
- **Semantic coherence:** measures if the bot's answer and the bot's answer with empathy are semantically similar. The question used is: "La risposta con empatia ha lo stesso significato semantico della risposta del chatbot. Non ci sono concetti mancanti o contraddittori"
- **Empathy increase:** measures if the bot's answer with empathy has an effective increase of empathy compared to the bot's answer. The question used is: "La frase nella colonna ANSWER WITH EMPATHY esprime più empatia rispetto alla frase nella colonna ANSWER"

4. Dataset Analysis

This section analyses data quality by examining both the distribution of agreement scores and the level of inter-annotator agreement (IAA).

Due to a limited pool of available evaluators, the dataset was constrained to 480 annotated sentences. These sentences were evenly distributed among 12 volunteers, each assessing 40 sentences (excluding the 30 sentences used for measuring agreement). This approach was made to ensure the quality of the annotations while preventing evaluator fatigue. Nevertheless, a more in-depth analysis reveals that 223 sentences, equal to 46.5% of the total, have the score grater or equal to 3 on all the metrics considered. This means that these sentences were judged to be of high quality in every aspect analysed. This subset of data can be used to finetune language models.

To obtain a more robust analysis and less subject to small variations, the annotation categories were grouped into three macro-categories: scores 1 and 2, score 3 (neutral) and score 4 and 5.

The analysis of sentences with lower score (1 and 2) revealed three key factors: grammatical errors, the presence of non-Italian words and lack of a significant increase in empathy as shown in Figure 2.

Grammatical Errors: A substantial portion of sentences with lower score exhibited grammatical errors (words in red). This highlights the importance of incorporating robust grammar checks during the generation process. Example: *"Ohimini, cara/o utente, è comprensibile che durante il trattamento del tumore possa esserti difficile gestire i sintomi. Sono qui per aiutarti a trovare soluzioni e supporti per farcela insieme".* *"Ohimini"* is a made-up word and *"supporti"* contains a typo.

Non-Italian Words: the lower score sentences frequently included non-Italian words (words in red), primarily English. This deviation from the dataset's focus on Italian-language interactions can be attributed to the underlying multilingual language model, which was predominantly trained on English text. This highlights the need for improved language model training to prioritize Italian vocabulary. Example: "*Per prevenire le infezioni after surgery, è importante seguire le istruzioni del medico e del personale ospedaliero, come ad esempio lavare le mani frequentemente, evitare di toccare la ferita e utilizzare dispositivi di protezione individuali."*

Lack of a significant increase in empathy: Among the lower score sentences (173, representing 36%), the transformed responses (indicated by the blue and orange columns) did not exhibit a significant rise in empathy or indecision compared to the original chatbot responses. This suggests that further refinement of the empathy-enhancing techniques might be necessary.

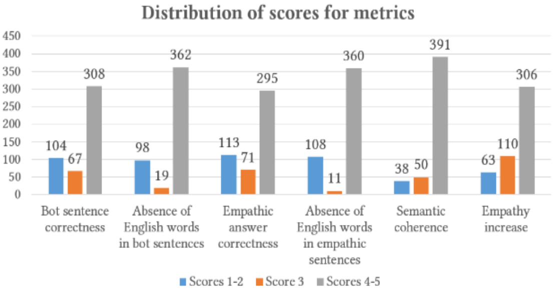


Figure 2: Scores Distribution for all metrics.

Regarding of analysis of the inter-annotator agreement (IAA) for the annotations generated as outlined in Section 3.2. Fleiss' kappa coefficient was employed to quantify the level of concordance between multiple annotators while accounting for potential chance agreement. Kappa values range from -1 to 1, with negative values indicating agreement below chance, values between 0 and 0.2 representing slight agreement, 0.21 to 0.4 fair agreement, 0.41 to 0.6 moderate agreement, 0.61 to 0.8 substantial agreement, and values exceeding 0.8 denoting almost perfect agreement.

The calculation of kappa coefficients on aggregated categories allowed to evaluate the inter-annotator agreement in a more robust way. The results are summarized in Table 2. Notably, the highest levels of agreement were observed for metrics related to the presence of English words. This finding is likely attributable to the relative simplicity of this specific annotation task. Conversely, metrics assessing other linguistic features exhibited lower, yet still acceptable, levels of agreement, generally falling within the moderate range.

Table 2
Agreement result

Metrics	Fleiss Kappa	Aggregate Fleiss Kappa
Bot sentence correctness	0.608	0.821
Absence of English words in bot sentences	0.781	0.927
Empathic answer correctness	0.566	0.807
Absence of English words in empathic sentences	0.782	0.948
Semantic coherence	0.587	0.881
Empathy increase	0.645	0.840

Figure 3 presents the distribution of annotations for three metrics: "Empathy increase", "Bot sentence correctness", and "Absence of English words in bot sentences". The distribution for "Absence of English words in bot sentences" exhibits a marked concentration towards the highest score (5), indicating a strong consensus among annotators regarding the absence of

English words in bot sentences. In contrast, the distribution for “Empathy increase” or “Bot sentence correctness” is more dispersed across the entire range of possible scores, suggesting a greater degree of variability in annotator assessments of bot empathy increase.

The observed disparity in distribution patterns between the metrics can be attributed to the inherent nature of the annotation tasks. The task of identifying the absence of English words in bot sentences is relatively straightforward and objective, leading to a higher degree of agreement among annotators. On the other hand, assessing bot empathy increase involves a more subjective judgment of factors such as grammatical accuracy, coherence, and relevance, resulting in a wider range of annotations.

The same behaviour can be noticed with metric “Empathic answer correctness”.

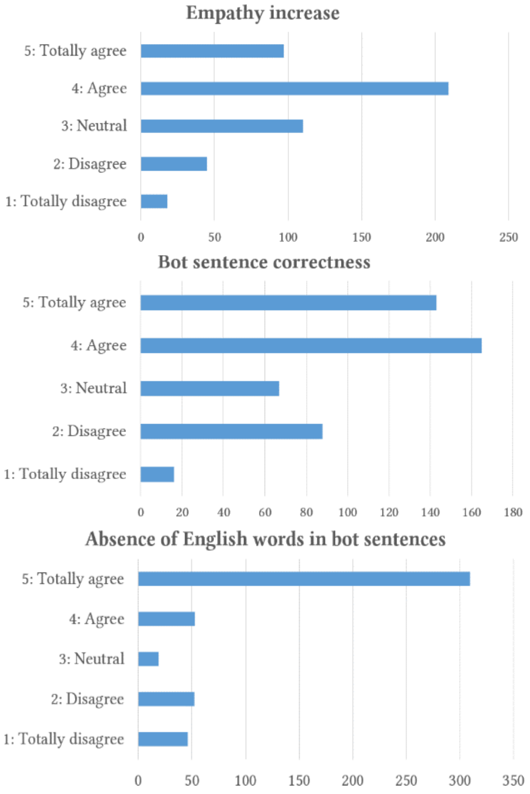


Figure 3: Distribution of "Empathy increase", “Bot sentence correctness”, and "Absence of English words in bot sentences"

5. Discussion and Conclusion

In this work, we have presented the creation of a dataset of sentences representing typical interactions with a healthcare chatbot. The dataset includes both user input sentences and empathetic responses generated by the chatbot. Human validation has confirmed the quality and usefulness of the dataset for developing and evaluating empathetic chatbots in the healthcare domain.

This work presents a two-pronged contribution to the field of empathetic chatbots, specifically focusing on the Italian language.

Firstly, it addresses the critical issue of data scarcity by providing a high-quality, annotated dataset for training and evaluating empathetic chatbots within a healthcare context. This dataset can be employed to fine-tune large language models (LLMs) such as Llama2, enabling them to generate responses with demonstrably enhanced empathetic qualities. The limitations of non-fine-tuned models are exemplified through the observation that they can produce factually incorrect or unempathetic sentences (e.g., *" Il tuo corpo è vulnerabile al rischio del tumore al seno a causa della tua età avanzata, nonostante la tua vitalità e forza interiori. La storia familiare di tumori al seno nella tua famiglia e la tua condizione di obesità possono aumentare il rischio, come pure l'abuso di tabacco e alcool. Inoltre, la tua scelta di non avere figli o di averli dopo l'età di 35 anni può aggiungere ulteriore rischio al tuo corpo."*). By leveraging the proposed dataset and selecting sentences with demonstrably high empathy scores, a targeted training set can be constructed specifically for this purpose. This, in turn, allows for the fine-tuning of the LLM, significantly improving its ability to generate empathetic responses in a healthcare setting.

Secondly, the work contributes a rigorous human validation methodology for evaluating the effectiveness of empathy expression in chatbots. This methodology provides a valuable tool for researchers and developers working in this domain.

5.1. Future Work

In the future, we intend to expand the work in two main directions:

Domain expansion: We will explore the creation of similar datasets for other domains, such as customer service or education, to assess the applicability of our approach in different contexts.

Comparison of language models: We will conduct a comparative study to evaluate the performance of different language models in generating empathetic chatbot responses. This study will allow us to identify the most suitable language model for this specific task.

We believe that this work represents an important step towards the development of empathetic chatbots capable of offering a more natural and engaging user experience, especially in sensitive contexts such as healthcare.

Acknowledgements

The authors would like to express their sincere gratitude to Lutech-Softjam colleagues for their invaluable contributions to the evaluation of the dataset used in this study. Their expertise and meticulous work in assessing the dataset's quality and relevance were instrumental in ensuring the robustness of our research findings.

References

[1] Fellous, Jean-Marc and M. A. Arbib, "Who needs emotions?: The brain meets the robot.," *Oxford University Press,* 2005.

[2] Z. Emmanouil, G. Paraskevopoulos, A. Katsamanis and A. Potamianos., "EmpBot: A T5-based Empathetic Chatbot focusing on Sentiments," arXiv preprint arXiv:2111.00310., 2021.

[3] S. Jamin, P. Xu, A. Madotto and P. Fung, "Generating empathetic responses by looking ahead the user's sentiment.," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[4] F. Liu, Q. Mao, LiangjunWang, N. Ruwa, J. Gou and Y. Zhan, "An emotion-based responding model for natural language conversation," *Springer Science+Business Media,* 2018.

[5] Q. Guo, Z. Zhu, Q. Lu, D. Zhang and W. Wu, "A Dynamic Emotional Session Generation Model Based on Seq2Seq and a Dictionary-Based Attention Mechanism," *Appl. Sci.,* p. 10, 2020.

[6] R. Sprugnoli, "MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion," in *Proceedings of the Seventh Italian Conference on Computational Linguistics*, 2020.

[7] S. M. Mohammad, "Practical and ethical considerations in the effective use of emotion and sentiment lexicons," arXiv preprint arXiv, 2020.

[8] A. Welivita, Y. Xie and P. Pu, "A Large-Scale Dataset for Empathetic Response Generation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* pp. 1251-1264, 2021.

[9] H. Rashkin, E. M. Smith, M. Li and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207,* 2018.

[10] A. Welivita, Y. Xie and P. Pu, "A Large-Scale Dataset for Empathetic Response Generation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* pp. 1251--1264, 2021.

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale and others, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288,* 2023.

[12] S. C. Gadanho, "Learning Behavior-Selection by Emotions and Cognition in a Multi-Goal Robot Task.," *Journal of Machine Learning Research,* vol. 1, pp. 385-412, 2003.

A. Online Resource

The dataset can be downloaded at https://github.com/smanai/idre

B. Appendix

Below the prompts used for both steps of dataset creation are shown.

Prompt for QnA Sentence Generation: """genera {} coppie di domande utente e risposta di un assistente virtuale.

Le domande devono essere in lingua italiana e rappresentano frasi tipiche una persona che vuole informazioni nel dominio "{}".

Le risposte sono quelle di un tipico chatbot di un call center di un'azienda ospedaliera.

Le risposte devono solo esporre dei fatti oggettivi e scientifici ma prive di empatia.

la struttura del output deve essere:

#

utente:

assistente:"""

Prompt for Empathy Enhancement: """La seguente frase è la risposta di un chatbot di un call center di un ospedale ad una persona che richiede informazioni. La frase è informativa, ma non trasmette empatia per la situazione della persona che chiama. Puoi modificare la seguente frase aggiungendo l'empatia mancante?

Puoi modificare la frase aggiungendo testo o modificandolo ma deve mantenere lo stesso significato semantico.

la frase modificata deve essere scritta in lingua italiana.

Non devi scrivere altro testo oltre alla frase trasformata.

inizia la modifica della frase con il carattere "-" come in un elenco puntato.

"""